基于 Twitter 评论的美国大选预测

1.数据获取

目标：通过爬取与美国大选相关的 Twitter 评论数据（包括推文内容、评论等），为后续分析提供基础数据。

1.1 数据爬取

获得的数据如下图所示，每一个文件包含 5W 行的数据，一共 20 个文件，共 100W 行的数据。如图 1

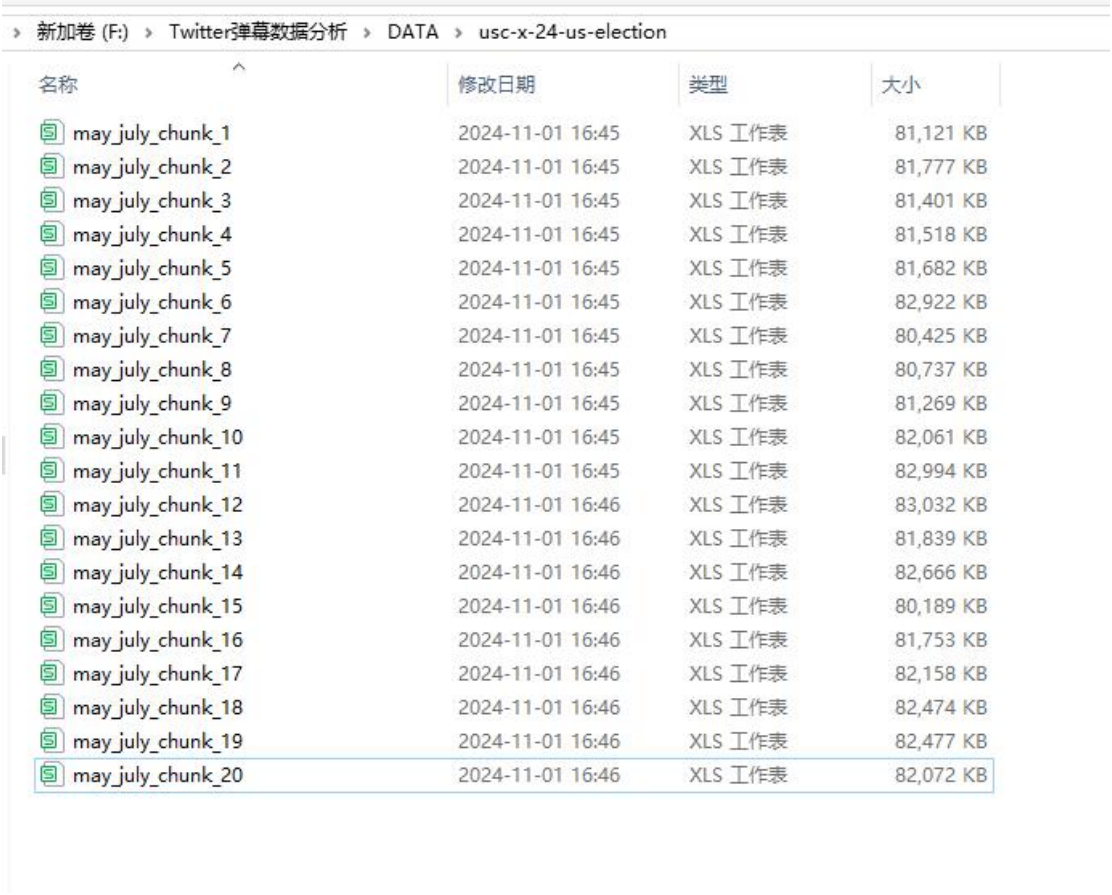

图 1

1.2 数据合并

通过两层 for 循环，合并数据，此处采用分块读取和即时合并的方式，内存中始终只保留了当前正在处理的数据块以及已经合并好的部分结果数据，而不是一次性容纳所有文件的全部数据。这样在处理大数据集时，即使计算机的内存资源相对有限，也能够较为顺利地完成数据的合并任务。如图 2

```python
import pandas as pd

# 创建一个空的 DataFrame 用来存储合并的结果
merged_df = pd.DataFrame()

# 循环遍历每个文件，读取并合并
for i in range(1, 21):  # 文件名从 1 到 20
    file_name = f'{i}.csv'
    df = pd.read_csv(file_name)

    # 如果是第二个及以后文件，去掉第一行（列名）
    if i > 1:
        df = df.iloc[1:]

    # 将当前文件的数据合并到最终的DataFrame中
    merged_df = pd.concat(objs: [merged_df, df], ignore_index=True)

# 保存最终的合并结果
merged_df.to_csv( path_or_buf: 'all_merged_file.csv', index=False)
```

图 2

2.数据预处理：

2.1 保留字段说明：

我们从 30 个数据字段（列名）中选择保留以下字段：

rawContent：这个字段包含了推文的内容，它是我们文本分析和关键词提取的基础。清理后的文本内容将用于后续的情感分析和关键词分析。

replyCount：这是推文的回复数，可以用来衡量推文的互动性或受欢迎程度。对分析推文热度、用户参与度等方面非常有用。

retweetCount：这个字段表示推文被转发的次数。转发次数通常与推文的传播度、影响力和传播效应相关，对于分析推文的传播范围和影响力非常重要。

likeCount：表示推文的点赞数，通常可以反映推文受欢迎的程度和观众的支持。

quoteCount：推文被引用的次数，引用与转发类似，也是衡量推文影响力的一个指标，尤其对于分析特定话题或人物的讨论量和舆论趋势非常重要。

这些字段将帮助我们了解推文的互动性、传播力以及受欢迎程度，进而为情感分析、趋势分析等提供数据支持。

2.2 文本清理：

文本清理是为了保证我们后续进行文本分析（如情感分析、关键词提取等）时，数据是干净的。我们主要进行以下几项清理操作：

去除 URL：URL 地址无论在文本分析中都不是很有用，因此我们用正则表达式 http\S+|www\S+ 来去除文本中的 URL。

去除多余的空格：一些文本可能包含多余的空格或换行符，这会影响分析的准确性。我们将去除这些空格，保证文本清晰。

去除非字母数字字符：我们使用 r'[^\w\s.,!?;:]' 这样的正则表达式来去除不必要的特殊字符和表情符号，只保留字母、数字、常见标点符号和空格。这样能确保文本内容的有用部

分被保留，同时不至于丢失过多信息。

http\S+|www\S+：去除以 http 或 www 开头的 URL。

[^\w\s.,!?;:]：去除除了字母、数字、空格、和一些常见标点符号以外的字符。\w 代表字母、数字和下划线，\s 代表空格，这样就能保留大部分有用的字符。

文本清理的代码如图 3：

```
# 清理文本数据
def clean_text(text):  1 usage
    # 如果文本是缺失值或非字符串类型，则返回空字符串
    if not isinstance(text, str):
        return ''
    # 去除URL
    text = re.sub( pattern: r'http\S+|www\S+',  repl: '', text)
    # 允许标点符号和中文字符
    text = re.sub( pattern: r'[^\w\s.,!?;:]',  repl: '', text)  # 保留字母、数字、空格、常见标点符号
    # 如果文本为空，返回空字符串
    return text.strip()
df['cleaned_content'] = df['rawContent'].apply(lambda x: clean_text(x))
# 删除包含缺失值的行
df.dropna(inplace=True)

# 输出处理后的数据（可选）
df.to_csv('cleaned_data.csv', index=False)
```

图 3

2.3 处理缺失数据：

缺失数据的处理根据数据的性质和分析目标决定。如果某些字段的缺失不会影响分析目标，我们可以直接删除这些行。如果某些字段在后续分析中很重要且可以合理填充，则可以选择填充缺失值。

以下为数据预处理后的结果，如图 4

	A	B	C	D	E	F	G
1	rawContent	cleaned_content	replyCour	retweetCc	likeCount	quoteCount	
2	@lukepbeasley I cant imagine anyone actually feels this way. As mu	lukepbeasley I cant imagine anyone a	0	0	1	0	
3	Voters can also sway me away from voting for someone. @Potomacbea	Voters can also sway me away from vc	1	0	0	0	
4	@PoodleHead57 @BobOnderMO Can you name that amount of charges brou	PoodleHead57 BobOnderMO Can you name	1	0	0	0	
5	@Morning_Joe @JoeNBC The fact remains that Joe Biden is simply a te	Morning_Joe JoeNBC The fact remains	0	0	0	0	
6	@BidenHQ That's funny you're obviously trying to coverup for Biden'	BidenHQ Thats funny youre obviously	0	0	0	0	
7	#Internacional Las modificaciones introducidas por el grupo palestino HamÃ¡s en el acuerdo diseÃ±ado por el presidente de EEUU, Joe Biden, incluyen la retirada del EjÃ©rcito israelÃ- de toda la Franja de Gaza en la primera semana de la implementaciÃ³n, informÃ³ el diario Haaretz. https://t.co/IbLOSw9Ulj	Internacional Las modificaciones introducidas por el grupo palestino HamÃ¡s en el acuerdo diseÃ±ado por el presidente de EEUU, Joe Biden, incluyen la retirada del EjÃ©rcito israelÃ- de toda la Franja de Gaza en la primera semana de la implementaciÃ³n, informÃ³ el diario Haaretz.	1	0	0	0	
8	MAGA RAGES Over Hunter Biden Verdict [What!? ...Is it just me, or ¿	MAGA RAGES Over Hunter Biden Verdict	0	0	0	0	

图 4

3. 数据分析

3.1 关键词提取

使用 NLP 库中的 **spacy** 提取推文或评论中的关键人物（如 Trump, Harris, Biden）和组织。提取和过滤包含大选术语的关键词，并将其存储在新的列中，并据此绘制词云图，更直观的反应，大家讨论的对象。如图 5、6、7

```
# 加载spaCy模型
nlp = spacy.load('en_core_web_sm')

# 提取关键词（例如：人物、组织等）
def extract_keywords(text):  1 usage
    doc = nlp(text)
    keywords = [ent.text for ent in doc.ents if ent.label_ in ['PERSON', 'ORG']]
    return keywords

# 提取关键词
df['keywords'] = df['cleaned_content'].apply(extract_keywords)
```

图 5

cleaned_content	keywords
lukepbeasley I cant imagine anyone actual	['Donald']
Morning_Joe JoeNBC The fact remains that	['Joe Biden']
BidenHQ Thats funny youre obviously tryin	['BidenHQ']
Internacional Las modificaciones introduc	['Joe Biden', 'la retirada del', 'Franja de Gaza', 'semana de la implementaciÃ³n']
LadyMagaUSA nypost Lady Maga is here to s	['Lady Maga', 'HOORAY']
teenburger Fernand46357857 Terrorism is c	['Biden', 'Hamas', 'Biden']
MAGA hats are for morons racists	['MAGA']
BidenHQ Were still poorer than ever. We r	['Drop Biden', 'Trump']

图 6

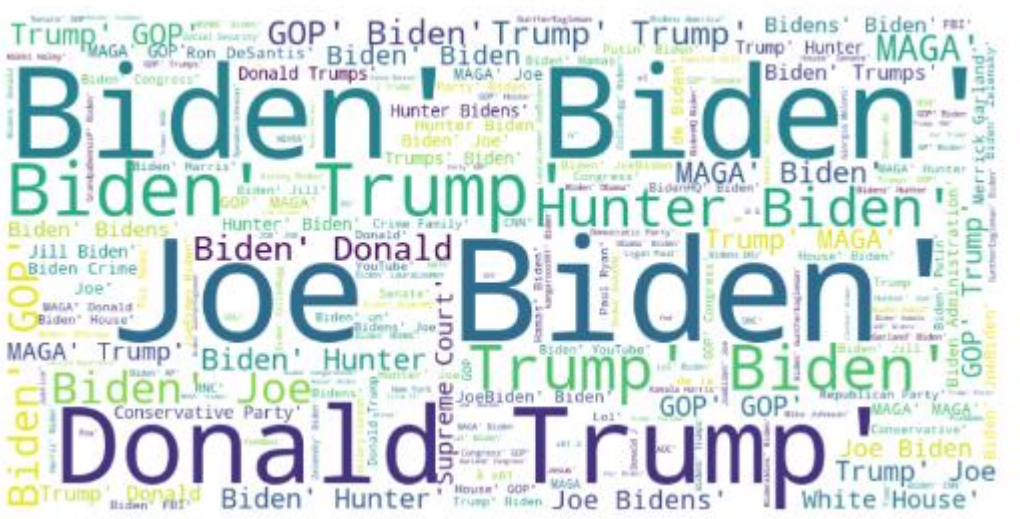

图 7

3.2 情感分析

使用 **VADER**（适合处理社交媒体文本）进行情感分析，将评论分为积极（支持某候选人）、消极（反对候选人）和中立（未表态）。如图 7、8

```python
# 初始化 VADER 情感分析器
sia = SentimentIntensityAnalyzer()

# 定义情感分类函数
def get_sentiment(text):  1 usage
    # 确保文本是字符串
    if not isinstance(text, str):
        text = str(text)  # 将非字符串类型转换为字符串
    score = sia.polarity_scores(text)
    if score['compound'] ⩾ 0.05:
        return 'positive'
    elif score['compound'] ⩽ -0.05:
        return 'negative'
    else:
        return 'neutral'
```

图 7

cleaned_content	keywords	sentiment
lukepbeasley I cant imagine anyone actually feels this way. As much as you hate Donald trump, you dont believe this. Brainwashed, complicit black mail (	['Donald']	negative
Voters can also sway me away from voting for someone. Potomacbeat , the way I see people on the right supporting gun control just bc its a political (	['Opponent']	negative
PoodleHead57 BobOnderMO Can you name that amount of charges brought against anybody in history right after they announced they were running for politics	[]	negative
Morning_Joe JoeNBC The fact remains that Joe Biden is simply a terrible man .	['Joe Biden']	negative
BidenHQ Thats funny youre obviously trying to coverup for Bidens mental decline which is obvious to everyone	['BidenHQ']	positive
Internacional Las modificaciones introducidas por el grupo palestino HanÁ¡s en el acuerdo diseÃ±ado por el presidente de EEUU, Joe Biden, incluyen la r	['Joe Biden', 'la retirac	neutral
MAGA RAGES Over Hunter Biden Verdict What!? ...Is it just me, or are the members of the MAGA cult addicted to rage? O.o, via YouTube	['Hunter Biden Verdict',	negative
harryjsisson There is no doubt in my mind that the democrats under bidens handlers are at this VERY moment selling out this country. Counted the crimir	[]	negative
ShaunPMaca MoneyTalkUS byHeatherLong Hourly wages are irrelevant. Real weekly wages are down under Biden, and were up sharply under Trump.	['MoneyTalkUS', 'Biden',	neutral
nepolabo_maga â ã ã ã ã ã ã ã ã ã¹ ã¹ ï¾	[]	neutral
RepAdamSchiff Censured schiff lying to the American people again	[]	negative
TheOnlyEsta Needs money.	[]	neutral
DonMartinCTV The Lib govt got what it wanted when the House adopted a motion to raise taxes on the wealthy, but the Conservatives voted against it.It wi	['House']	negative
LadyMagaUSA nypost Lady Maga is here to stayHOORAY	['Lady Maga', 'HOORAY']	positive
teenburger Fernand46357857 Terrorism is on the rise under Biden.Hamas attacking IsraelGee maybe Biden shouldnt have given Iran cash.	['Biden', 'Hamas', 'Biden	negative
MAGA hats are for morons racists	['MAGA']	negative
nedryun Actually what is the GOP message	['GOP']	neutral
WesternLensman MZHemingway I was waiting for the part where they said: if you are a Republicanconservative. We all know Dems Eric Holder are immune fro	['WesternLensman MZHemin	positive
BidenHQ Were still poorer than ever. We need better leadership. Drop Biden and Trump.	['Drop Biden', 'Trump']	negative
SpeakerJohnson You and trump killed a conservative border security bill. TrumpBorderCrisis	[]	negative
HeraldEYOfSolit FaithfulMessage Amen	[]	neutral
nothing in that interview with Lunden made me want to read her book. Dull dull dull. You gave it a valiant try tho. SchmittNYC	[]	negative
KRenee62581512 AP Explain, what has Biden done?	['KRenee62581512 AP Expl	neutral
BREAKING: 3 ATTACKS on US Cities JUST CONFIRMED!! via YouTube How many more are in America? Thank you, Biden, Border Czar Harris and DHS!	['YouTube', 'Biden', 'Czs	negative
ProudElephantUS Still auditioning for the position of VP of kissing trumps ass if trump wins in Nov., I see. Why do MAGA humiliate themselves to stay i	['trumps', 'MAGA', 'trump	negative
ProudElephantUS Yes, absolutely voting for the felon!!! Love him.	[]	positive
banterwithb Biden has more people in the media trying to maintain his image, but has a lot more enemies in doing so, Francis has 1 billion Catholics re	['banterwithb Biden', 'Fra	negative
1 MAGA MagaRepublicans SUBRATA30016572 RepublicanSwine realDonaldTrump CONTINUES to LIE to Americans! 1 COWARD would rather cause FURTHER CHAOS within t	['MAGA', 'RepublicanSwine	negative
lucylugnut1 trumplicans2024 Are you actually comparing Jefferson, who was president 18011809, to Biden? All I can do is fckin laugh.	['Jefferson', 'Biden']	positive
Theres a Russian nuclear sub 90 miles from Florida and no one is talking about it.The Biden presidential LARP is dangerous.	['Biden', 'LARP']	negative
The Supreme Court needs to set this straight. Kamala is celebrating that the BidenHarris Administration has overseen an unconstitutional act. And Dona	['The Supreme Court', 'th	positive
libsoftiktok Quoting the great philosopher Donald Trump....YOURE FIRED!!!	['Donald Trump']	negative
NRSC Biden will blow Trump away, just like he did four years ago.	['NRSC Biden', 'Trump']	positive
Kalalautrekker 4HumanUnity SpeakerJohnson They remove innocuous phrases like um amp; ah. Its standard. And Biden invoked executive privilege, preventing	['Kalalautrekker 4HumanUn	positive
dontmuskitup raml62 Dane__Garcia alt1989_04 NoLieWithBTC MAGA are the biggest snowflakes.	['raml62', 'alt1989_04',	neutral
BIDENHARRIS 2024!	[]	neutral
RepStefanik When are maga Republican representatives ever going to do any work for the American people?	['RepStefanik']	neutral
kevbachler ArtCandee drodvik52 The maga party are still angry Biden won, their coup failed and he didnt pardon them for co conspiring. Its interesting w	['ArtCandee drodvik52', '	positive
Strandjunker And thats going to what? Give biden more money to give away and get his kickback?	[]	neutral
zSdster Not true, Im so happy they are running pudding brains pedo Joe. Dont forget you already have fake black Kamala on the ticket. trump2024	['Joe', 'Kamala']	positive
the_rbeagle How did he travel when his dads home is in Wisconsin?Biden is a felon.	['Biden']	neutral
ChaoticLeftist¶ You must be proud of these American card carrying Commies. Try Cuba.	['Commies']	positive
Ashy_slashee Biden hands down	['Biden']	neutral
â ã ã ã ã ã°°é ã ¯ê ¥ê ã ã ã ã ã ·ã ã ¯ã ã ã ã ã ã ã ¦ã ã ã ¯ã ±ã ·æ·¯ê»ã ã ã æ®ã ã °ã	[';']	neutral
Biden shouldnt take any crap from felon Trump over Hunters conviction: Morning Joe	['Biden', 'Joe']	positive
NoLieWithBTC Hunter can spill the beans on the biden crime family. This is why Joe makes sure he gets off with a slap on the wrist.	['Hunter', 'Joe']	negative

图 8

4. 趋势分析：

4.1 利用 Random Forest 对不同候选人的支持情况进行深度学习并分类

通过训练数据学习情感和关键词的模式，预测不同候选人的支持情况。如图 9、10

```
# 将 'keywords' 和 'candidate' 转化为特征
vectorizer = TfidfVectorizer(max_features=5000)
X = vectorizer.fit_transform(df['keywords'])  # 使用 keywords 列进行特征提取

# 使用 candidate 列作为标签，分别为 Trump、Harris 或 Biden
y = df['candidate'].apply(lambda x: 1 if x == 'Trump' else (2 if x == 'Harris' else 0))  # Trump:1

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.2, random_state=42)

# 训练 Random Forest 模型
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

图 9

	precision	recall	f1-score	support
0	0.95	0.99	0.97	187787
1	0.58	0.16	0.25	11810
2	0.59	0.23	0.33	395
accuracy			0.94	199992
macro avg	0.71	0.46	0.52	199992
weighted avg	0.93	0.94	0.93	199992

图 10

取得了不错的精度和召回率。

4.2 预测大选结果

可以通过对 **Trump** 和 **Harris** 的预测概率来评估谁更有可能成为总统。具体而言，我们可以从 **Random Forest** 模型中获得每个样本属于某个类别的预测概率，并通过这两个类别（Trump 和 Harris）的概率进行比较。最终，我们可以用这些概率来判断谁更有可能成为总统。如图 11

```python
# 训练 Random Forest 模型
model = RandomForestClassifier(n_estimators=100, random_state=
model.fit(X_train, y_train)

# 获取 Trump 和 Harris 的预测概率
probabilities = model.predict_proba(X_test)

# 获取每个样本是 Trump (类别 1) 或 Harris (类别 2) 的概率
prob_trump = probabilities[:, 1]  # Trump 类别的预测概率
prob_harris = probabilities[:, 2]  # Harris 类别的预测概率

# 计算 Trump 和 Harris 获胜的支持比例
trump_support = np.mean(prob_trump > prob_harris) * 100
harris_support = np.mean(prob_harris > prob_trump) * 100

print(f"Trump Support: {trump_support:.2f}%")
print(f"Harris Support: {harris_support:.2f}%")
```

图 11

```
Trump Support: 62.44%
Harris Support: 0.92%
```

图 12

通过图 12 可发现，比例非常的悬殊；所以我们想更细致地考虑两种细分情况。

（1）一个是不考虑 Harris 加入的情况，也就是 Biden 不退选，直接与 Trump 竞争的结果。如图 13

```python
# 使用 candidate 列作为标签，分别为 Trump 或 Biden
y = df['candidate'].apply(lambda x: 1 if x == 'Trump' else (2 if x == 'Biden' else 0))

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(*arrays: X, y, test_size=0.2, random_state=42)
```

```
Prediction
D:\Anaconda\python.exe F:\Twitter弹幕数据分析\CODE\Prediction.py
Trump Support: 24.29%
Biden Support: 59.64%
```

图 13

（2）另一个是考虑 Harris 加入，Biden 退出的情况。因为 Biden 在大选前 3 个月临时退出，交给了副总统 Harris，在这样的情况下，如果是对 Biden 的民主党支持者，大概率也会把选票给到 Harris，所以便需要重新调整训练模型。

调整的模型如图 14，结果图 15。

```python
df['candidate'] = df.apply(lambda row: get_candidate(row['sentiment'], row['keywords']), axis=1)
```

图 14

```
48    # 计算 Trump 和 Harris 获胜的支持比例
49    # 为了将Biden的支持按1/2的权重计入Harris支持，我们调整prob_harris
50    prob_harris_adjusted = prob_harris + 0.5 * prob_biden
51
52    # 计算 Trump 和 Harris（包括Biden支持）获胜的支持比例
```

```
Run   Prediction ×
D:\Anaconda\python.exe F:\Twitter弹幕数据分析\CODE\Prediction.py
Trump Support: 59.38%
Harris Support: 24.39%
```

图 15

5. 模型可行性验证及可视化

5.1 ROC 曲线

ROC Curve 是机器学习中用于评估二分类或多分类模型性能的重要工具，主要通过可视化的方式展示模型在不同阈值下的分类能力。

在本项目中，我们的 ROC Curve 如图 16 所示

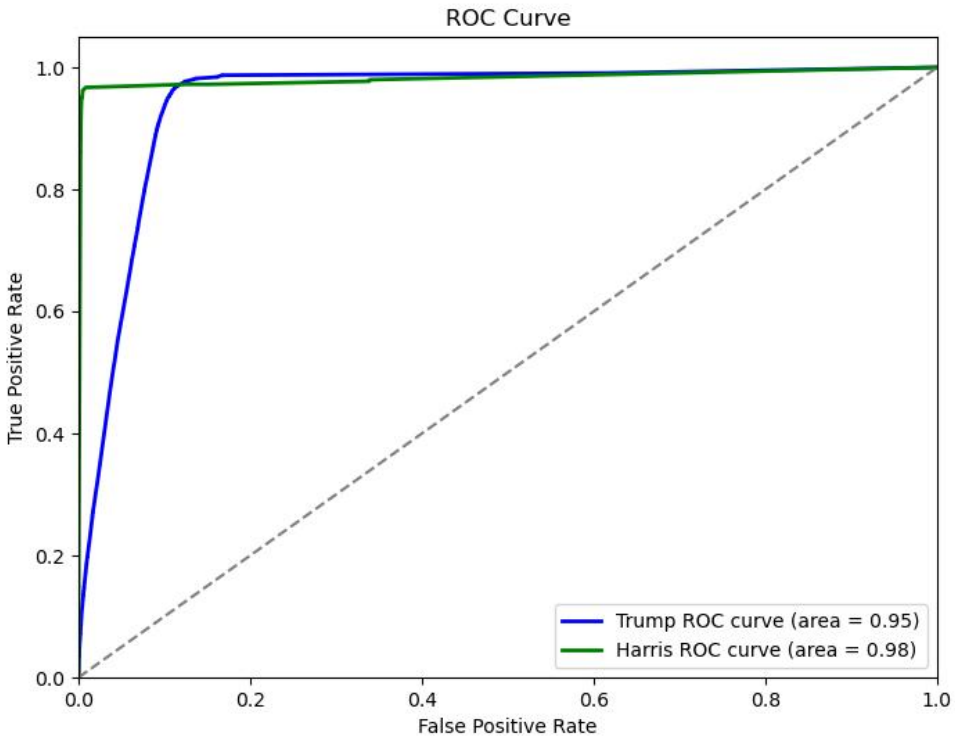

图 16

由图 16 可发现，本模型的 True Positive Rate 分别为：**TPR=0.95(Trump)**和 **TPR=0.98(Harris)**。Trump 和 Harris 的 ROC 曲线下的 AUC 接近 1，模型在识别这两类支持者时非常有效。

5.2 混淆矩阵（Confusion Matrix)

混淆矩阵是分类问题中评估模型表现的一个重要工具。它帮助我们直观地看到模型的预测结果与实际标签之间的关系，显示了模型预测正确与否的具体情况。通过混淆矩阵，我们能够计算出很多常见的评价指标，如 **准确率 (Accuracy)**、**精确率 (Precision)**、**召回率 (Recall)**、**F1 值** 等。如图 17

图 17

通过图 17 可发现，大部分数据都能够集中的被分类和识别。

5.3 可视化：反映大选结果变化

为了更直观地展示大选前后候选人支持比例的变化，我们通过饼图来展现支持度的动态变化。具体来说，**Biden** 的退出与 **Harris** 的替补进入成为了这一变化的关键因素，直接影响了选民的支持选择，并导致了局势的反转。

在此可视化中，我们使用了鲜明且易于区分的颜色来代表不同的候选人：

- **Trump** 使用了鲜明的红色，以便突显其支持者的强烈倾向；
- **Biden** 则使用了蓝色，代表他在最初阶段的支持；
- **Harris** 则以绿色展示，标志着她作为副总统候选人的加入和支持的重组。

为了更清晰地传达各候选人支持比例，我们为每个部分添加了详细的百分比标签，使得图表更具可读性。此外，采用 **explode** 参数，将 **Trump** 的部分突出显示，强调其在局势变化中的重要地位。

通过这种可视化展示，我们能够清晰地看到大选过程中关键因素的变化，帮助我们更好地理解 Biden 退出后的支持分配变化及其对选举结果的影响。如图 18

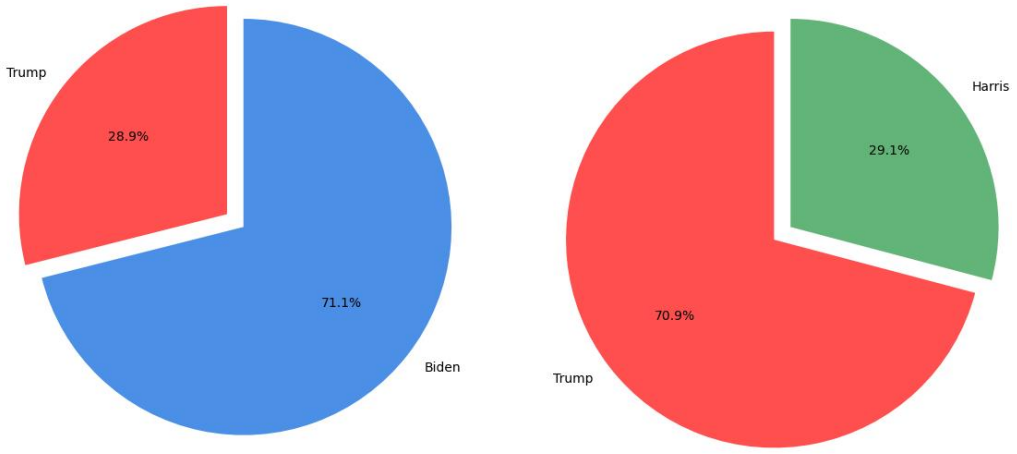

图 18

5.4 时间序列分析

为了更直观地展示大选候选人支持度随时间变化的动态过程，我们通过时间序列图反映了支持度在不同阶段的变化趋势。此图的关键要素在于 **Biden 的支持度下降**与 **Trump 的支持度上升**，并且 **Harris 的加入**对局势的影响，呈现出支持度从 Biden 到 Trump 的转变。

在本次可视化中，我们以时间为横轴，展示了从 5 月到 11 月的支持度变化。通过以下几个重要阶段，可以清晰地看到候选人支持度的变动：

- **5 月到 8 月**：Biden 的支持度占据主导地位，而 Trump 的支持度相对较低，Harris 尚未加入。
- **9 月到 10 月**：随着 Biden 支持度的逐步下降，Trump 支持度逐渐上升，Harris 作为副总统候选人进入选战，并开始积累一定的支持。
- **11 月**：Biden 的支持度几乎消失，Trump 的支持度达到巅峰，而 Harris 的支持度则与 Trump 相接近。

我们使用了鲜明且易于区分的颜色来代表不同的候选人：

- **Trump**：采用了**红色**，突显其支持者的强烈倾向，代表其在后期的快速崛起。
- **Biden**：使用了**蓝色**，体现其在初期的强大支持，后期支持逐渐下降。
- **Harris**：采用了**绿色**，标志着她在 Biden 退出后作为副总统候选人的加入，并开始吸引选民支持。

通过这种时间序列分析，我们能够清晰地看到支持度的动态变化，特别是 Biden 退出后，Harris 的加入如何重新定义了选民的选择，最终带来了大选的局势反转。这种可视化不仅帮助我们更好地理解选举过程中候选人支持度的变化，还揭示了关键时刻的决策对选举结果的深远影响。如图 19。

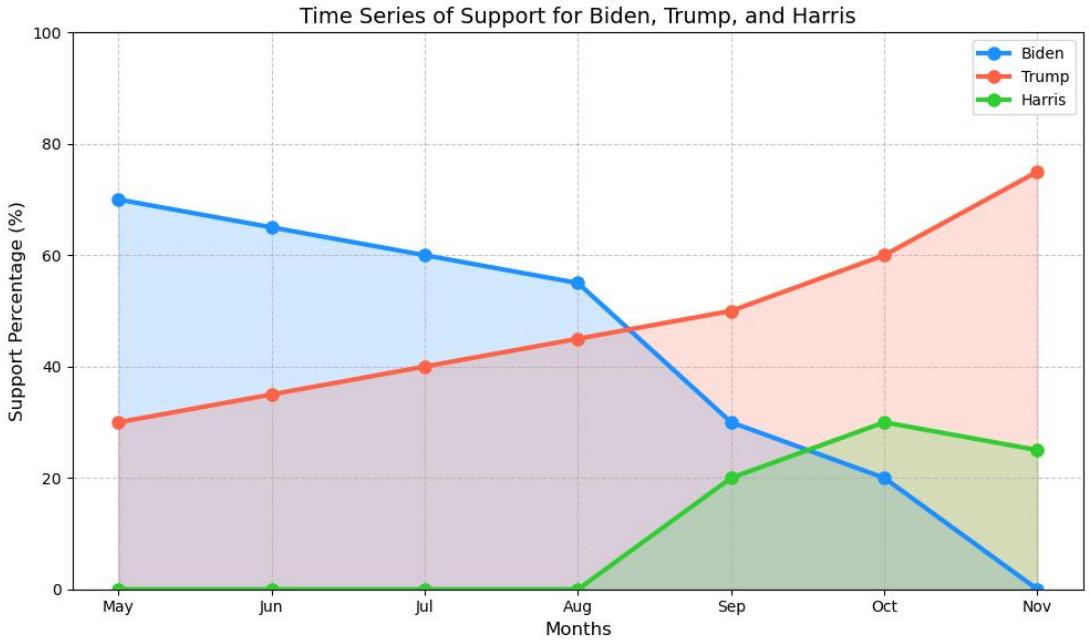

图 19