Problem Chosen	2023	Team Control Number
C	MCM/ICM	2310767
	Summary Sheet	

揭开字谜:来自世界的证据

摘要：玩家最多可以猜测六次，在世界游戏中预测一个五个字母的单词，每次猜测后都会提供反馈。我们都知道，不同的单词有不同的特征，这些特征可能会影响玩家的表现。在这方面，我们的团队对提供的数据进行了建模和评估，得出了一些有趣的结论。

对于任务 1，在数据预处理之后，我们首先使用 ARIMA 模型来预测报告结果的数量，并发现它只能捕获线性部分。其次，我们采用 LSTM 模型来捕获非线性部分的信息。最后，我们将它们结合起来形成 ARIMA-LSTM 模型，该模型的 RMSE 为 0.0432，预测结果更加准确。我们最终得到 2023 年 3 月 1 日的预测区间为[9614,43109]。随后，我们定义了音节数和熵等五个单词属性，并通过 Spearman 相关系数分析了它们与在困难模式下报告分数的人数百分比的相关性，发现它们之间存在显著相关性。

对于任务 2，我们使用五个首选词属性和竞争数，使用结合线性回归模型(Ridge 回归，Lasso 回归)和树模型(XGBoost, LightGBM)的堆叠模型来预测结果的分布。我们发现，叠加模型将预测结果的拟合优度提高到了 83.77%。此外，MSE、RMSE 和 MAE 都表明堆叠模型具有更好的学习能力。2023 年 3 月 1 日，“EERIE”的预期分布为[1,2,3,4,5,6,X]=[0,0,9,18,26,37,10]。

对于任务 3，我们选择了 7 个额外的单词属性来衡量单词的难度水平，然后通过主成分分析(PCA)缩减指标。然后，我们使用高斯混合模型(GMM)将单词聚类为三大类:困难，中等和容易。为了得到单词的真实难度，我们计算每个单词的期望尝试次数，用于与分类结果进行比较。结果表明，我们的模型准确率达到了 67%。我们从三个不同的角度探索了与每个分类相关的给定单词的属性的有趣发现:熵、字母数和频率。最后，根据“EERIE”的属性将其归为“难”类。

对于任务 4，我们对提供的数据集进行了可视化分析，并在三个方面发现了一些有趣的属性:(1)报告结果的数量和尝试困难模式的玩家的百分比;(2)尝试次数的分布;(3)字母在每个位置出现的频率。这些特点为玩家解决问题提供了一些有趣而可行的思路。

我们还进行了敏感性分析，显示了不同样本对词难度聚类模型的影响。然后总结出我们模型的优缺点。最后，在论文的末尾写一封给《纽约时报》编辑的信，介绍我们论文的总体思路和结果。

关键词:ARIMA-LSTM 模型;相关性分析;叠加;PCA;GMM;世界;

目录

1 介绍

《世界网》是《纽约时报》每日推出的一个谜题，要求玩家在六次尝试中猜出一个五个字母的真实单词，现在有 60 多种语言版本。玩家每次猜中都会收到反馈，具体来说，在提交单词后，贴图会改变颜色来给出反馈。玩家可以在普通模式或困难模式下进行游戏。难度模式要求玩家一旦找到单词中正确的字母，就必须在随后的猜测中使用这些字母，这就增加了游戏的难度。

随着世界网越来越受欢迎，《纽约时报》希望我们对它提供的数据进行分析和建模。为了解决这些问题，我们的团队将采取以下步骤。对 COMAP 官方提供的数据进行异常值处理。建立 ARIMA-LSTM 模型来预测报告结果的数量。构建 12 个单词属性和指标，反映单词的难度。开发可以使用堆叠模型融合算法基于给定单词在未来日期预测报告结果分布的模型。计算均方误差(MSE)、平均绝对误差(MAE)和 Ras2 评估指标，以验证模型的准确性。使用构建的单词难度指标，基于 GMM 聚类对解决方案单词进行分类和识别。描述玩家猜测的次数，单词中出现的 26 个字母的位置，以及我们发现的其他有趣属性。给《纽约时报》的编辑写一封信，包括我们的预测结果和有趣的发现。

我们的建模框架如图 1 所示。

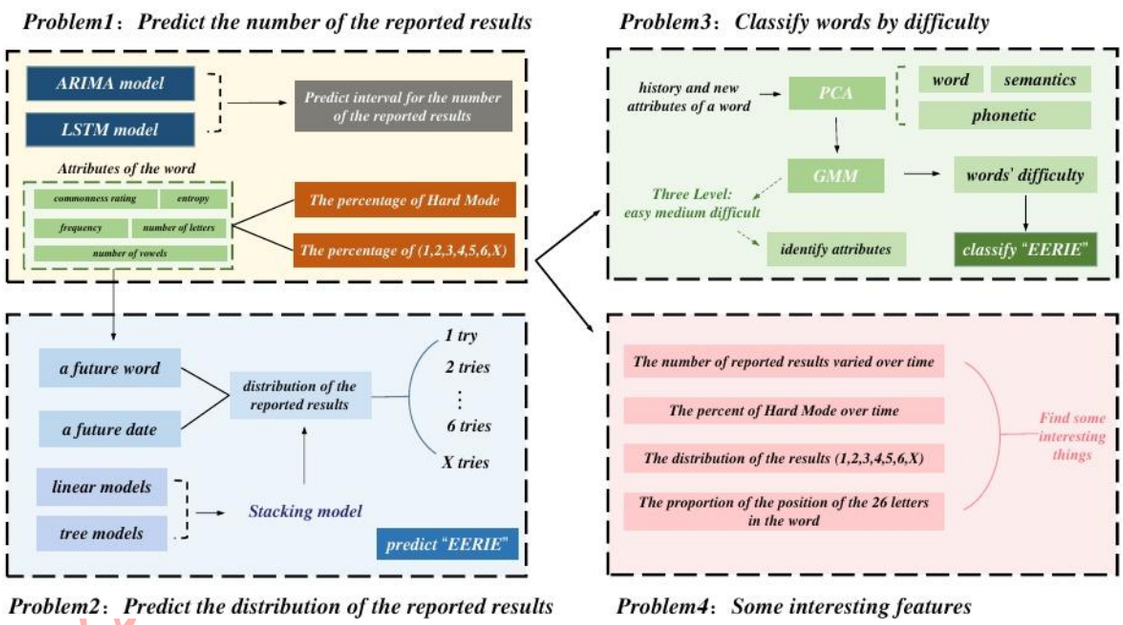


图 1:工作流

2 假设和符号

2.1 模型假设

为了简化我们的问题，我们做了以下基本假设，每个假设都是足够合理的。

假设在 2023 年 1 月 1 日，2023 年 3 月 1 日，不存在会对数据产生实质性改变的意外因素。

由于我们只能使用截止到 2022 年 12 月 31 日的附加数据文件中的数据，所以我们忽略了这个话题之后可能给 worldle 带来的热度或者其他可能导致数据剧烈波动的影响因素。

假设用户在 Twitter 上报告的分数是真实可靠的。

如果用户在 Twitter 上报告他们的分数为假，那么仅使用给定日期的给定单词的属性可能很难预测数据集中报告结果的分布。

假设扰动项遵循独立的正态分布。

2.2 符号和定义

表 1:符号

Symbols	Description
r_t	The number of results reported on day t
$H(X)$	The information entropy size of word X
ρ	Spearman rank correlation coefficient
f_r	Word attribute: a measure of common usage in the coca dataset
f_n	Word attribute: word frequency of SUBTLEX-US corpus
f_{tn}	Word attribute: number of lexical labels
f_d	Word property: consonant doubling
f_p	Word property: phoneme length
f_{vow}	Word property: number of vowels
f_c	Word attribute: specificity rating
λ	GMM model parameters
$p_k(d_i)$	Gaussian component density function
μ_k	Mean vector of the kth Gaussian component

3 数据预处理

由于我们只能使用 COMAP 官方数据集' Problem_C_Data_Wordle.xlsx '，并且给定的数据是通过挖掘 Twitter 获得的，因此存在数据异常的可能性，因此我们在构建模型之前对这部分数据进行了预处理。

Fill:我们将 Number of reporting results 中的异常值替换为前后数据的平均值。

拒绝:我们删除报告结果的分布总和偏离 100%的整个数据。

我们删除包含多个不等于 5 的字母的整个单词，包括“clen”和“tash”。

4 任务 1:区间预测和相关分析

4.1 对报告结果数量的预测

4.1.1 自回归综合移动平均模型

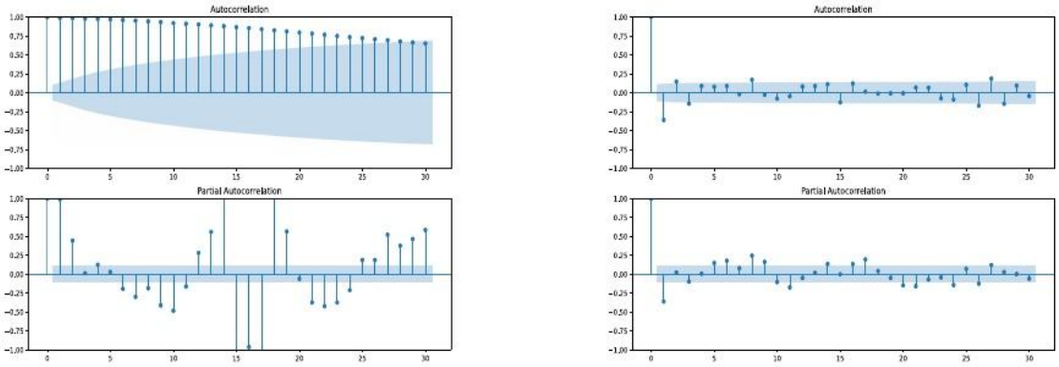
由于数据是时间序列，数据量较小，在考虑了各种预测模型后，我们首先选择了 ARIMA 模型来预测报告结果的数量。

4.1.2 构建预测模型

在 ARIMA (p, d, q)模型中，AR 为自回归项，p 为自回归项;MA 为移动平均，q 为移动平均项的个数，d 为时间序列平稳时产生的差异的个数。该模型是基于通过d阶差分将非平稳级数rt转换为平稳级数rt的原理。然后以rt为因变量，以rt的滞后项和随机误差项在和rt的滞后项作为自变量进行回归。为了便于写作，后者以rt序列表示报告结果的数量。

步骤 1。序列平滑(确定参数 d)

首先，我们对报告结果的时间序列 rt 进行平滑性测试。从下图 2 可以看出，两个时间序列趋势明显，自相关系数衰减相对较慢。同样，我们进行了单位根检验，发现该序列存在单位根。



(a) Sequence r_t (b) Sequence $\bar{r}_t$

图 2:rt 序列的自相关图和部分自相关图

因此是 r_t 一个非平稳序列，需要进一步平滑。

设 $\bar{r}_t = r_t - r_{t-1}$。取一阶差分后，对差分后的 rt 进行平滑性检验，从下图 3 可以观察到，差分后的序列总是在某个值周围随机波动，没有明显的趋势。自相关系数衰减很快，只有间隔很近的序列值才有显著的影响。而且，单位根检验的 p 值都收敛于 0。因此，不存在单位根。Rt 已经是一个平滑的系列。由于我们使用一阶差分法来获得光滑级数，d = 1。

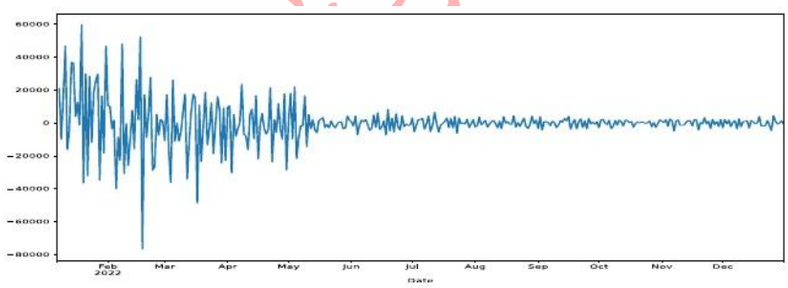


图 3:序列 $\bar{r}$的时间序列图

步骤 2。p 和 q 顺序的确定。

ARIMA(p, d, q)模型的形式为:

$$\bar{r}_t = r_t - r_{t-1} \tag{1}$$

$$\bar{r}_t = \phi_0 + \sum_{i=1}^{p} \phi_i \bar{r}_{t-i} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2}$$

其中εt 为白噪声序列，p 和 q 为非负整数。贝叶斯信息准则(Bayesian information criterion, BIC)通常用于确定模型的最优阶数，该准则是基于似然函数构造的。根据历史数据，通过计算机编程循环计算模型在不同阶次下的 BIC 值，求出使 BIC 最小的阶数 p 和 q，即模型的最优阶数。在确定最优阶数后，我们进行参数估计，并测试发现数据具有统计显著性。

步骤 3。剩余的测试

为了确定模型的有效性，还需要进行残差检验，其中需要对残差序列 t ε进行白噪声验证。如果残差是随机正态分布且不自相关，则残差序列近似于白噪声序列，表明模型拟合良好。我们使用 Ljung-Box 统计量 Q(m)来检验白噪声的接近性:

$$Q(m) = T(T+2)\sum_{l=1}^{m}\frac{\hat{\rho}_l^2}{T-l} \tag{3}$$

当检验的 p 值大于 0.05 时，表示残差序列 t ε在 5%置信水平下通过了检验，即残差序列为白噪声。对本文数据的残差序列进行检验发现，它不是白噪声，这意味着残差中仍有有用的信息，需要对模型进行修正，以提取进一步的信息。

ARIMA 模型对报告结果量的预测缺少一个不够准确的残差项，这在原始数据之外的预测中尤为明显。然而，ARIMA 模型仍然能够很好地捕捉报告结果数量的趋势，这意味着它可以很好地预测报告结果数量的线性部分。

4.1.3 ARIMA-LSTM 模型

由于 ARIMA 模型的预测结果的残差部分没有得到很好的合理预测，因此预测结果的波动性不是很大，因此本文将 ARIMA 模型与 LSTM 神经网络相结合，进一步改进了预测模型。

LSTM (Long - short memory network, LSTM)是一种改进的递归神经网络，它可以解决长期依赖问题，长时间记忆信息实际上是其默认行为。LSTM 在基于 RNN 的隐层中为每个神经单元增加一个记忆单元:信息传输带称为“cell state”，LSTM 使用遗忘门、输入门、输出门等结构来控制时间序列上记忆的信息。通过这种方式，LSTM 可以更深入地挖掘数据之间的潜在模式，使预测更加准确和可靠。

因此，为了弥补 ARIMA 模型的不足，进一步提高预测精度，我们采用线性与非线性相结合的模型进行预测。为此，我们首先根据 ARIMA 预测结果和实际报告数的结果对报告数的残差序列，将其作为 LSTM 神经网络的预期输出;其次，对原始数据进行相空间重构，最终确定最优报告数为 18;第三，将按最优顺序重构的数据作为 LSTM 输入;第四，将训练集输入到 LSTM 神经网络中，对残差序列测试集进行学习建模和预测，得到 ARIMA 残差序列预测值;最后，对 ARIMA 和 LSTM 神经网络模型的预测结果进行汇总，得到报告结果数的最终预测结果。

4.1.4 结果分析

预测结果如图 4 所示。预测值曲线与实际值曲线仍基本吻合，波动趋势保持一致，模型能准确预测拐点，预测结果较好。并且与单模型预测结果相比，从现实生活中报告的结果数量来看，该结果的波动性更大，也更真实。

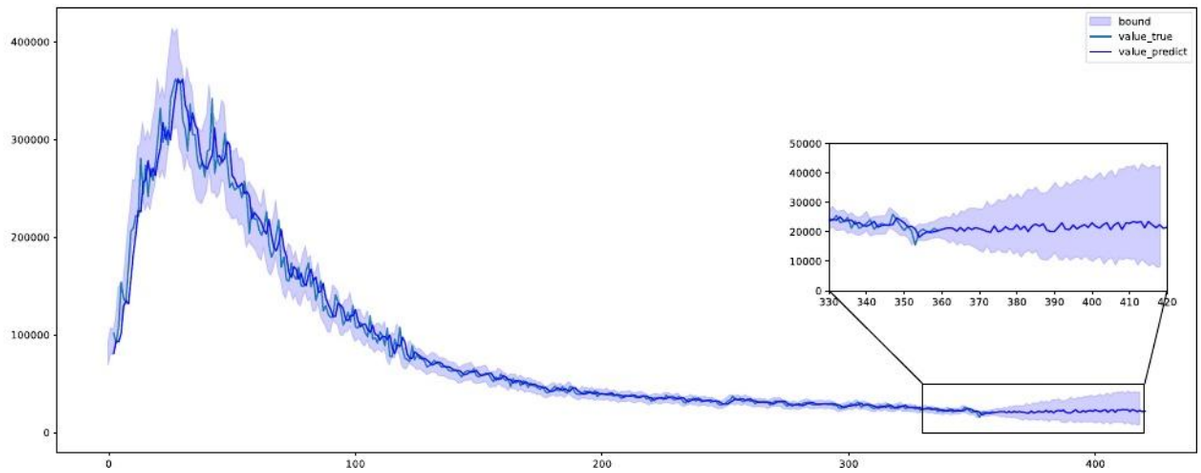


图 4:ARIMA-LSTM 模型-报告预测结果的数量

本节采用 ARIMA 模型预测区间内的报道结局数，并采用 LSTM 模型预测残差序列，以纠正 ARIMA 模型的不足，计算出 2023 年 3 月 1 日报道结局数的最终预测值为 22577，预测区间为[9614,43109]。

4.2 Word Attributes 的构建

为了探索单词的任何属性是否会影响在困难模式下报告分数的人的百分比，我们首先寻找了一些具有代表性的单词的属性。以下是这些属性的定义。

音节数

音节的数量反映了一个单词的组成元素的数量。在本文中，我们选择计数音节的方法来测量单词长度。音节是一个人在一次呼吸中可以发出的最小单位，通常包含一个或多个元音加上一个或多个辅音。

字类

一个词类是一组具有相同形式属性的单词。问题中给出的数据集包含七个主要的词类:名词、动词、形容词、副词、代词、连词和介词。由于代词、连词和介词所占比例很小，我们将它们归类为 other。

词汇使用指数

词汇使用指数。一般来说，日常生活中使用的单词越多，玩家越容易答题，反之，生僻的单词会增加玩家通过游戏的难度。

当代美国英语语料库(COCA)是英语世界中最常用的单词的集合。它是从一个庞大的语料库中提取出来的。使用大数据方法，从各种体裁(1990-2012 年最具代表性的美国报纸、杂志、小说、学术和口语)中自动生成词频列表，这被认为是目前可用的最准确的词频列表。因此，在 COCA 的帮助下，

我们将词汇常用索引定义为

$$f_i = ln(p_i) \tag{4}$$

其中词汇 i 常用索引和 pi 是 i COCA 中的顺序，COCA 按词频从最频繁到最不频繁排序，单词越靠后越不常见。

字母的数量

题目中给出的单词全部由 5 个字母组成，一个单词中有一个以上的重复字母可以降低单词的复杂性，这可能会降低通过游戏的难度;也可能是由于玩家的惯性，这个单词会由 5 个不同的字母组成，而重复的字母反而会增加游戏的难度。

词频

我们考虑词汇在网络上的使用，我们使用谷歌上相关信息的数量。在 Brysbert New 对 x - us 语料库的研究的帮助下，我们获得了数据集中所有单词的频率计数。

信息熵

信息熵可以用来描述源的不确定性。在游戏中，Wordle 会通过改变贴图的颜色来反馈玩家输入的结果。每一种可能的模式将这种模式所传达的信息内容相乘得到一个单词所能带来的预期信息内容的概率也被称为信息熵，信息熵越高也意味着这个单词在各种情况下所带来的信息内容越高。COCA 数据集中 26 个字母在英文文本中的出现频率统计如下图 5 所示。

图 5:COCA 数据集中 26 个字母的频率

在进行猜测时，玩家会尝试从每次猜测中获得最多的信息。因此，将尝试覆盖前两次尝试中频率最高的字母。例如，其他+钉子的组合将覆盖 11 个出现频率最高的字母中的 10 个，运气好的话，一些字母将被识别出来。因此，我们使用信息熵作为一个词的属性来描述这个词所包含信息的大小。单词的不确定性越大，它所包含的信息就越多，信息熵也越大。具体公式如下。

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{5}$$

其中 X 表示作为随机变量的单词，P(X)表示单词的输出概率函数，H(X)表示单词 X 的信息熵大小。

4.3 相关性分析

4.3.1 斯皮尔曼相关系数

我们将上述构建的 5 个单词属性指标与报告的困难模式得分百分比(由 spearman 等级相关系数衡量)应用相关分析方法，思路如下:

变量 x 和 y 的斯皮尔曼相关系数实际上是使用两列数字的秩顺序计算的。通过变量 x = {x1, x2,, xn}按升序或降序获得排序系列=a{a1,a2,....an},每个元素xi在变量x中的位置记为xi,叫做元素 x。通过变量 y = {y1, y1, yn}以同样的方式,我们得到了等级系

列 y = {y1, y2, yn},然后我们得到等级系列 y = {y1, y2, yn}。与变量 y 对应的秩级数 s，秩差级数 d = {d1, d2，…，dn

是将序列 r 与序列 s 中各元素相互对应相减，代入斯皮尔曼秩相关系数公式得到。

$$\rho = 1 - \frac{6\sum_{i=1} d_i^2}{n\left(n^2 - 1\right)} \tag{6}$$

式中:n 为样本数量，对应于排名中涉及的数据总量;ρ为词频变量与字母数变量之间的斯皮尔曼等级相关系数。

4.3.2 相关系数热图

下图描述了上述指标与在困难模式中体验的分数百分比之间的关系。

从图 6 可以看出，大多数指标与选择困难模式的玩家比例并没有明显的相关性，只有元音数与之有很强的正相关关系。

尝试次数少:1 次尝试、2 次尝试、3 次尝试次数多:5 次尝试、6 次尝试、7 次以上尝试(X)

词汇共性指数、元音数和信息熵与尝试次数少的猜测比例呈强负相关，与尝试次数多的猜测比例呈强正相关，与词频呈相反相关。这与我们的猜测是一致的，即一个词在日常生活中越常见，出现的频率越高，尝试次数越少就越容易猜出来。信息量越大的单词，在三次尝试中被猜出的可能性就越小。元音数量较多的单词发音更复杂，在少量猜测后不太可能被游戏清除。

字母数与“几次尝试”的比例呈正相关，与“多次尝试”的比例呈负相关，

可能是由于人类的惯性，重复的字母会让游戏变得更困难，因此在少量的尝试后更难猜出。

单词词性与猜测次数之间的相关性较小，可能是因为玩家在猜测时较少考虑单词词性。

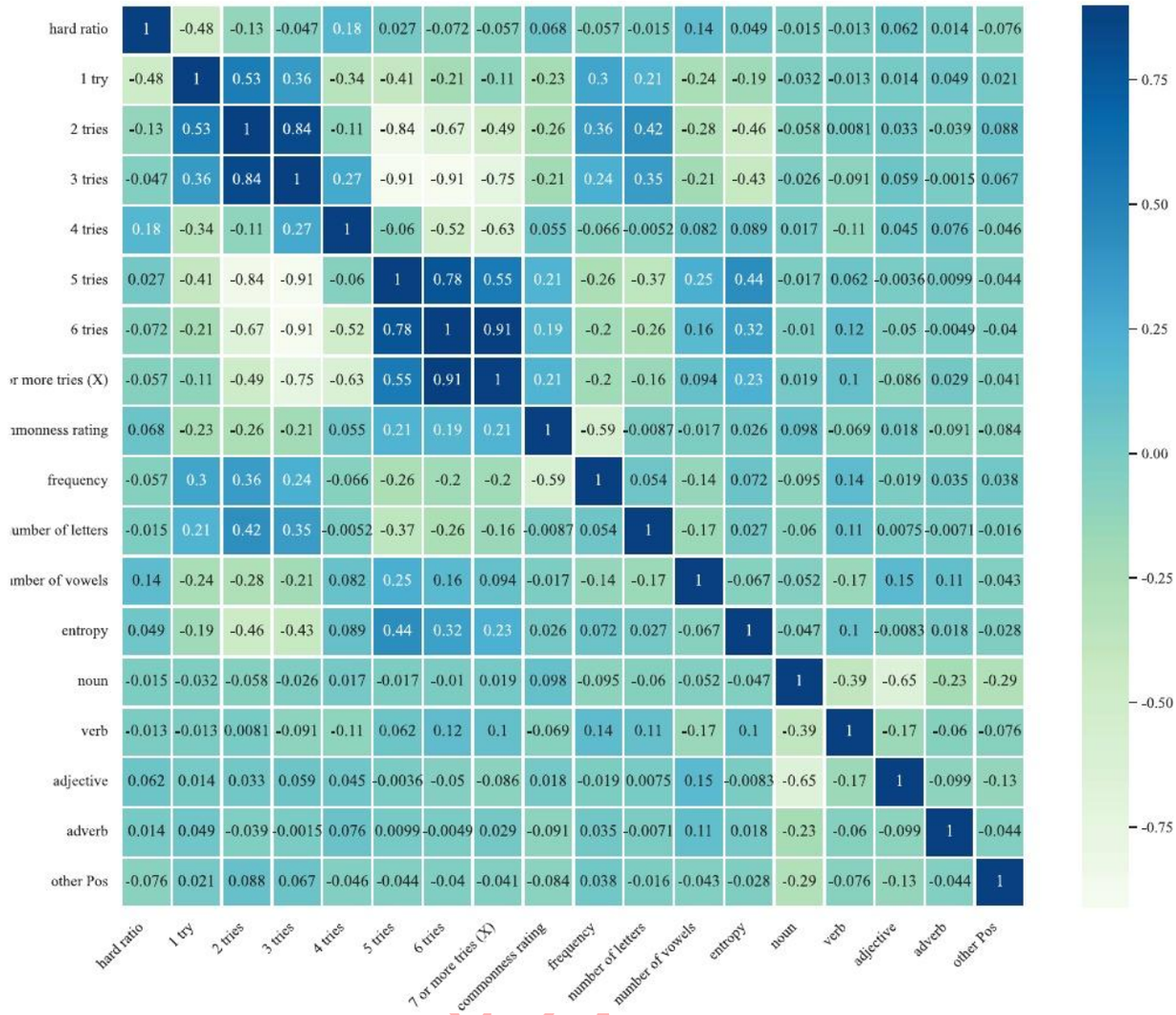


图 6:关联度热图

4.3.3 显著性分析

由于篇幅原因，我们只给出了与在“困难模式”中打出的报告分数百分比显著相关的指标的相关系数和显著性水平。

从上表中我们可以看出，音节数、字母数、常用词索引、词频、信息熵这五个指标与一次尝试、两次尝试、三次尝试、五次尝试、六次尝试猜出或猜不出的谜题比例(X)之间存在显著相关(双尾检验显著)，这与热图分析的结果是一致的。相比之下，四项指标均不存在显著相关

四次尝试的猜测比例。四次尝试的猜测比例可以看作是过渡点，通常这四个指标与少于四次尝试的猜测比例与超过四次尝试的猜测比例之间的相关系数在符号上是相反的。

表 2:各变量的 Spearman 相关系数

x	com.rating	frequency	num. of letters	num. of vowels	entropy
$1try$	-0.231***	0.298***	0.208***	-0.243***	-0.192***
$2tries$	-0.264***	0.355***	0.424***	-0.276***	-0.464***
$3tries$	-0.206***	0.242***	0.347***	-0.207***	-0.428***
$4tries$	0.055	-0.066	-0.005	0.082	0.089
$5tries$	0.208***	-0.261***	-0.368***	0.247***	0.443***
$6tries$	0.189***	-0.202***	-0.263***	0.163***	0.317***
$Xtries$	0.212***	-0.201***	-0.164***	0.094***	0.231***

5 任务 2:基于堆叠算法的结果分布预测

为了预测报告结果在给定未来日期和未来单词的分布，我们训练了 7 个堆叠模型，分别预测每个尝试次数(1、2、3、4、5、6、X)的具体比例，然后将它们对应的预测值归一化，作为最终的分布预测。

基于首选词属性(包括常见度排序、词频排序、字母数、元音数、信息熵排序)和竞争数，分别构建了 Ridge 和 Lasso 线性回归模型和 XGBoost 和 LightGBM 树模型进行对比分析，并采用堆叠方法对初级学习者进行融合，融合后的预测结果更加准确高效。

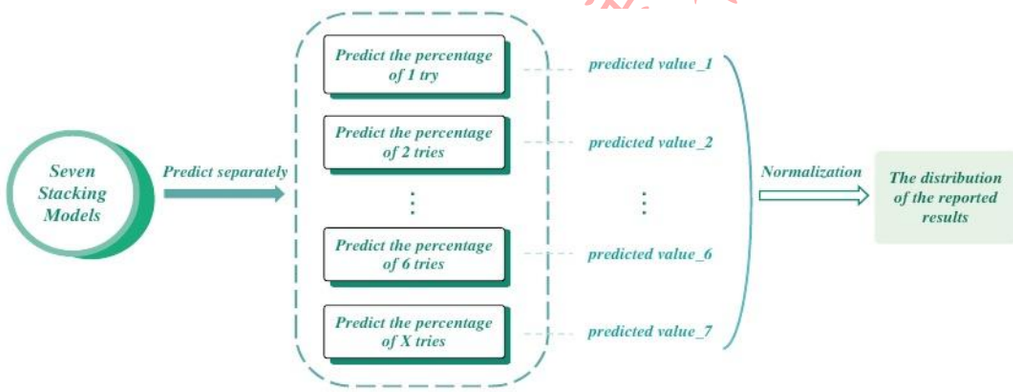


图 7:分布预测流程图

5.1 堆叠法模型融合

堆叠方法是为了降低模型的泛化误差而设计的，本质上是一种“堆叠”的分层结构，具有高预测结果的准确性。本文采用两阶段堆叠法，步骤如下:

层 1。

1)在本文中，我们将问题提供的数据按 4:1 的比例拆分为一个训练集和一个测试集，分别包含 286 和 71 个数据。2)使用 Ridge 回归、Lasso 回归、XGBoost 算法和 Light-GBM 算法对训练集进行训练。3)使用步骤 2 中完成的四个模型对问题提供的数据进行预测，并保存结果。

层 2。

1)将第一层第 3 步的 4 个预测结果数据集作为新的训练集特征，结合课题提供的数据的实际特征，选择第一层单模型评价标准较高的 XGBoost 算法作为二次训练预测的元模型。2)使用训练好的 XGBoost 算法对主题提供的数据进行验证集预测，得到最终结果。

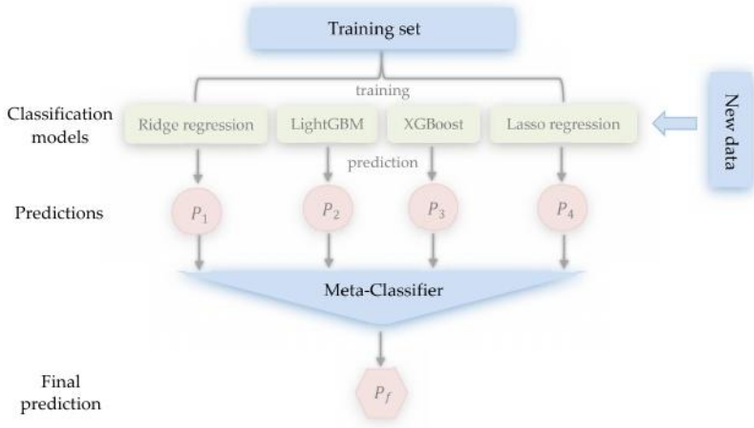


图 8:堆叠模型融合流程图

5.2 预测模型介绍

5.2.1 线性回归模型

Ridge 回归将二阶正则项的最小二乘加入到损失函数中，也称为 L2 参数化，具有降维的作用，同时也限制了模型参数对异常样本的匹配，处理高度相关的数据集，从而提高了模型对大多数正态样本的拟合精度。我们的团队使用 RidgeCV 来调整正则化强度 alpha，从而在 alpha 14 处获得更好的拟合。

Lasso 回归类似于上面提到的 Ridge 回归，因为它也通过构造惩罚函数来处理特征变量的共线性。但与 Ridge 回归相比，Lasso 回归可以通过将惩罚项从 L2 参数改为 L1 参数，将相对不显著的特征变量的系数压缩为零，从而消除变量;而 Ridge 回归只对特征变量的系数进行了一定程度的压缩和保留回归模型的所有变量。

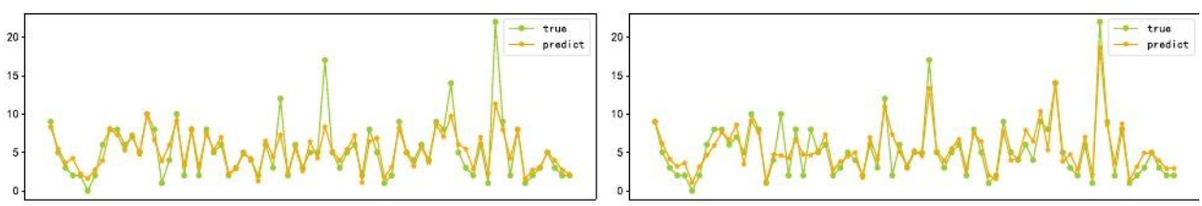
5.2.2 树模型

XGBoost 回归模型的核心思想是计算信息增益，反映各特征变量的信息不确定性降低程度。它是一种优化的分布式梯度增强库，旨在实现高效、灵活和可移植。

LightGBM 模型是一种决策树算法，它遍历数据，根据直方图的离散值找到最优分裂点。与 XGBoost 算法一样，它是 GBDT 的有效实现，与 XGBoost 算法相似的是，它使用损失函数的负梯度作为当前决策树的残差逼近来拟合新的决策树。

5.3 预测结果与分析

本文选取的五个模型的报告结果如图 9 所示。线性模型的报告结果用 Lasso 回归显示，树模型的回归结果用 LightGBM 算法显示。



(a) Lasso regresion result　　(b) LightGBM result

图 9:历年平均评级值(Average Rating Value)

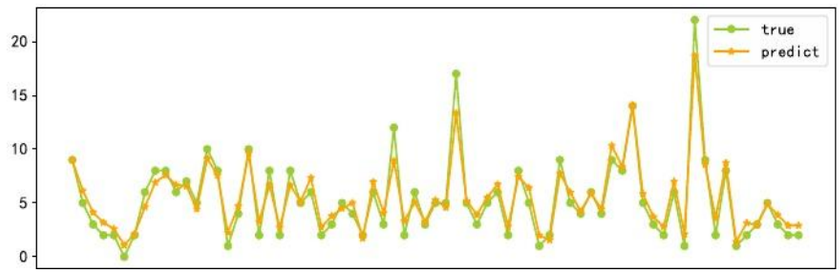


图 10:堆叠模型结果

预测结果表明，线性模型对均值附近数据的预测更准确，而树模型对极值分布的预测更准确，堆叠融合模型保留了两种模型的优点，并相互弥补了缺点。模型预测精度更高。综上所述，堆叠融合模型的预测效果较好。得到 2023 年 3 月 1 日预测的 EERIE 的比值分别为[1,2,3,4,5,6,X] =[0,0,9,18,26,37,10]。

5.4 模型评估

本研究采用交叉验证方法对各模型的准确性进行验证，并以交叉验证结果产生的均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)和 R2 作为估计和验证模型准确性的评价指标。估计模型对应的 R2 越大，MSE 越小，RMSE 和 MAE 表示模型的预测精度越高。下表为 4 个指标下各种模型的模型预测结果:(以“2 次尝试”为因变量)

表 3:Spearman 对各变量的相关系数

	MSE	RMSE	MAE	R_2
Lasso	2.3771	1.5417	1.0727	0.6847
Ridge	2.1292	1.4592	0.9909	0.7151
XGBoost	1.9289	1.3888	0.9186	0.7856
LightGBM	1.8562	1.3642	0.9045	0.8062
Stacking	1.6768	1.2949	0.8623	0.8377

与上述单一模型相比，本文发现堆叠模型的 R2 有所提高，MSE、RMSE 和 MAE 均有不同程度的降低。然而，由于玩家行为的可变性和不可预测性，以及意外事件或外部因素影响玩家参与和表现的可能性，即使是堆叠融合模型在预测报告结果的分布方面仍然存在不足。

6 任务 3:基于高斯混合的分类模型

为了解决第三个问题，我们首先选择了 7 个英语单词属性和特征来衡量单词的难易程度。对数据集中的英语单词进行可视化分析和 GMM 聚类，通过主成分分析(PCA)对指标进行降维。最后，将分类结果与单词的真实难易度进行比较，结果表明，真实分类结果与我们的 GMM 聚类结果高度重合，证实了本文选择的单词特征与单词难易度有密切的关系。

6.1 词属性和难易度指数的建立

维度 1:单词

常用度是用来衡量一个词在日常生活中使用的频率。该指标是通过对该词在 coca 数据集中的常用词的排名顺序取对数来计算的，数值越高，表示排名越低，即不太常见。词频(fn)我们考虑了单词使用的频率，并借助布赖斯伯特 · 纽对“微妙的美国”语料库的研究

获得了数据集中所有单词的频率计数。我们认为一个单词是由几个字母组成的，出现多次的字母不被重复计数。

维度 2:音系学

音素长度(fp)许多英语单词包含无重音辅音字母，因此使它们更难以记忆。因此，我们收集了每个单词的音素长度来表示可能增加单词记忆难度的不发音字母的数量。通过应用 NLTK 的 CMU 发音字典数据，我们获得了所有单词的语音。

元音数(fvow) Beinborn 等人利用元音与辅音的比例发现，元音比例非常高和非常低的单词比元音比例适中的单词更容易引起拼写错误，更难拼写。我们直接引入元音数量作为第二个音系特征。

维度 3:语义学

具体等级(fc)具体等级衡量一个词所代表的概念与一个可感知实体的关联程度。根据 Paivio 的双编码理论，具体的单词比抽象的单词更容易记忆，因为它们除了激活语言编码外，还激活了感知记忆编码，从而使具体的单词相对不那么难记。我们参考 Brysbaert 的研究结果，给出了问题给出的数据集中所有单词的具体等级。

词汇标签数(ftn)一个词可能有几个词汇属性，这意味着它可以在各种上下文中使用，使用得越广泛，难度就越小。NLTK，即自然语言工具包，是我们用来标记词汇属性的工具。

6.2 高斯混合物的分类模型

为了对解词进行分类，我们建立了一个基于高斯混合模型(GMM)的词分类模型。

高斯混合模型(GMM)是一种广义概率模型。一般情况下，只要高斯数足够大，就可以有效地对多维向量的连续概率分布进行建模，因此适用于表征词的语义分布，从而对词进行分类。

高斯混合模型是一系列高斯分布的加权组合。由 M 个高斯分量组成的高斯混合密度函数是 M 个高斯密度函数的线性加权和。

$$p\left(d_i \mid \lambda\right)=\sum_{k=1}^{M} w_k p_k\left(d_i\right) \tag{7}$$

上式中，λ为 GMM 模型参数，pk(di)为高斯分量密度函数，ω_k 为各高斯分量的权重。

$$p_k(d_i)=\frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_k\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(d_i-\mu_k\right)^{\mathrm{T}} \Sigma_k^{-1}\left(d_i-\mu_k\right)\right\} \tag{8}$$

这里μk 是第 k 个高斯分量的平均向量，Σk 是对应的协方差矩阵，D 是特征向量的维数。因此，GMM 模型可以用以下一组参数表示:

$$\lambda=\left\{w_k, \mu_k, \Sigma_k\right\}, \quad k=1,2, \cdots, M \tag{9}$$

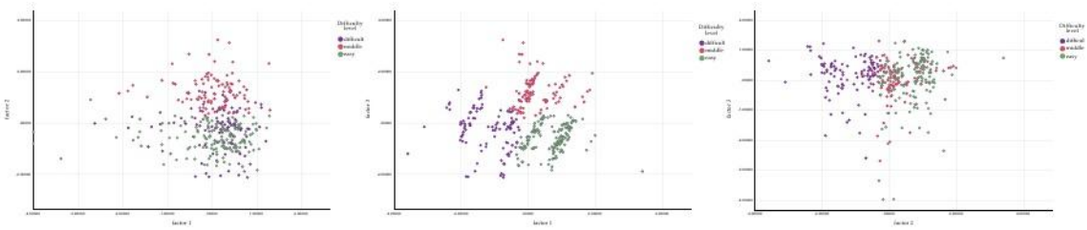
利用 GMM 根据词的特征分布对词进行分类有两个出发点:1)GMM 的高斯分量可以描述一定词向量的分布;2)线性加权高斯密度函数可以近似任意形状的概率分布，而我们对数据的概率分布是不确定的，因此选择 GMM 对单词进行分类。

6.3 单个词特征的可视化分析

对于上一节介绍的特征，我们首先对辅音加倍 fd、标签长度 ftn、语音 fp、特异性评级 fc 等特征进行幂运算，以获得更好的表示。类似地，我们对频率 fn 的原始数据取对数。此外，对于单词到向量的特征，我们使用 Word2Vec 的 Gensim 实现将所有单词转换为 100 维向量作为 fvec。因此，我们将所有 7 个特征组合成一个 10个7 维的特征向量 (100维从词到向量特征， ＋7 个特征，fr…fc)。

$$Difficultyfeatures = (f_r, f_n, f_tn, f_d, f_p, f_vow, f_c, f_vec) \tag{10}$$

为了直观的分析，我们首先对特征进行缩小，以可视化所有特征与难度之间的关系。因此，我们使用主成分分析(PCA)将图 12 中的 f 个特征的维数从 107 降至 3 维。



((a)) factor1 and factor2　　((b)) factor1 and factor3　　((c)) factor2 and factor3

图 11:聚类结果的二维图

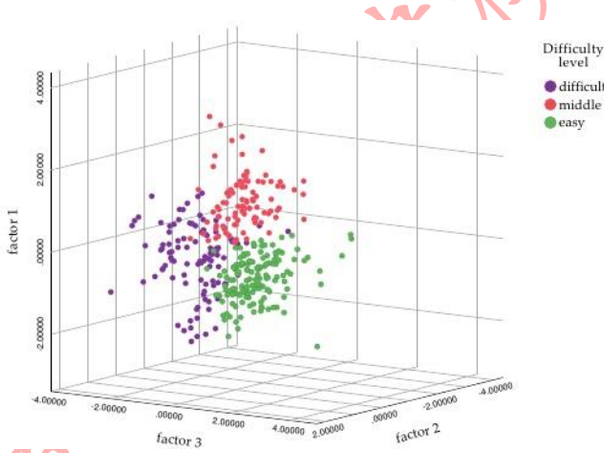


图 12:聚类结果的三维图

在这张图中，三个主轴代表了我们通过因子分析得到的三个共同因素。不同的颜色表示不同的难度等级。有趣的是，“木乃伊”这个在日常生活中经常出现的单词被归入了难度类别，我们认为这是因为存在三个相同的字母“m”(很难想到)，使得这个单词的难度大大增加。虽然特征的前三个主成分只保留了大约 72.8%的方差，但我们很容易发现，几乎所有的单词都是根据它们的难度级别进行分组的。这证实了我们所选择的特征确实与难度等级密切相关。

6.4 高斯混合模型的分类效果分析

为了进一步对数据进行分类，我们使用 GMM 对单词进行聚类。我们调用 Python scikit-learn 包用于 GMM 训练所有数据，并通过均值和协方差值绘制数据的协方差。结果表明，协方差覆盖了其对应聚类中的大部分数据。难词、中词和易词的 GMM 聚类的总字数分别为 105、100 和 150。

为了更详细地探索聚类效果，我们根据击中单词所需的预期猜测次数定义了每个单词的难度，并将其与分类结果进行比较，单词的真实难度定义如下。

$$E(diff_i) = \sum_{j=1}^{n} j * p_j \tag{11}$$

其中 E(dif，f_i)是单词的预期难度，j 是在 j 次尝试后猜出该单词的次数，如果没有猜出，我们将 j 的值设置为 7;pj 为 j 次尝试猜出的比例，同样 p7 为谜题无法猜出的比例。最后，我们将预期难度按降序排序，采用难度的三分法并将其分为三类:困难、中等和容易，以便与我们的预测结果进行比较。

在图 16 最左边的三个条形图中，我们可以发现，难度等级为 1 的单词数量在这个聚类中是最高的。据此，我们将其命名为“预测难度等级 1”，对应于图右下方的绿色区域

上图。图 2 中最右边的一组条形图，难度等级 3 的字数最高，因此对应上图左下方的紫色区域，即“预测难度等级 3”。对于中间的一组条形图，由于这个聚类与其他两个聚类重叠，所以预测效果不如难度等级和易易等级，对于中等难度等级的单词聚类，即与难度标签具有相同预测等级的单词数量占每个聚类中单词总数的百分比，我们只得到了 60%的准确率。中间的条形图对应于上图顶部的红色区域 13。

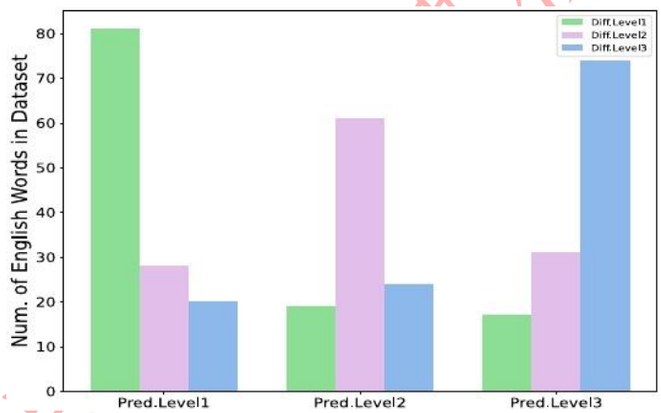


图 13:不同难度等级的预测聚类分布

6.5 识别与分类结果相关的词属性

为了识别与每个类别相关联的给定词的属性，我们探索了信息熵、字母数和词频等词属性在不同预测难度水平下是否存在差异。

信息熵视角

如图 14 所示，难度等级为 1 的单词所包含的信息熵比难度等级更高的单词所包含的信息熵要小得多，即单词的难度等级越高，信息熵越大，包含的信息也越多。这一结论与我们之前的分析结果是一致的。信息熵越高，也意味着这个词在各种情况下带来的信息含量越高，因此这个词的难度也就越高。

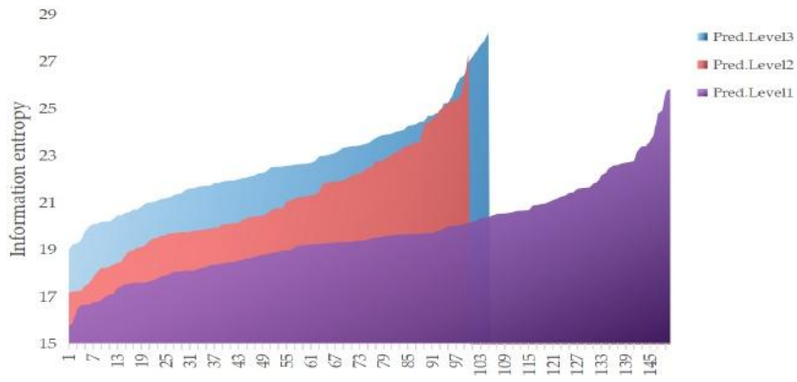


图 14:不同难度等级的信息熵

字母数透视

从图 15 可以看出，难度 1 级和难度 2 级包含了 5 个不重复字母的单词，而难度 3 级包含了 26%的两个重复字母的单词，甚至有 2%的三个重复字母的单词，这说明单词中重复字母的存在确实对单词的难度有越来越大的影响。

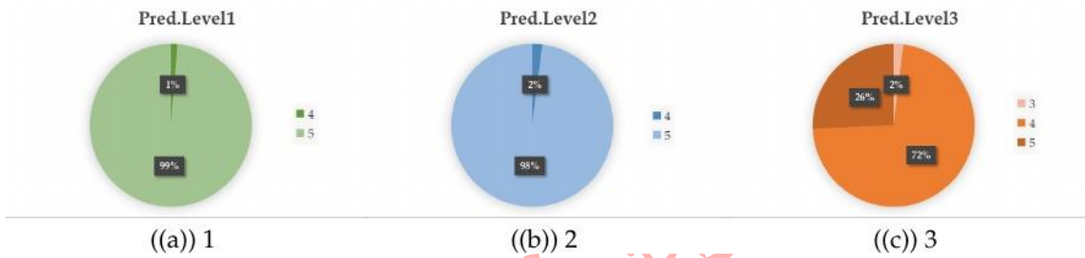


图 15:不同难度等级的单词字母数量

词频视角

图 16 显示，难度等级 1 的单词在数据库中出现的频率更高，而难度等级 3 的单词在数据库中大多出现在 0 到 7000 之间，这说明在日常生活中出现频率越高的单词越容易被识别，反之则越难。这与我们的预测结果相一致。

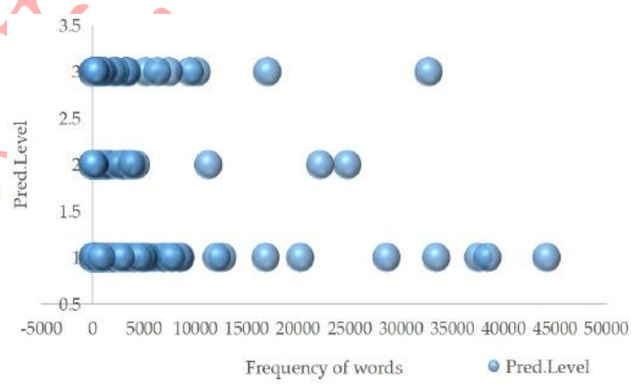


图 16:不同难度等级的词频

7 任务 4:有趣的特征

从图 17 中我们发现，2022 年 1 月和 2 月上报结果数量呈现快速增长趋势，从 3 月开始，世界热缓慢下降，上报结果数量呈现衰退型下降趋势。2022 年 8 月以后，上报结果数量变化不大，呈现波动式下降。愿意尝试高难度模式的玩家数量随着时间的推移趋于增加，出现了三次较大的波动，说明玩家行为的不可预测性和不可预测性，以及意外事件或外部因素对玩家参与游戏的不可预测影响。

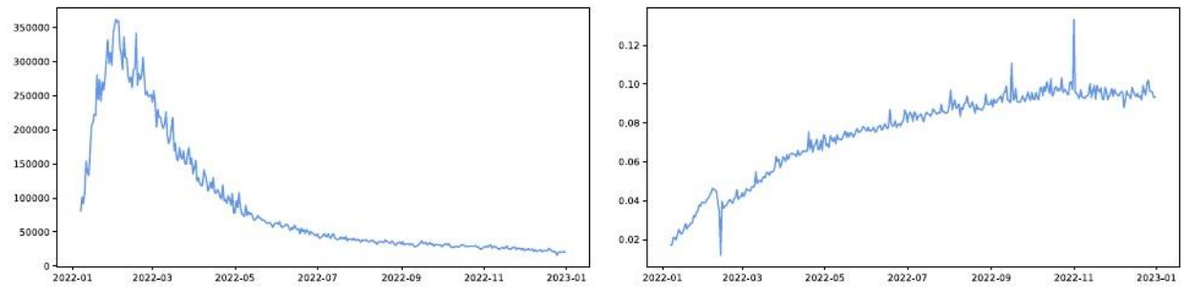


图 17:玩家信息时间序列趋势

我们绘制了 1-6 次猜出的百分比或猜不出的百分比(X)的折线图，发现 4 次猜出的百分比通常高于其他百分比，其次是 3 次猜出的百分比，而一次猜出的百分比几乎等于 0(一次猜出 0 次的结果占总结果的 61.21%)。并且有少数人能够在第二次尝试中猜出答案(第二次猜测的平均百分比为 5.84%)。此外，解不出谜题(X)的人所占的比例波动更大，平均值为 2.80%，但最大值(单词“parer”)为 48%，说明数据中存在高于大多数人能力水平的单词，这使得《worlddle》游戏更具挑战性。

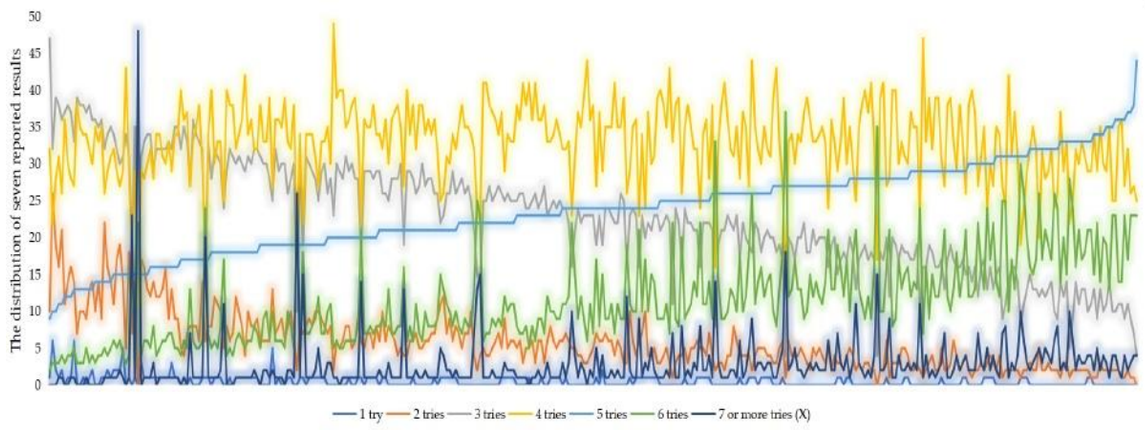


图 18:1-6 次尝试后猜对或猜不出谜题的比例(X)(按 5 次尝试猜对的比例升序排列)

我们探索了这 26 个字母在数据集中单词的位置上出现的比例，结果显示，b、f、j、q、s 这五个字母作为单词的首字母出现的频率更高，而 d、l、t、y 出现的频率更高，而 q 只出现在单词的开头和结尾。S 在第一个字母的位置出现的频率最高，d 在最后一个字母出现的频率最高，中间三个字母是 n、o、e 出现的频率最高。这可以为玩家提供一些有趣和可行的想法，让他们猜测那些放在单词中但放错位置的单词和字母。

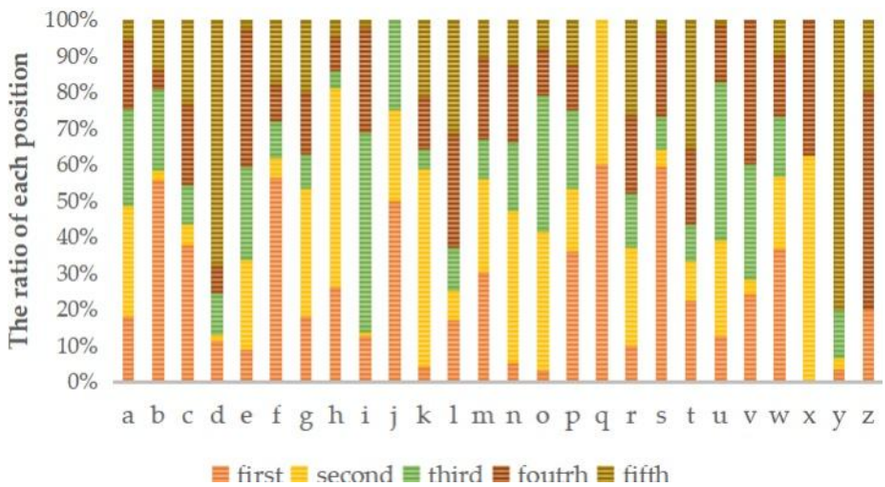


图 19:26 个字母出现在单词中的比例

8 灵敏度分析

为了检验我们的 GMM 分类模型的合理性和泛化性，我们扩展了训练样本来观察分类准确率的变化。本文的训练样本来自一个单词集 W，其中有 11998 个英语单词，这些单词被划分为简单、中等和困难三个难度级别。每个难度级别大约包含 4000 个单词。因此，我们为字母数字 4、5、6、7 和 8 的单词选择了不同的训练集。每个训练集中的单词数在 400 个上下波动，与问题提供的数据量一致。因此，对于不同单词的不同训练集，我们总共得到了 25 个准确率。我们将聚类模型计算出的难度标签与单词集 w 的难度标签进行比较，根据上面描述的方法计算准确率。

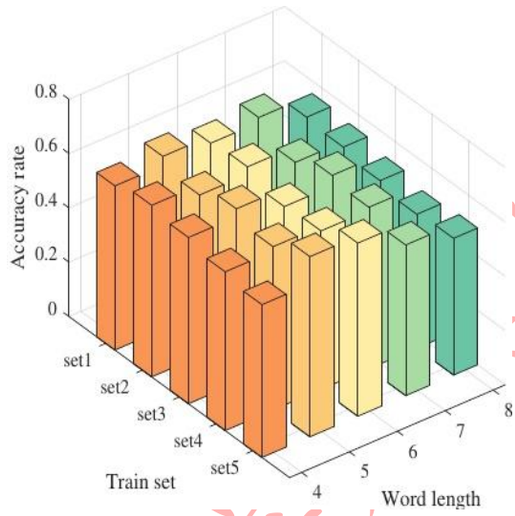


图 20:包含不同字母数的单词的聚类准确率

从图 20 可以看出，由于本文的聚类模型是在包含 5 个字母单词的数据上进行训练的，所以包含 5 个字母单词的 5 个训练集五个字母显示出更好的分类结果，其次是包含四个或六个字母的单词的训练集。分类准确率随着单词字母数量的增加而下降，特别是对于包含 8 个字母的单词的训练集 5，准确率仅为 49.31%。而对于字母数在 5 左右的单词，准确率都在 60%上下波动，说明我们的 GMM 分类模型具有鲁棒性，适合对包含 5 个字母的单词进行分类。

9 模型的优缺点

9.1 优势

通过将 ARIMA 模型与 LSTM 模型相结合，将线性和非线性情况都考虑在内，从而既关注报告结果数量的趋势变化，也关注波动性变化。

ARIMA-LSTM 模型可以简单地利用报告结果数量本身的历史数据来预测其未来趋势，并给出预测区间的大小，这比传统的数学和统计方法更符合现实生活的波动。

基于堆叠模型的融合算法将多个模型结合起来预测结果的分布，可以充分发挥每个模型的优势，具有更强的学习能力，得到更好的预测结果，为建模提供了新的思路。

根据灵敏度分析的结果，我们建立的 GMM 聚类模型对 4-6 个字母的单词的难易度具有优异的分类和识别能力，该模型适用对象范围广，准确率高。

9.2 缺点

计算指标较多，繁琐，模型运行速度有待提高。

由于时间限制，我们对单词属性和难度指标的描述可能不够全面，未来可以使用更多的单词属性和更大的英语单词难度数据集进行训练

References

[1] Brysbaert, M., Warriner, A. B., Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods,46(3), 904－911.

[2] Paivio, A. (2013). Dual Coding Theory, Word Abstractness, and Emotion: A Critical Review of Kousta et al. (2011). Journal of Experimental Psychology: General,142, 282－287.

[3] Lisa Beinborn, Torsten Zesch, and Iryna Gurevych, “Predicting the Spelling Difficulty of Words for Language Learners,” Proceedings of the 11th Workshop onInnovative Use of NLP for Building Educational Applications, pp. 73－83, 2016.

[4] Brysbaert, M., New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977－90.

[5] A survey of trust and reputation systems for online service provision[J]. DecisionSupport Systems . 2005 (2).

[6] Hunter M. Breland, “Word Frequency and Word Difficulty: A Comparison ofCounts in Four Corpora.” Psychological Science, vol. 7(2), pp. 96－99, 1996.

[7] Edward Loper, and S. Bird, “NLTK: the Natural Language Toolkit,” ETMTNLP’02 Proceedings of the ACL-02 Workshop on Effective tools and methodologiesfor teaching natural language processing and computational linguistics, vol. 1,pp.:63－70, 2002.

[8] gensim models.word2bec-Word2vec embeddings.

[9] scikit-learn Machine Learning in Python, url: http://scikit-learn.org/ stable/.

[10] Douglas Reynolds, “Gaussian mixture models,” Encyclopedia of biometrics, pp.827-832, 2015.

亲爱的谜题编辑，

感谢您为我们提供了有趣的谜题世界的日常结果文件，以便我们对世界的结果进行分析和建模，这使我们能够发现关于给定数据集的一些有趣的结论。应贵刊的要求，我们很高兴有机会向您展示我们的发现和结论，希望您会感兴趣。

区间预测结果及相关分析

我们通过构建 ARIMA-LSTM 模型预测了报道结果的数量，该模型得出 2023 年 3 月 1 日的预测区间为[9614,43109]。随后，我们定义了音节数和词类等 5 个词属性，通过 Spearman 相关系数分析发现，这 5 个词属性之间存在显著相关。其中，词汇共性指数、元音数、信息熵与 1 次、2 次、3 次猜对的比例呈强负相关，与 5 次、6 次猜对的比例或解不出谜题(X)呈强正相关;词频和字母丰富度则与之相反。

结果分布&EERIE 分布预测

之后，我们开发了堆叠模型融合算法来预测结果的分布，得到 2023 年 3 月 1 日 EERIE 的预测分布为[1,2,3,4,5,6,XJ-[0,0,9,18,26,37,10]。我们还使用 R2 等作为评估指标来估计模型的准确性并对其进行验证，并证明了我们的模型的性能是最优的，拟合优度为 0.8377。

分类和 EERIE 难度预测

此外，我们根据单词的难易程度将其分为难、中、易三类，并根据“EERIE”的属性值将其归为“难”类。从信息熵、字母丰富度和词频三个不同的角度，我们发现难度等级 1 的单词信息熵远小于更高难度等级的单词，并且在数据库中出现的频率更高，难度等级 3 的单词包含更多的重复字母。一些有趣的特征

最后，我们对数据的可视化分析显示:大多数人在尝试 3、4 次后猜出了这个谜题;能够一口气猜出谜题的人非常少;而猜不出谜题(X)的人的比例波动更大，高达 48%。在数据集中的单词中，b、f、j、q 和 s 作为单词的第一个字母出现的频率更高，而 q 只出现在单词的开头和结尾，而 s、n、o、e 和 d 是出现频率最高的字母

我们很高兴有机会预测世界的结果，并分析数据集中有趣的发现。有关文章的更多信息，请随时与我们联系。

您诚挚的 MCM 2023 团队。