基于药物评价的情感分析模型

小组成员： 申若曦 李润涵 姜来

1. 项目背景

在药物生产和研究过程中，临床用药反馈对于产品的改进至关重要。在精准医疗、个体化用药的大趋势下，人们更加追求用药过程中的精准度、感受与体验。随着社交媒体与信息技术的发展，网络平台成为了人们表达观点的重要途径，在线评论网站和意见论坛里包含大量关于用户在多个产品领域的偏好和体验的信息。这些信息可以利用数据挖掘方法（比如情感分析）来获得有价值的信息。

在药物应用过程中，越来越多的人在社交媒体上表达用药后的反馈，这对药企、用药医师等各个方面均有较高的参考价值。但这些数据的数量庞大，难以人工判断某一药物效果。因此，结合本组成员的学科背景，本项目采用制药领域的在线用户评论进行患者的情感分析。这些评论包含了药物的有效性、副作用、使用体验等多方面信息。分析药品评论所获得的信息，能帮助医生进行临床决策，并通过获得集体感受来改善公共卫生监测。本项目将基于这些数据，尝试一些机器学习模型来评估患者对药物的情感倾向，将对应药物进行等级分类。

1.1 研究思路

将对药物的评论转化为数值矩阵，通过数值的分析与预测对评价的情感倾向进行模型训练，进而得到对药物的总体评级。

1.2 数据来源

本项目所用数据来源于Felix Gräßer等人的会议论文《Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning》。该论文利用网络爬虫收集了多个药物评价网站的数据，对多种药物品类、针对病症、患者评价等数据进行了整理，为本项目数据的科学性和全面性提供了保障。

2.数据预处理

2.1 数据概览

本项目共纳入161,000条数据作为训练集，53,800条数据作为测试集。数据集中包含六类信息：药品名称、对应病症、患者评价、患者打分、评价日期、和“点赞数”（认为该评价有用的用户数量）。下图为数据包含信息的概览：

uniqueID	drugName	condition	review	rating	date	usefulCount
206461	Valsartan	Left Ventricular Dysfuncti	"It has no side effect, I take it in combination of Bystolic	9	20-May-12	27
95260	Guanfacine	ADHD	"My son is halfway through his fourth week of Intuniv. W	8	27-Apr-10	192
92703	Lybrel	Birth Control	"I used to take another oral contraceptive, which had 21 The positive side is that I didn't have any other si	5	14-Dec-09	17
138000	Ortho Evra	Birth Control	"This is my first time using any form of birth control. I&#	8	3-Nov-15	10
35696	Buprenorphine	Opiate Dependence	"Suboxone has completely turned my life around. I feel	9	27-Nov-16	37
155963	Cialis	Benign Prostatic Hyperpl	"2nd day on 5mg started to work with rock hard erectio	2	28-Nov-15	43
165907	Levonorgestrel	Emergency Contraceptio	"He pulled out, but he cummed a bit in me. I took the Pl	1	7-Mar-17	5
102654	Aripiprazole	Bipolar Disorde	"Abilify changed my life. There is hope. I was on Zoloft a	10	14-Mar-15	32
74811	Keppra	Epilepsy	" I Ve had nothing but problems with the Keppera : con	1	9-Aug-16	11
48928	Ethinyl estradio	Birth Control	"I had been on the pill for many years. When my doctor	8	8-Dec-16	1
29607	Topiramate	Migraine Prevention	"I have been on this medication almost two weeks, start	9	1-Jan-15	19

我们计划将数据预处理后，从数据分布中获得信息，建立分类标准，选择合适的模型进行训练并优化。

2.2 数据预处理

处理缺失值

数据的对应病症列有约七百行数据有缺失值，因为700对于数据总量（161000）较小，因此我们选择删除含有缺失值的行。此外，在文本信息处理时，我们发现数据量过大，一行处理代码需要七八小时的时间才能完成，因此随机选了一万条评论作为研究对象。

删除相关性较弱的信息

我们删除了对评价药物效果无关的列，如药品ID、对应病症、评价日期、点赞数等。因部分药物的评价很少，可能会因为评价人的主观性造成评价比较偏颇，缺乏代表性，因此我们删除了少于20个评价的药物。

文本信息处理

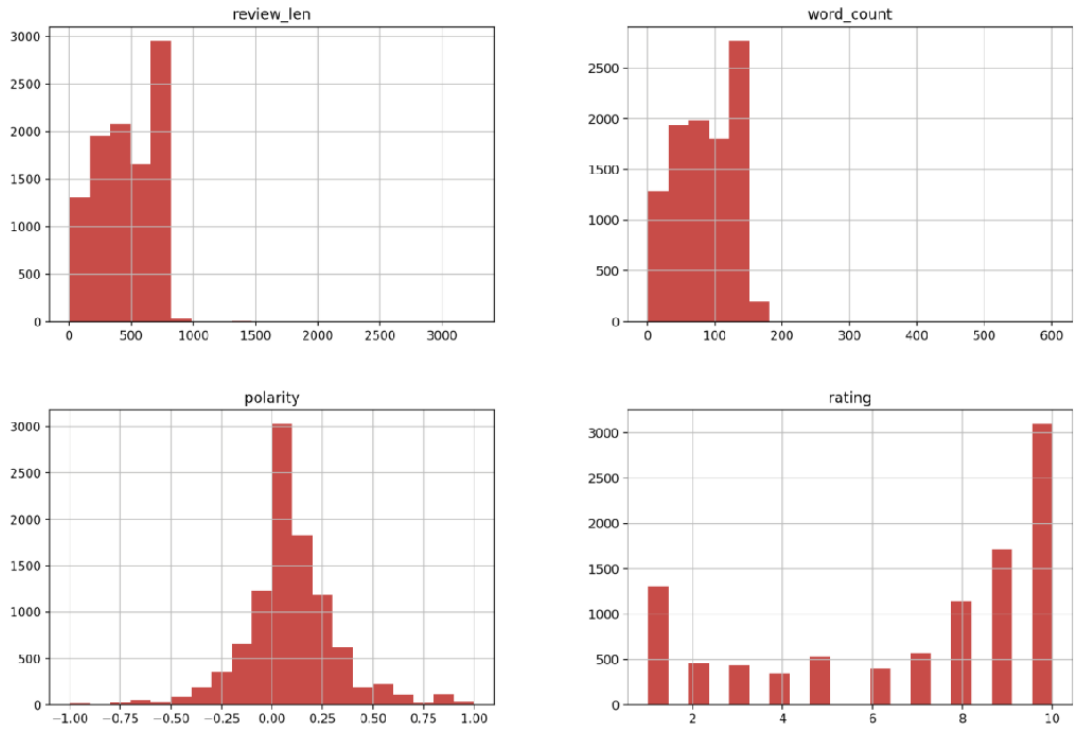
1. 我们首先对文本的格式进行处理，删除标点符号、空白符号，并将大写字母转化为小写字母。
2. 之后我们删除了Stop words。Stop words是指一些在句子中常见的词，如“the”、“and”、“in”等。虽然他们比较常见，但不能传达对文本的有用信息。 我们从给定语料库的文本中删除这些词，识别更罕见和可能与我们感兴趣的内容更相关的词。
3. 然后我们删除了频率过少的词，比如那些在所有评论中都只出现过一次的词。
4. 之后我们用了nltk里的两个包来提取词干，使相似种类的单词位于同一个词干之下。我们用的第一个包是SnowballStemmer，它可以删除相似单词。用的第二个包是PorterStemmer，它可以删除单词中常见的形态词尾和固定词尾。
5. 最后，我们想知道评论所带有的情绪偏向，因此我们加入了特征——情感极性。NLTK已经有一个内置的、经过预先训练的情感分析器，适用于社交媒体

的语言，比如带有一些俚语和缩写的短句。 因此这个包很适合用于药物评论的分析。 经过处理，我们得到的情感极性是处于-1 ~ 1之间的一个值，其中-1代表消极情绪，0代表中性，1代表积极情绪。

数据分布

在数据预处理完之后，我们利用图表展示了数据在各特征值上的分布情况。

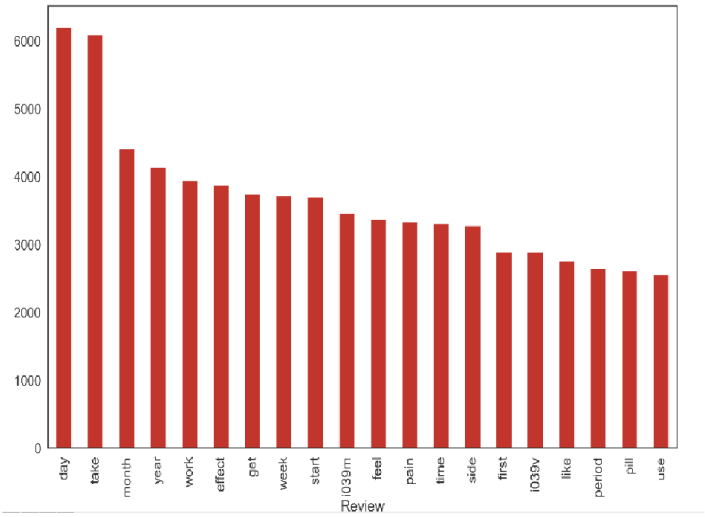
我们统计得到了评论的长度、单词数量、情感极性和打分的分布。



- 长度和词汇数量的分布都很类似，绝大部分评论长度都少于800，单词数量少于200；说明评论都较短。
- 情感极性呈钟形曲线，服从正态分布。打分的分布可以看出，极端分数，比如1和10的数量较多，而中间值的数量很少。但根据常理来看，评论的情感和打分的分布应该较相似，感情是消极情绪的打分应该较低，而感情是积极情绪的打分应该也更高。但最后的结果却不符合这一规律，因此我们推测是因为用户的打分的主观性较强，并没有一个统一标准。这也从侧面说明了，并

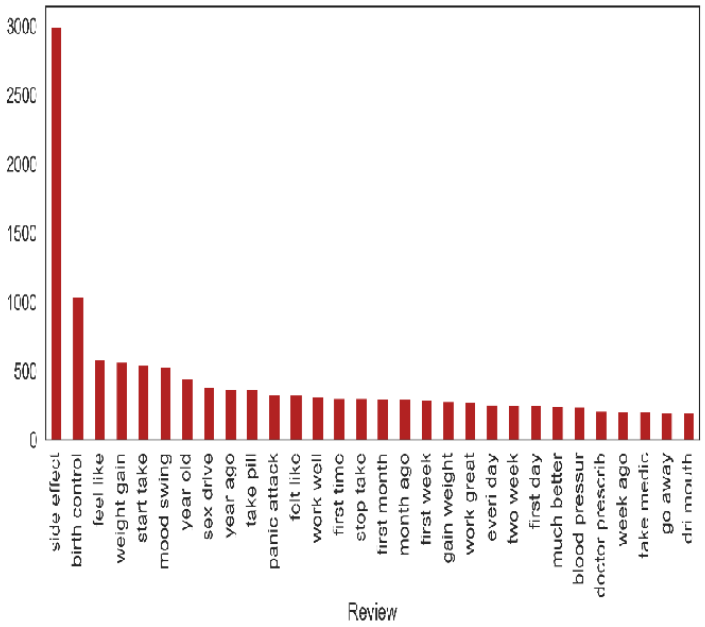
不能单纯根据用户的打分来进行药物评级，而应该进一步的数据挖掘，通过情感分析等方法，得到更加客观、全面的评级。

- 接下来，我们看一下词频的统计。
    - 1-words词频分析



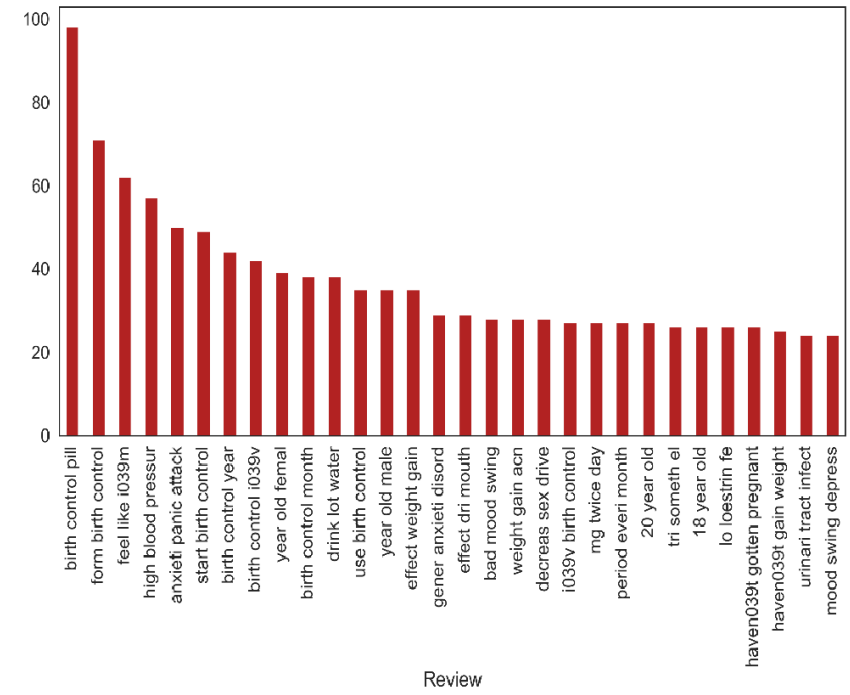
我们可以看到排名最前面的几个词是“day”、“take”，“month”，“year”，这说明患者对用药频率可能比较在意。

    - 2-words词频分析



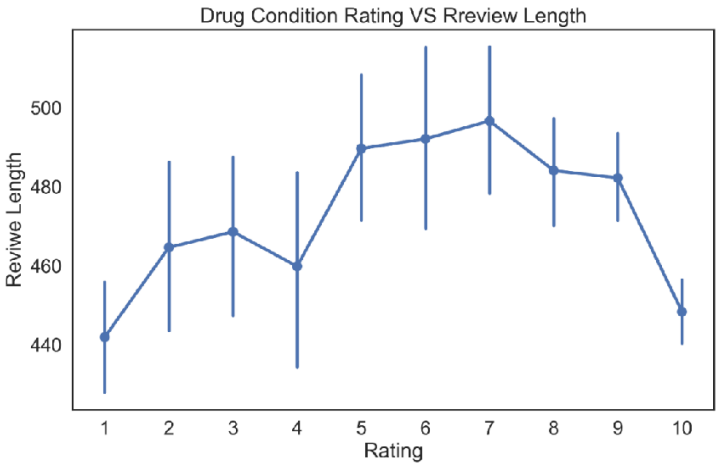
2-words词频分析排名最前面的几个词是副作用、避孕、用户感觉。

    - 3-words词频分析



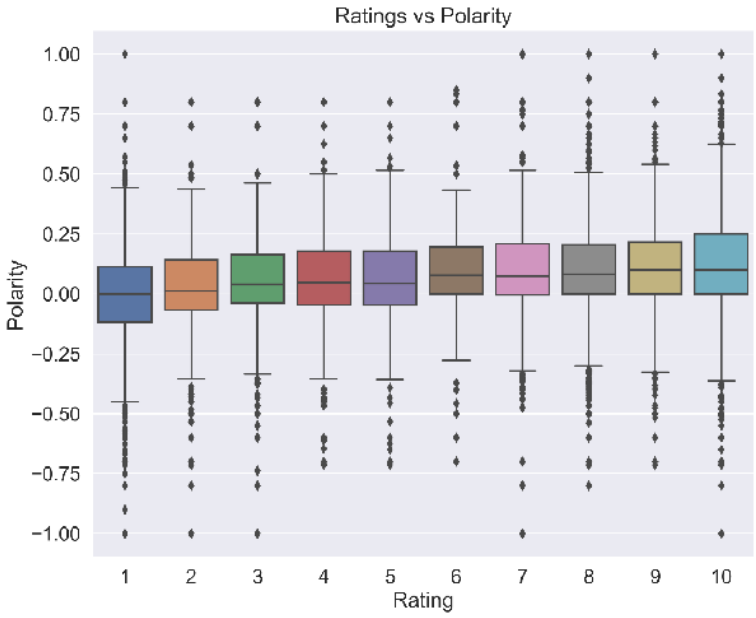
3-words词频分析和2-words词频分析的结果类似

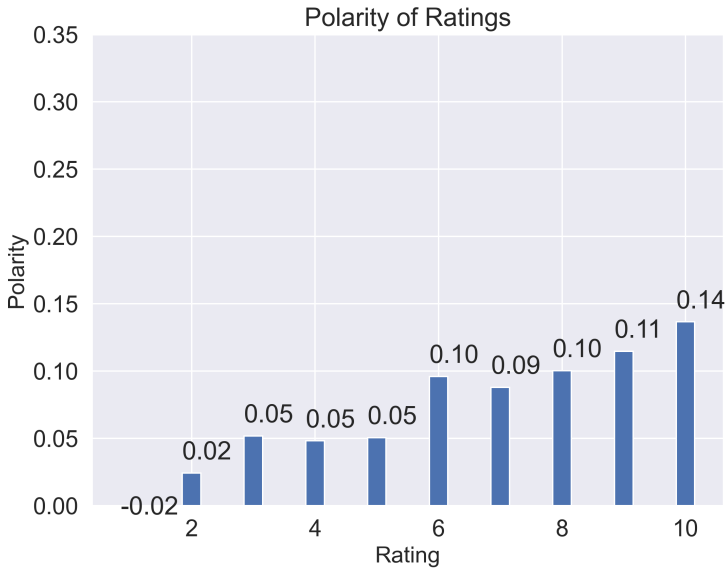
- 之后，我们又看了评论长度和患者打分的关系



可以看到10分评价的长度偏短，5-9分评价的长度偏长。

- 最后我们查看了打分与情感极性的关系





我们观察到，平均极性会随着打分的提高而上升，但是在1分评价中异常值较多。打分>3时，评论的平均情感极性为积极情绪；打分<3时，评论的平均情感极性为消极情绪。因此我们以以3为分界进行初步分类，>3的评价的标签为正向，<3的评论的标签为负向。

3. 模型部分

我们对文本信息向量化处理后，划分训练集和测试集，将Positively Rated(0-1)作为y值，向量化文本作为x值进行模型训练。我们选取了四个模型，神经网络、SVM、逻辑回归和朴素贝叶斯，通过对比指标选出表现最好的模型。

3.1 二分类模型初探

逻辑回归模型

- 定义sigmoid函数

$$\sigma_z = \frac{1}{1+e^{-z}}$$

- 计算损失函数

$$l(\theta) = \sum_{i=1}^{n}[y_i(\theta^T x_i) - log(+e^{\theta^T x_i})$$

- 实现梯度下降方法

  计算梯度，定义学习速率v

$$\frac{\partial l}{\partial \theta} = -\sum_{i=1}^{n}(x_i(y_i - P(y_i = 1|x_i;\beta)))$$

$$\theta_{new} = \theta_{old} - v\frac{\partial l}{\partial \theta}$$

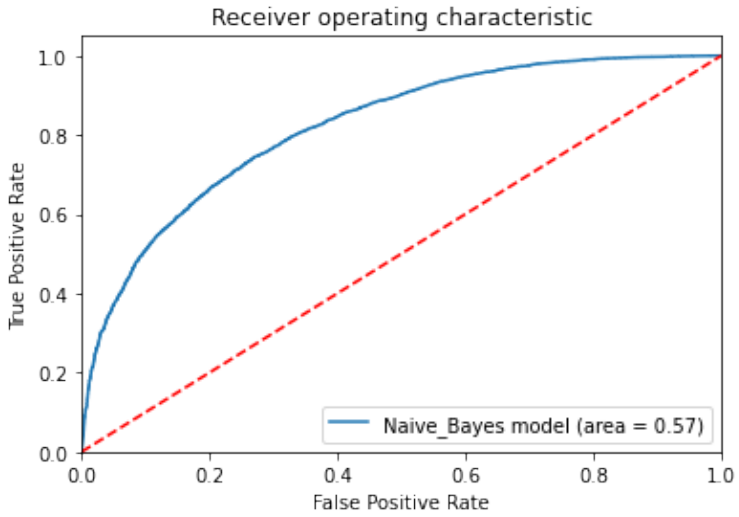
通过以上方法实现的模型可以实现二分类，初步得到的模型在训练集上准确率达到了95.2%，在测试集上准确率达到了83.7%；训练集正确率远高于测试集正确率，因此认为模型过拟合，后续将针对该问题进行优化。

支持向量机

支持向量机（Support Vector Machine, SVM）是一类按监督学习（supervised learning）方式对数据进行二元分类的广义线性分类器（generalized linear classifier），其决策边界是对学习样本求解的最大边距超平面（maximum-margin hyperplane）。本项目采用sklearn库的SVM模型，设置kernel类型为线性，进行SVM模型训练。结果得到SVM模型在测试集上的准确率约为0.818。

朴素贝叶斯

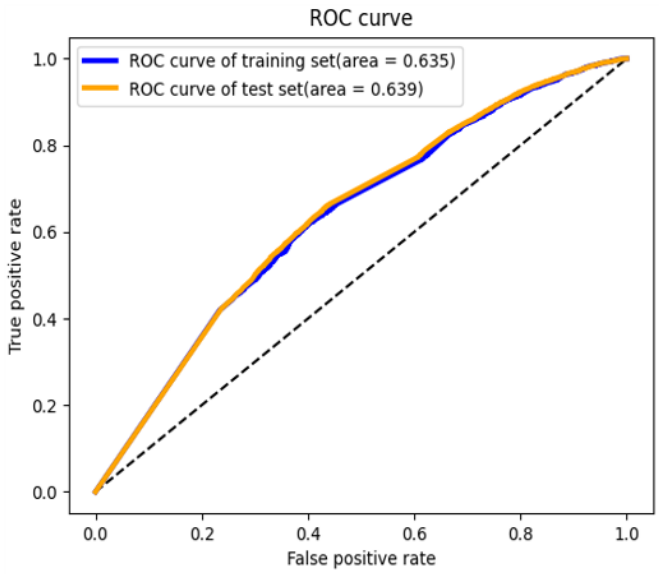
朴素贝叶斯模型是基于贝叶斯定理和特征条件独立假设的分类方法，他所需估计的参数很少，对缺失数据不太敏感，算法比较简单，同时结果又高效准确。经过训练，最终测试集的准确率在0.825，并没有理想得高。究其原因，因为模型设定时的相互独立的假设只在理论中成立，往往无法应用于实际，因此导致分类准确率并不算很高。如下是ROC曲线图。



神经网络

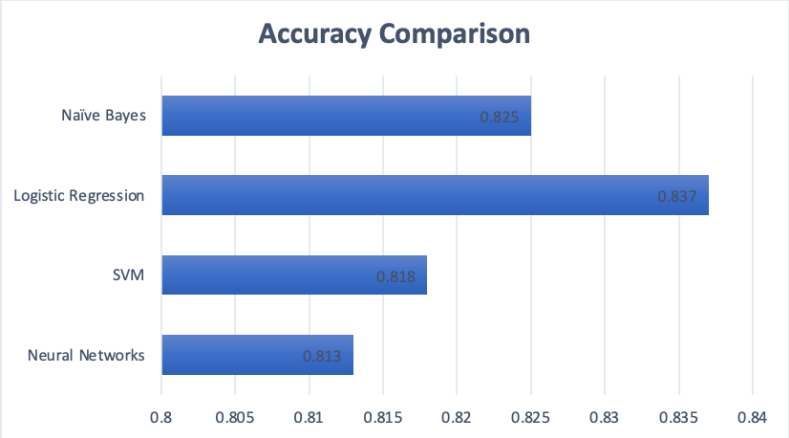
在神经网络模型的学习中，我们采用Keras作为基本框架。Keras是一个高级神经网络API，它能相对快速地处理大批量数据。建立模型时，我们预先设定了一个有300个节点的隐藏层，采用ReLU作为激活函数。在输出层我们采用Softmax函数作为激活函数。训练时，我们用Adam作为优化函数，设定epochs=12，即完成12次前向计算并反向传播更新权重的过程，每次训练的样本数为500个，并选取部分数据作为验证集，以避免过拟合的问题。最终我们的训练集准确率为0.813，测试集准确率为0.816。初步认为模型存在欠拟合的问题。

但是，在进一步优化调整模型时，我们发现，无论如何改变参数或激活函数，例如，设定两个或以上隐藏层、增加每个隐藏层的节点等等，训练集和测试集的准确率并未发生任何改变，并且代码运行一次的时间成本过高，因此，我们最终按最初设定训练模型。如下是它的ROC曲线。



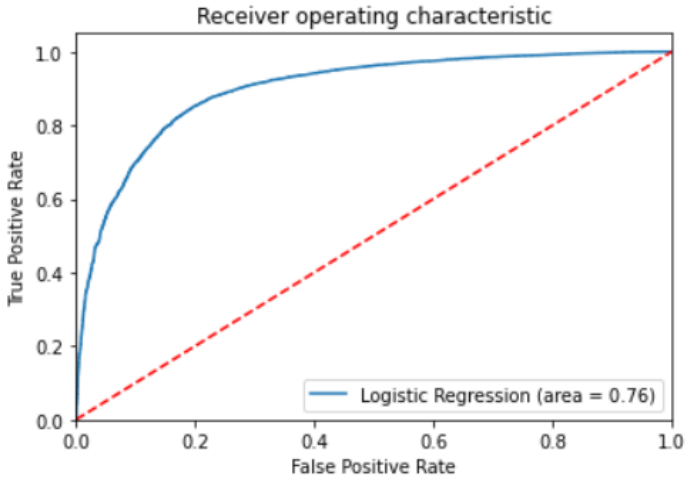
进一步改进思路

下图为四个模型的精度比较图。可以看到，四类模型中，逻辑回归的表现最好，故后期我们主要采用逻辑回归模型进行优化和改进。同时，目前本项目只涉及到二分类问题，但情感的倾向可能是一个更复杂的分类问题，因而进行多分类可能会进一步提升模型精度，提高预测效果。



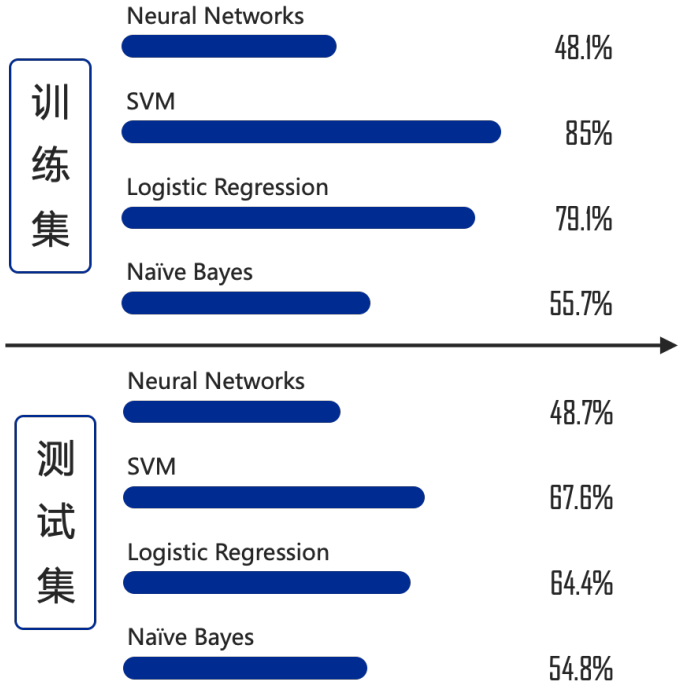
3.2 逻辑回归模型改进

我们发现在初步训练的模型中，有过拟合的问题，因此尝试如下几种方法来改善过拟合问题：正则化、早停等。在L1正则化和L2正则化中，L2正则化效果更好。我们猜测是由于L2正则化可以使趋于0的系数变为0，更适用于稀疏模型。最终我们通过添加L2正则化，并减少最大迭代次数，通过早停来避免过拟合。最终模型在训练集上准确率达到了92%，在测试集上准确率达到了87.5%。效果有部分提升。下面是它的ROC曲线。



3.3 多分类模型训练

为了探究多分类是否能更好地解决该问题，本项目重新定义评级规则，进行五星级打分。将原来的1-2分记为1星，3-4分记为2星，5-6分记为3星，7-8分记为4星，9-10分记为5星。不同模型的精度比较如下图所示。



由图可以看出，SVM模型在四种模型中表现最佳，但总体来看，四个模型的精度较二分类时有不同程度的下降。

4. 结果分析与讨论

4.1 多分类模型精度下降问题

直观上而言，四个模型二分类的准确率远高于五分类的准确率，这是由于问题变得更复杂，对模型的训练难度也会一定程度上增加。但我们不能仅从准确率来评判二分类模型比多分类模型好，毕竟在随机猜测时，五分类的期望准确率20%也远低于二分类的期望准确率50%。

4.2 基于评论的情感分析意义

我们将评论文本的情感表达作为特征，建立模型并学习了其与评级之间的关系，实现了二分类和多分类的预测。文本方式的评价与数字方式的评分往往是有出入的。具体来说，文本评价能透露出患者对药物个性化的主观感受，而较简单明了的数字化评价更具有普适性。通过我们的模型建立，可以更清晰地理解患者对药品打分的情感依据，从而转换为客观的数字评级。在现实很多情况下，例如临床治疗时，患者对药物的评价往往通过语言表达，我们的模型可以通过抓取患者对药品感受的最普适重要的特征进行客观打分，避免医师或家属理解上的偏差，从而有助于提高患者治疗的心理状态。

5. 总结与展望

本项目的主要问题是针对药物评论的自然语言情感倾向研究。主要完成了数据的初步发掘与情感分析，然后利用逻辑回归、支持向量机、朴素贝叶斯、神经网络等模型进行了二分类、多分类的训练，对该问题进行了较为系统的探索。其中添加正则化之后的逻辑回归模型在二分类问题中表现最佳，而多分类当中SVM表现最佳。

为了对本研究问题进行进一步拓展，我们认为未来可以对数据中的其他特征进行训练，例如建立起患者评价与“点赞数”之间的联系，研究通过使用者的主观评价对网友“点赞数”分类的结果；或者在模型中加入多个特征进行训练，探究增加预测准确率的可能性。

在模型构建方面，未来可以进一步优化现有模型，增加其准确率和稳定性，并以此为基础建立一个完善的药品预测系统，以实现通过使用者的主观评价对药品进行评估，同时也能向医师提供一个临床决策的支持工具，进而针对药物的有效性、安全性等进行研究。另外这也对保险公司与药厂制造上有所帮助。

参考文献

Felix Gräßer et al. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125.

附：小组分工表

- 申若曦：数据预处理、数据可视化、SVM模型分析、报告撰写与汇总
- 李润涵：数据预处理、逻辑回归模型分析、报告撰写
- 姜来：数据预处理、神经网络模型分析、报告撰写
- 柯岱霆：数据预处理、朴素贝叶斯机率模型分析（已退课）