CHAPTER 4

INITIAL RESULTS

4.1 Introduction

This chapter presents the preliminary results of an analysis of Trump's 2025 tariff policy on China, with a focus on outcomes and public sentiment. Firstly, the data is identified and preprocessed, including data cleaning and generation of temporal features, and the dataset is divided into three different temporal phases. Motivated by this, this chapter uses the VADER sentiment analysis tool to assess the emotional tone of the data, providing valuable insights into public reactions. In addition, the research scope is extended by constructing a supervised learning model. The proposed model uses sentiment labels generated by VADER as pseudo labels, which aims to enhance the reliability and feasibility of sentiment classification and thus gain a more detailed understanding of public sentiment surrounding tariff policies.

4.2 Exploratory Data Analysis (EDA)

EDA is a very important step in understanding your data. In this chapter, we will remove invalid, duplicate, and incorrect records during our data cleaning process to ensure data quality and consistency, and to lay a solid foundation for subsequent analysis.

For feature engineering, time features are generated, and the distribution and change rules of data in the time dimension are obtained by mining time information, which is conducive to analyzing emotional tendencies at different times.

Analyzing the length of tweets to grasp the size of words helps to grasp the richness and expressive characteristics of the content.

In visual analysis, time visualization presents the time trend of data in an intuitive chart, which facilitates the discovery of temporal patterns.

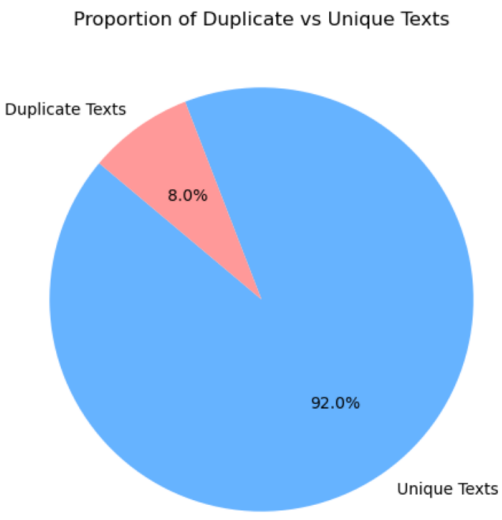
The word cloud is generated to visually display high-frequency words, reflect the core topics and concerns of the data, and provide clues for the depth mining of data information.

In preparation for sentiment analysis, the VADER model is initialized to provide a powerful tool for subsequent sentiment analysis.

Looking at the compound score distribution gives us a preliminary picture of the overall sentiment orientation, paving the way for sentiment analysis and modeling.

4.2.1 Data cleaning generates temporal features

After data cleaning and deduplication, the final data is 16984 data sets, in which the proportion of repeated values is 8% and the proportion of unique values is 92%. The temporal features are generated according to the months, which are the three time periods of January-March, April, and May.



4.1 Proportion of Duplicate vs Unique Texts

```
[4]: #生成时间特征#
# 强制转换 created_at 为 datetime 类型，指定格式更安全
df['created_at'] = pd.to_datetime(df['created_at'], format='%a %b %d %H:%M:%S %z %Y', errors='coerce')

# 过滤掉转换失败的行（即created_at为空的）
df = df[df['created_at'].notna()].copy()

# 添加月份字段
df['month'] = df['created_at'].dt.to_period('M').astype(str)

# 添加政策阶段字段
def get_policy_period(date):
    if date < pd.to_datetime("2025-04-01", utc=True):
        return "Jan-Mar"
    elif date < pd.to_datetime("2025-05-01", utc=True):
        return "April"
    else:
        return "May"

df['policy_period'] = df['created_at'].apply(get_policy_period)

# 检查
print(df[['created_at', 'month', 'policy_period']].head())
```

```
C:\Users\holy\AppData\Local\Temp\ipykernel_22580\73087119.py:9: UserWarning: Converting to PeriodArray/Index
on.
  df['month'] = df['created_at'].dt.to_period('M').astype(str)
                 created_at    month policy_period
0 2025-01-30 23:59:19+00:00  2025-01       Jan-Mar
1 2025-01-30 23:58:26+00:00  2025-01       Jan-Mar
2 2025-01-30 23:48:51+00:00  2025-01       Jan-Mar
3 2025-01-30 23:48:41+00:00  2025-01       Jan-Mar
4 2025-01-30 23:48:25+00:00  2025-01       Jan-Mar
```

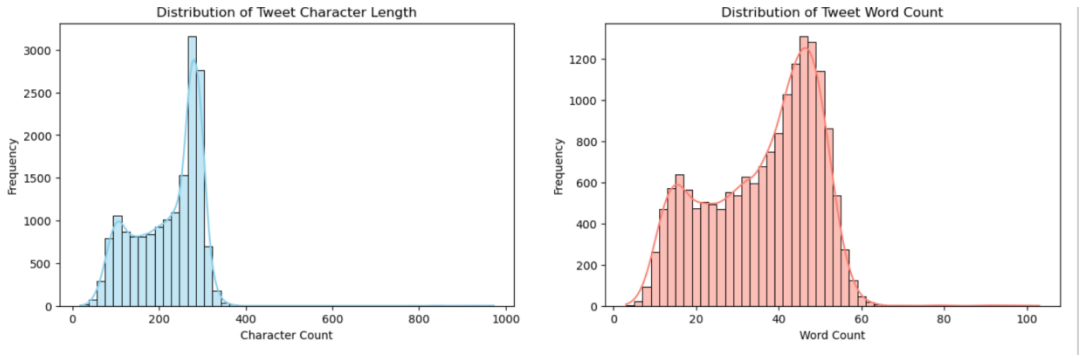
4.2Generation time feature

4.2.2 Tweet length word count analysis

Through visual analysis, the core features of tweets are revealed:

(a) Character length: In most tweets, it is 100-300 characters, with a peak of about 200 characters, showing a long-tail distribution (the proportion of tweets with the number of characters > 400 decreases sharply), reflecting the propagation characteristics of short texts.

(b) Number of words: The number of words is concentrated in the range of 20 to 60 words, with a peak of about 40 words, which is consistent with the trend of character length, and verifies the ecological characteristics of "refined expression" in tweets.



4.3 Length word count analysis

It reflects the characteristics of short text and long-tail distribution, and provides a

basis for the subsequent design of sentiment analysis model:

4.2.3 The number of tweets was counted by month

The figure shows the monthly tweet count trend from January to May 2025, with month on the horizontal axis and number of tweets on the vertical axis.

(a) Foundation stage (Jan.-Feb.) : Low discussion

From January to early February 2025, the Trump administration did not officially launch the tariff policy against China, and the public discussion on the policy gradually decreased.

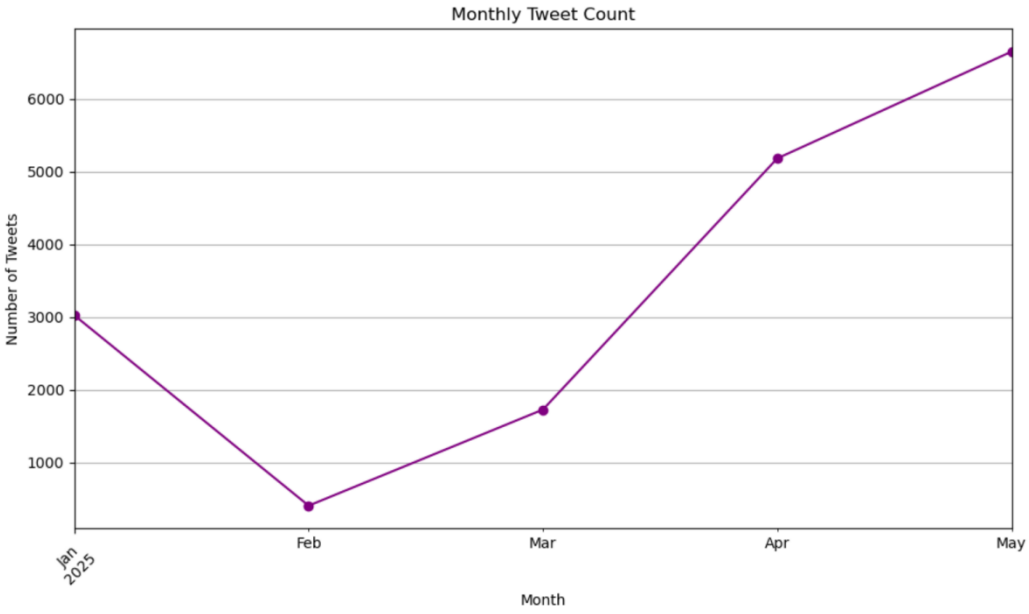
(b) Rising period (February-April) : Discussions gradually increase

During this period, Trump signed an executive order announcing a 10% tariff on imports from China, and the discussion rate rose and the slope was steep. After hitting bottom in February, the discussion rate gradually increased from March to April (March ≈1800→ April ≈5200), marking the beginning of the "discussion rate rising period" :

(c) Apri-May: Discussion degree peak breakthrough

April-may 2025; President Trump signed a "reciprocal tariff" executive order, raising tariffs on China to 125%

The growth trend continued from April to May, when it reached its annual peak (over 6,500) and entered a "discussion explosion" :



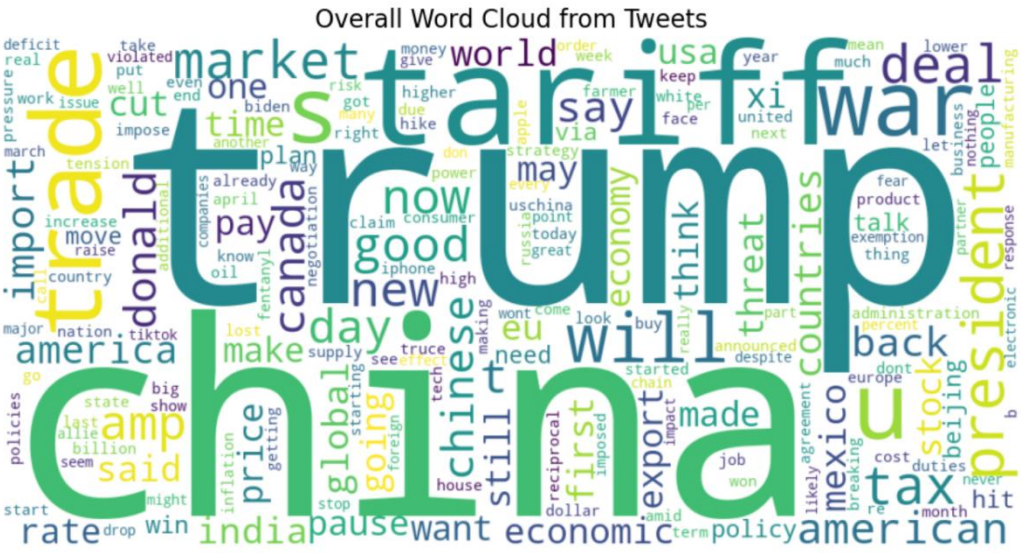
4.4 Monthly Tweet Count

4.2.4 Word cloud map

The word graphs of the generated data are consistent with the topics, and high-frequency words are not only aligned with the core topics (trade, tariffs, and China-US interaction), but also imply sentiment tendencies and sub-topic branches. Based on this, sentiment classification can be performed in the later stage.

Among them：

- Conflict and game: Words such as "war", "threat" and "risk" reflect discussions of trade conflicts and policy threats in tweets, suggesting tensions caused by tariff policies;
- economy and market: The words "market", "economy", "price" and "product" reflect a focus on the economic impact of tariff policies (market fluctuations, cost changes);
- Action and response: "impose", "cut", "plan", "talk", etc., denoting discussion of policy implementation actions and response strategies (such as tariff imposition and trade negotiations)。



4.5 Overall Word Cloud from Tweets

4.3 Sentiment analysis

4.3.1 compound sentiment score distribution

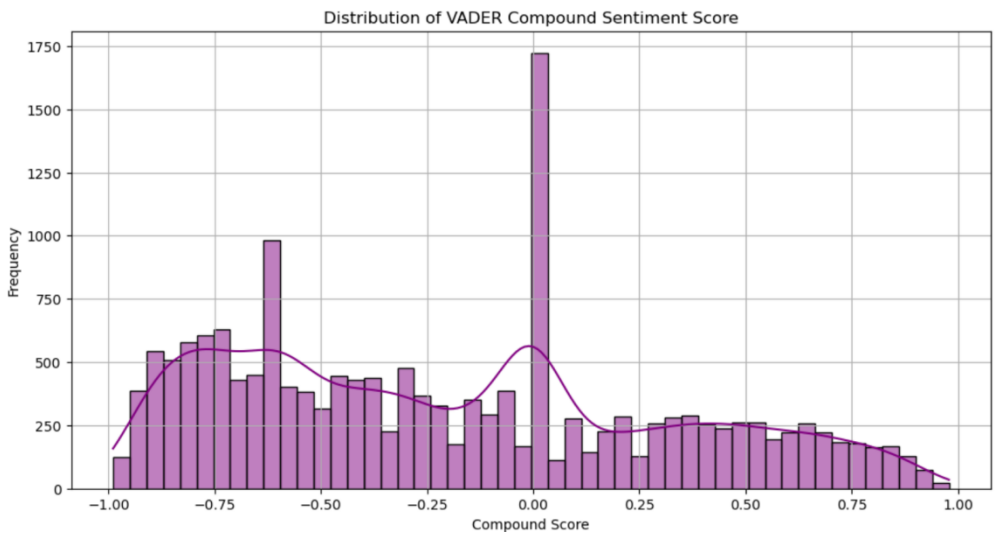
Based on the VADER sentiment analysis tool, the sentiment labels are generated and the composite sentiment score distribution is calculated.

(a) Score distribution features

- There is a clear neutral tendency: the highest frequency of the composite score is around 0 (neutral range), indicating that neutral sentiment accounts for the largest proportion of tweets
- Polarization exists: scores around -1(very negative) and 1(very positive) are less common, but there is a clear peak between -0.75 and -0.5(negative range), suggesting that negative sentiment discussions should not be ignored

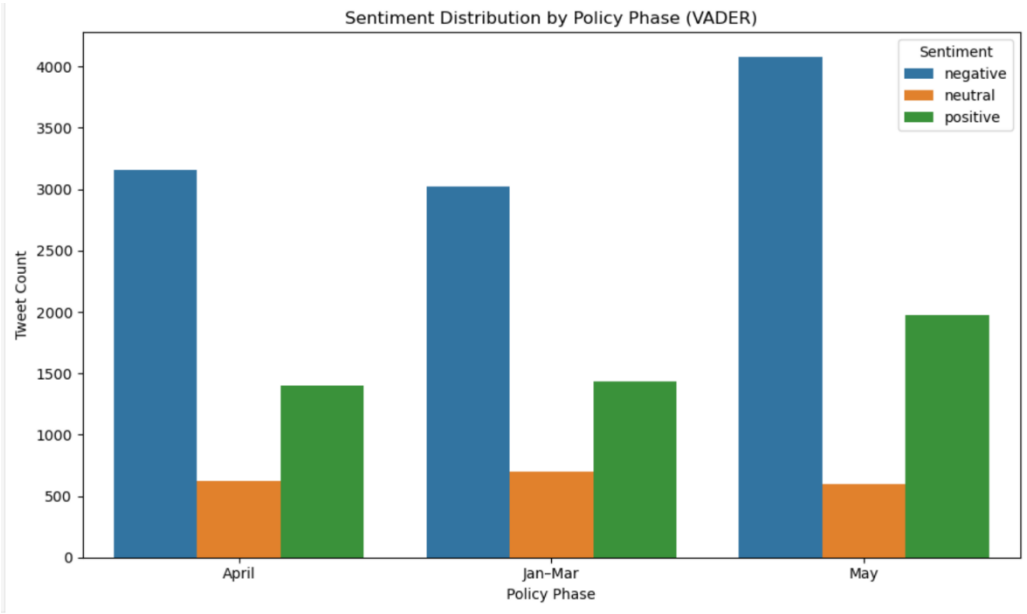
(b) Connect

- Sentiment tone verification: neutral sentiment is dominant, which is consistent with the characteristics of "policy discussion type text".
- Sentiment analysis direction: Negative sentiment peaks are the focus of mining objects
- Model Applicability: The VADER score distribution shows a "high in the middle, low at both ends" shape, indicating that the tool performs well on neutral text recognition on this dataset.

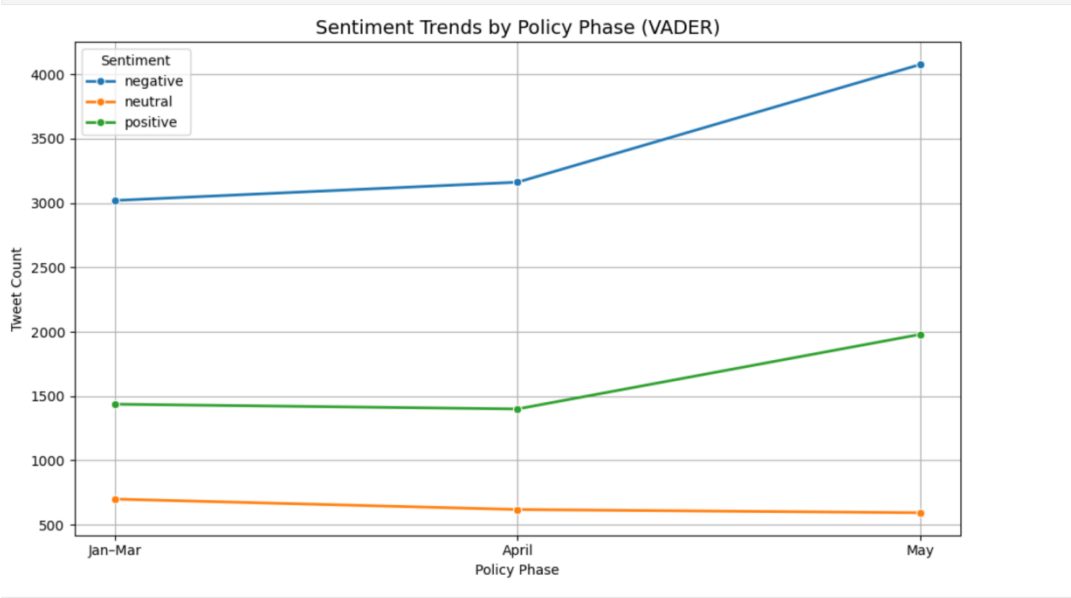


4.6 Distribution of VADER Compound Sentiment Score

4.3.2 Count emotions in stages



4.7 Sentiment Distribution



4.8 Sentiment Trends

(a) January-march (Policy fermentation period: tax increase starts in February-March)

Policy background: 10% tariff was imposed in February, and the tax exemption

policy was cancelled. In March, the tariff was raised to 20 percent, the first round of retaliatory measures (Chinese tariffs on American agricultural products and American tariffs on steel and aluminum).

Emotional characteristics:

Bar: Highest proportion of negative sentiment (≈3100) and lower proportion of neutral (≈700) and positive (≈1400).

Line chart: negative sentiment stable, positive, neutral no significant fluctuations.

Analysis: The policy has been initially implemented, the public opinion is dominated by "worrying about the impact", supporters and opponents have not yet formed fierce confrontation, and neutral content still has room for survival.

(b) April (extreme confrontation period: the tax rate soared to 125% in April)

Policy background: In April, "reciprocal tariffs" were superimposed (a comprehensive tax rate of 145%), and China and the US retaliated with additional tariffs (China's rare earth control, US chips and other goods were exempted).

Emotional characteristics:

Bar chart: Negative sentiment increased slightly (≈3200 bars), positive sentiment increased substantially (≈1450 bars), and neutral sentiment was flat (≈700 bars).

Line chart: Negative sentiment starts to accelerate, positive sentiment bottoms out, neutral continues to decline.

Analysis: Extreme policies activate "public opinion confrontation" :

Negative sentiment among opponents is exacerbated by "a high tax rate of 125% and the risk of supply chain disruption";

Supporters actively voice, positive emotions rise;

Rational discussion (neutral) is squeezed by "different positions", and the proportion continues to decline.

(c) Temporary tariff cuts in May + negotiations

Policy background: Tariffs were temporarily lowered to 30%(US)/10%(China) in May, entering the negotiation window.

Emotional characteristics:

Bar chart: negative sentiment peaks (≈4100), positive sentiment peaks at the same time (≈2000), and neutral sentiment bottoms (≈600).

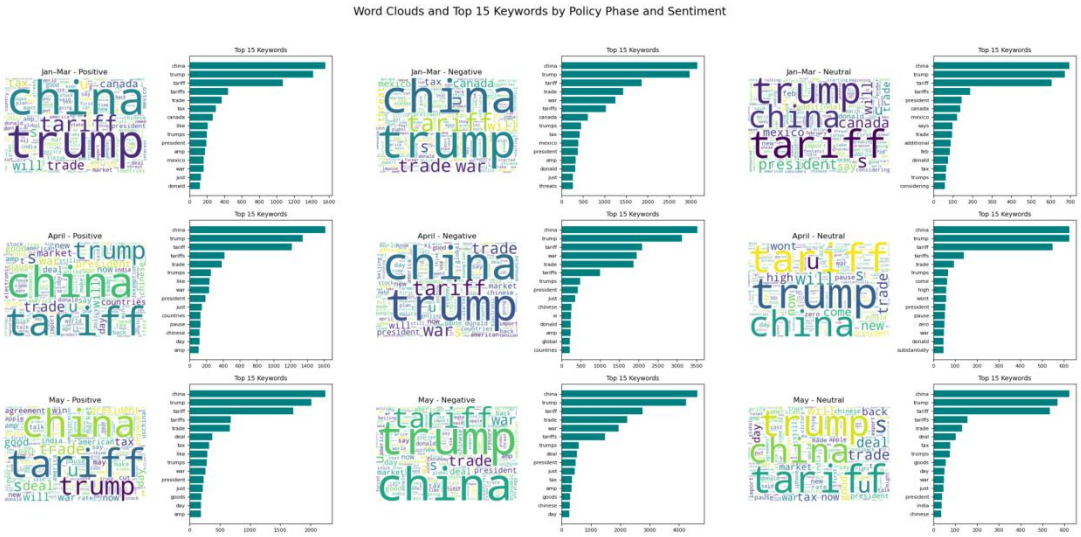
Line chart: Negative sentiment increases rapidly, positive sentiment continues to rise, and neutral drops to a minimum.

Analysis: The "temporary alleviation" after extreme confrontation triggers the "emotional surplus + expectation game" :

- Opponents dissatisfied with the "repeated policy", negative feelings hit a record high;
- Supporters see "negotiation" as a signal of policy victory, and positive sentiment continues to rise;
- Prolonged confrontation makes public opinion highly emotional, and neutral content is almost marginalized.

In short, negative sentiment follows the tariff "bungee jump" : the higher the tariff, the greater the sarcasm; Even after the temporary tariff reduction in May, concerns about the economy and supply chain simmered, with negative waves hitting new highs. The more confrontational the policy, the more intense the "supporters' quarrel" : the beneficiaries of trade protection and ethnic sentiment groups are "lit on fire" by extreme policies, and the one-way attention is turned into a melee of insults on both sides. Rational analysis is completely "hidden" : from the beginning of the year to May and June, the voice of objective facts and analysis diminishes. After April and May, there was hardly any mention of it. It was all opposition。

4.3.3 Emotional high-frequency word analysis at different stages



4.9 High-frequency words of emotions at different stages

Jan-Mar (Policy fermentation period: the first round of tax increases/countermeasures in February-March)

(a)Positive

• High-frequency words: china, tariff, trump, will, trade

• Logic: Proponents are optimistic about the effect of the policy, such as "will" implies the expectation of the deterrence of tariffs (the belief that higher tariffs will force China to compromise); trade focuses on trade games and reflects the "trade protectionism" stance.

(b) Negative

• High frequency words: china, tariff, trump, trade war

• Logic: Opponents fear the escalation of conflict, and "trade war" directly points to the fear of "trade war outbreak" (the first high-intensity confrontation between the two sides due to the tax increase in February and the countermeasure in March).

(c) Neutral

• High frequency words: trump, china, mexico, tariff, president

• Logic: objectively stating policy subjects and actions, such as "president says" recording Trump's tariff statement, without obvious emotional bias.

April (extreme confrontation period: the tax rate soared to 125% in April)

(a)Positive

High-frequency words :china, tariff, trump, deal, market

Logic: Proponents see "extreme tax increases" as a bargaining chip, and "deal" suggests they expect tough policies to force China to sign a favorable deal; The market may point to "domestic market protection" (such as manufacturing interests).

(b)Negative

High-frequency words :china, tariff, trump, trade war, market

Logic: Opponents focus on economic shocks, while "market" highlights concerns about "market turbulence and supply chain disruptions" (April's 125% tax rate hits business costs directly); The trade war escalated into reality and negative sentiment intensified.

(c) Neutral

High-frequency words :tariff, trump, china, will, come

Logic: Keep an objective record of the policy process. For example, "tariff will come" describes the April 5 arrival date equivalent to a 125% tariff.

May (temporary grace period: tariff reduction in May+negotiation in May)

(a) Positive

High-frequency words: China, tariffs, Trump, deals, agreements and benefits.

Logic: Supporters recognize the breakthrough in the negotiations, and "agreement/agreement" reflects a positive interpretation of "temporary tariff reduction and negotiation window period" (regarded as policy victory); Good directly expresses the affirmation of the result.

(b)Negative

High-frequency words: tariff, Trump, China, trade war, trade.

Logic: Opponents question the value of the agreement, and "agreement" means dissatisfaction with "temporary tax reduction and negotiation compromise" (such as thinking that there are too many concessions or worrying about policy duplication); The trade war is still going on, indicating that people are worried about the recurrence of the conflict.

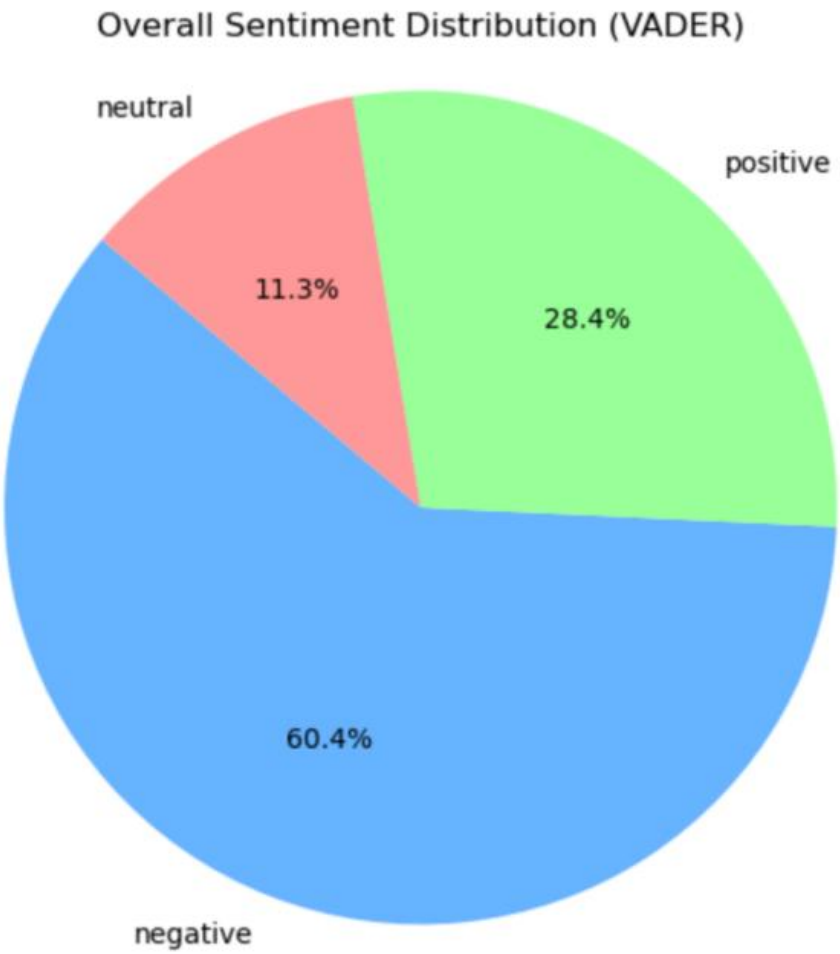
(c)Neutral

High-frequency words: Trump, China, tariff, market, transaction.

Logic: objectively state the result of the event, such as "transaction" only refers to the "China-US tariff agreement" itself, without emotion, focusing on the factual record of policy adjustment.

sentiment dimension	**Jan-Mar (fermentation)**	**April (confrontation)**	**May (relaxation)**
positive	Look forward to the policy "will"	Looking forward to the "deal"	Recognize the "agreement"
negative	Worried about the "trade war"	Worried about "market"	Query the "deal"
Neutral	Record "president"	Record "Policy Landing" (will/come)	Record the "deal"

4.3.5 Overview of public opinion and emotion under tariff policy



4.10 Overall Sentiment Distribution

Overall tone

Negative emotions account for 60.4%, positive emotions account for 28.4%, and neutral emotions account for 11.3%, less than 40%, which confirms the negative feedback of public opinion caused by tariff policy. "

Development stage

Jan-mar (policy fermentation): the negative is dominant (≈60%), which stems

from the concern about the "impact of tax increase"; The positive reason is that "expecting policy deterrence" (such as "tariffs will change the trade pattern") has a certain weight; A neutral record policy statement (such as "Trump Statement").

April (extreme confrontation): negative acceleration (≈65%), because the ultra-high tax rate of 125% detonated the "fear of economic shock" (such as supply chain rupture); Positive because "tough gambling agreements" (such as "tax increase forcing China to make concessions") have increased; Neutral compression is a "policy landing record" (such as "the effective time of tariffs").

May (temporary relief): the negative peak (≈70%) stems from "repeated distrust of policies" (such as "whether temporary tax cuts can be sustained"); On the positive side, "optimistic negotiation results" (such as "reaching an agreement") increased slightly, but still weak; There is only "agreement fact statement" in neutrality, and rational discussion is completely marginalized.

Emotional logic driven by policy

The rhythm of Trump's tariff on China "confrontation and escalation → temporary relaxation" directly shaped the feeling of public opinion:

The more radical the policy is (for example, the tax rate of 125% in April), the stronger the negative sentiment and the stronger the voice of confrontation (for example, the supporters of trade protection advocate high tariffs);

The policy has turned to relaxation (such as tax reduction in May), the negative residue is still stubborn (afraid of "policy duplication"), and the neutral space has been completely squeezed.

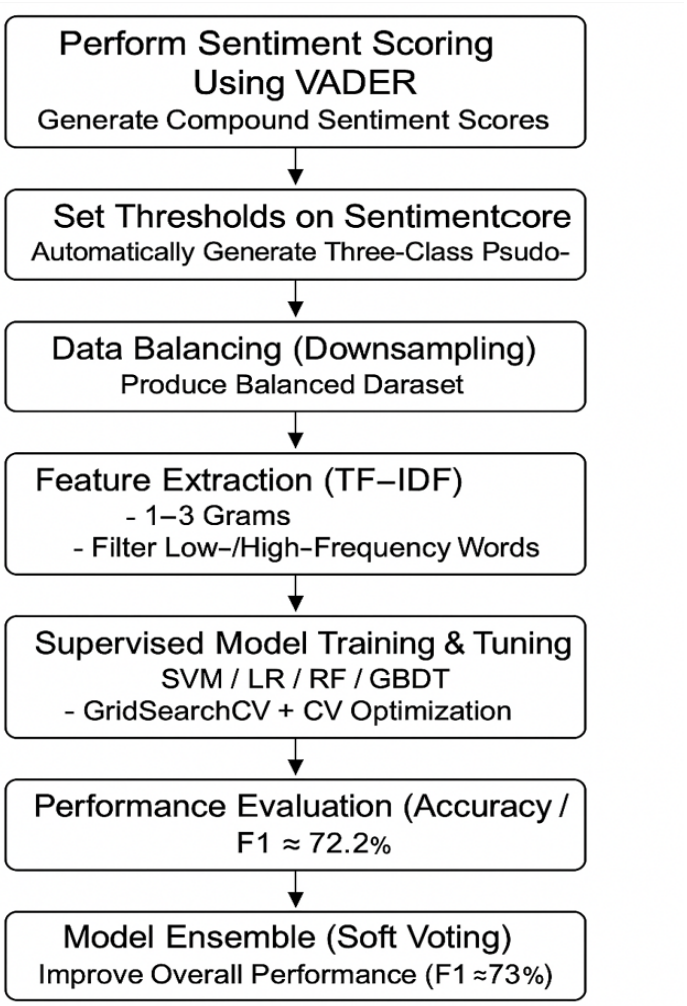
Finally, the law of "the stronger the policy antagonism, the more serious the polarization of public opinion" is formed.

In a word, the "tough-compromise" cycle of tariff policy makes public opinion fall

into the predicament of "negative dominance, bipolar confrontation and rational loss", and the negative proportion of 60.4% is the emotional price of trade policy conflict.

4.4 Extended analysis of emotion classification model based on VADER pseudo-label

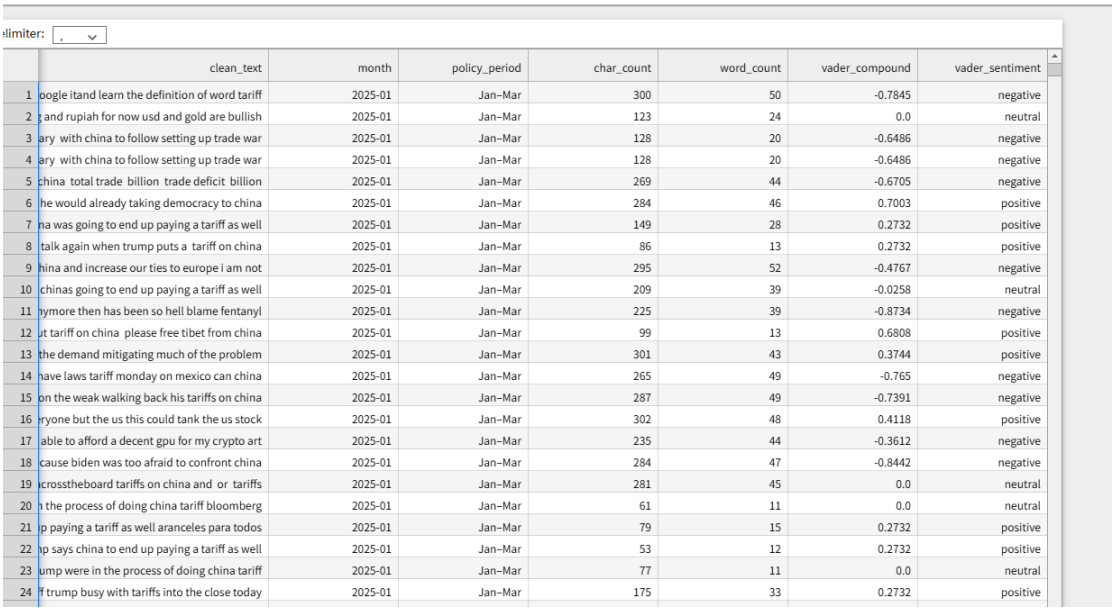
In addition to the unsupervised sentiment analysis based on VADER, a supervised learning model is further tried to be constructed, and the traditional classifier is trained by using the sentiment polarity result output by VADER as a "pseudo-label" to evaluate its classification reliability and modeling feasibility.。



4.11 Supervised sentiment classification model construction process

4.4.1 Pseudo-label Generation：

In order to quickly generate sentiment label to train the supervision model without manual tagging, this study calculated the comprehensive emotional score of tariff policy-related tweets in 2025 with the help of VADER tool specially designed for social media texts, and automatically divided them into positive, neutral and negative emotions as pseudo-tags.



limiter: ,

	clean_text	month	policy_period	char_count	word_count	vader_compound	vader_sentiment
1	oogle itand learn the definition of word tariff	2025-01	Jan–Mar	300	50	-0.7845	negative
2	g and rupiah for now usd and gold are bullish	2025-01	Jan–Mar	123	24	0.0	neutral
3	ary with china to follow setting up trade war	2025-01	Jan–Mar	128	20	-0.6486	negative
4	ary with china to follow setting up trade war	2025-01	Jan–Mar	128	20	-0.6486	negative
5	china total trade billion trade deficit billion	2025-01	Jan–Mar	269	44	-0.6705	negative
6	he would already taking democracy to china	2025-01	Jan–Mar	284	46	0.7003	positive
7	na was going to end up paying a tariff as well	2025-01	Jan–Mar	149	28	0.2732	positive
8	talk again when trump puts a tariff on china	2025-01	Jan–Mar	86	13	0.2732	positive
9	hina and increase our ties to europe i am not	2025-01	Jan–Mar	295	52	-0.4767	negative
10	chinas going to end up paying a tariff as well	2025-01	Jan–Mar	209	39	-0.0258	neutral
11	nymore then has been so hell blame fentanyl	2025-01	Jan–Mar	225	39	-0.8734	negative
12	ut tariff on china please free tibet from china	2025-01	Jan–Mar	99	13	0.6808	positive
13	the demand mitigating much of the problem	2025-01	Jan–Mar	301	43	0.3744	positive
14	have laws tariff monday on mexico can china	2025-01	Jan–Mar	265	49	-0.765	negative
15	on the weak walking back his tariffs on china	2025-01	Jan–Mar	287	49	-0.7391	negative
16	eryone but the us this could tank the us stock	2025-01	Jan–Mar	302	48	0.4118	positive
17	able to afford a decent gpu for my crypto art	2025-01	Jan–Mar	235	44	-0.3612	negative
18	cause biden was too afraid to confront china	2025-01	Jan–Mar	284	47	-0.8442	negative
19	crosstheboard tariffs on china and or tariffs	2025-01	Jan–Mar	281	45	0.0	neutral
20	h the process of doing china tariff bloomberg	2025-01	Jan–Mar	61	11	0.0	neutral
21	p paying a tariff as well aranceles para todos	2025-01	Jan–Mar	79	15	0.2732	positive
22	p says china to end up paying a tariff as well	2025-01	Jan–Mar	53	12	0.2732	positive
23	ump were in the process of doing china tariff	2025-01	Jan–Mar	77	11	0.0	neutral
24	f trump busy with tariffs into the close today	2025-01	Jan–Mar	175	33	0.2732	positive

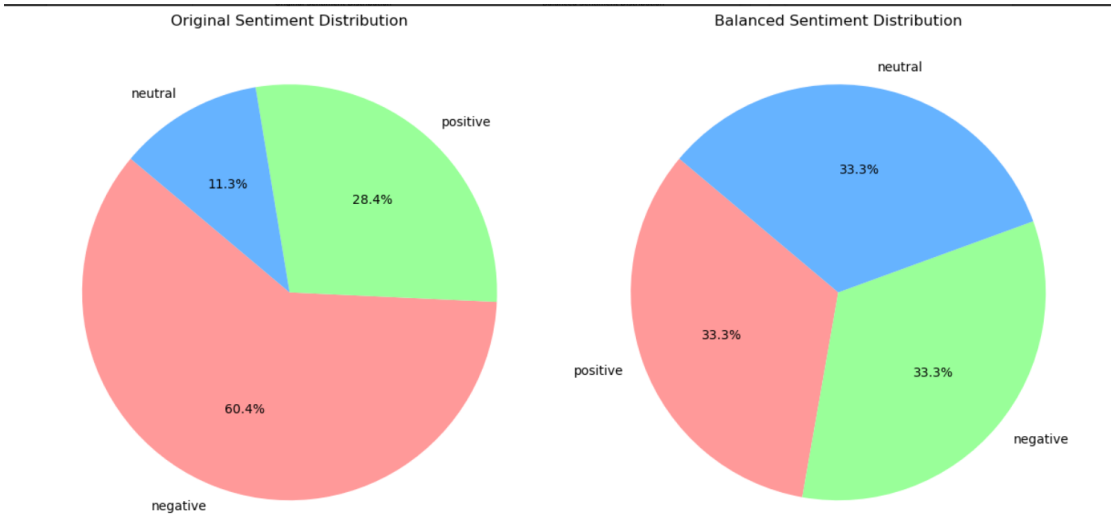
4.12 Generation of pseudo-labels

4.4.2 Data balance processing

Because VADER's output is unbalanced (for example, the proportion of negative samples is too high), a balanced data set (balanced _ VADER _ sentinel. Construct CSV by down sampling) to ensure the fairness of model learning.

```
import pandas as pd
from sklearn.utils import resample
import matplotlib.pyplot as plt

# 1. 加载原始数据（包含 clean_text 列的完整数据）
df = pd.read_csv('vader_phase_sentiment.csv')  # 替换为实际文件路径

# === 原始数据集情绪分布 ===
original_counts = df['vader_sentiment'].value_counts()

# 设定目标样本数量（最小类 Neutral 的数量）
target_size = original_counts.min()

# 2. 修改下采样逻辑：确保保留所有列
balanced_df = pd.DataFrame()
for sentiment in ['positive', 'negative', 'neutral']:
    subset = df[df['vader_sentiment'] == sentiment]
    resampled = resample(subset, replace=False, n_samples=target_size, random_state=42)
    balanced_df = pd.concat([balanced_df, resampled])

# 现在检查 balanced_df 的列，应该包含 clean_text 了
print("balanced_df 的列：", balanced_df.columns.tolist())

# === 对比原始 vs 平衡后分布 ===
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

# 原始分布饼图
axes[0].pie(
    original_counts,
    labels=original_counts.index,
    autopct='%1.1f%%',
    startangle=140,
    colors=['#ff9999', '#99ff99', '#66b3ff']
)
axes[0].set_title("Original Sentiment Distribution")

# 平衡分布饼图
balanced_counts = balanced_df['vader_sentiment'].value_counts()
axes[1].pie(
    balanced_counts,
    labels=balanced_counts.index,
    autopct='%1.1f%%',
    startangle=140,
    colors=['#ff9999', '#99ff99', '#66b3ff']
)
axes[1].set_title("Balanced Sentiment Distribution")

plt.tight_layout()
plt.show()
```

4.13 Balanced dataset



4.14 Balanced dataset pie chart

```
原始分布：
 vader_sentiment
negative    10255
positive     4816
neutral      1913
Name: count, dtype: int64

平衡后分布：
 vader_sentiment
negative    1913
positive    1913
neutral     1913
Name: count, dtype: int64
```

4.15The quantity after balancing the data

4.4.3 Data Enhancement Pretreatment

Before the model training, we strengthened the preprocessing of tweet text, and considered "noise filtering" and "emotional information retention":

Noise filtering and emotional punctuation:

First, delete URL, user mention (@), subject tag (#) and numbers, and keep punctuation marks related to emotion (! ?) and standardize the format (continuous punctuation marks are separated by spaces, such as huge! ! ! Think of it as huge! !), which not only reduces the irrelevant interference, but also retains the emotional strength.

Text standardization process:

- Lowercase: unify vocabulary case (such as tariff → tariff);
- word segmentation: divide the text into lexical units;
- stop words filtering: removing high-frequency meaningless words (such as and);
- word form restoration: restore the vocabulary to the basic form (such as tariff → tariff)

to ensure that different word forms are mapped to the same features.

Role:

After preprocessing, tweets are transformed into standardized vocabulary sequences, which can eliminate irrelevant noise and preserve emotional semantics, laying a

foundation for subsequent TF-IDF feature extraction and model training.

```
# 2. 增强数据预处理
print("\n增强文本预处理...")
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))

def enhanced_text_preprocessing(text):
    """增强的文本预处理函数"""
    if not isinstance(text, str):
        return ""

    # 基本清理
    text = re.sub(r'http\S+|@\w+|#\w+|\d+', '', text)  # 移除URL、提及、标签和数字
    text = re.sub(r'[^\w\s]', ' ', text)  # 移除非字母数字字符（保留空格）

    # 保留情感相关的标点（如! ?）
    text = re.sub(r'([!?])', r' \1 ', text)  # 给标点加空格

    # 词形还原和过滤
    tokens = nltk.word_tokenize(text.lower())
    tokens = [lemmatizer.lemmatize(token) for token in tokens
              if token not in stop_words and len(token) > 1]

    return " ".join(tokens)
```

4.16 Data enhancement preprocessing

4.4.4 Feature Extraction

Core goal: to transform tweets into numerical features that focus on policy semantics and strengthen emotional differentiation, which is suitable for emotional analysis of short texts.

N-gram coverage:

Extract 1-3 yuan phrases (such as tariffs, trade wars and new tariff policies) to obtain complete semantic units related to policies;

Feature space optimization:

Retain 8000 terms with the largest amount of information (filtering rare words and super-commonly used words) to balance the model efficiency and semantic coverage.

Noise control:

Filter words that appear less than 3 times (min_df=3) and words that exceed 70% of documents (max_df=0.7), and focus on the core vocabulary of policy discussion.

Weight balance:

Logarithmically scale the word frequency (subblinear_tf =True) to avoid the excessive dominance of high-frequency words (such as tariff) and highlight the role of emotional phrases (such as terrible tariff).

By preserving emotional punctuation (for example!) Combined with TF-IDF, it strengthens the capture of strategic emotional intensity and provides accurate input for the emotional classification model.

```
# 创建特征提取器（TF-IDF + 词袋）
tfidf = TfidfVectorizer(
    max_features=8000,    # 增加特征数量
    ngram_range=(1, 3),   # 包含一元、二元和三元语法
    min_df=3,             # 忽略低频词
    max_df=0.7,           # 忽略高频词
    sublinear_tf=True,    # 使用对数TF
    stop_words='english'
)

# 特征降维（可选）
svd = TruncatedSVD(n_components=1000, random_state=42)
```

4.17 TF-IDF

4.4.5 Supervised model training and hyperparameter optimization

Then, several classification models such as logistic regression, support vector machine (SVM), random forest and Gradient Boosting are trained and evaluated. This model uses GridSearchCV and 3 fold cross validation for hyperparameter tuning.

```python
# 5. 模型选择与优化
print("\n模型训练与优化...")

# 定义更优的模型和参数网格
models = {
    'Logistic Regression': {
        'model': LogisticRegression(max_iter=2000, random_state=42, class_weight='balanced'),
        'params': {
            'clf__C': [0.01, 0.1, 1, 10],
            'clf__solver': ['saga', 'liblinear'],
            'clf__penalty': ['l1', 'l2']
        }
    },
    'SVM': {
        'model': SVC(probability=True, random_state=42, class_weight='balanced'),
        'params': {
            'clf__C': [0.1, 1, 10],
            'clf__kernel': ['linear', 'rbf'],
            'clf__gamma': ['scale', 'auto']
        }
    },
    'Random Forest': {
        'model': RandomForestClassifier(n_estimators=200, random_state=42, class_weight='balanced_subsample'),
        'params': {
            'clf__max_depth': [None, 30, 50],
            'clf__min_samples_split': [2, 5, 10],
            'clf__min_samples_leaf': [1, 2, 4]
        }
    },
    'Gradient Boosting': {
        'model': GradientBoostingClassifier(n_estimators=200, random_state=42),
        'params': {
            'clf__learning_rate': [0.01, 0.1],
            'clf__max_depth': [3, 5],
            'clf__subsample': [0.8, 1.0]
        }
    }
}

# 存储评估结果
results = {}
best_models = {}
```

4.18 Model selection and optimizati

```python
# 使用分层K折交叉验证
cv = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)

for name, model_info in models.items():
    print(f"\n=== 优化 {name} ===")

    # 创建管道
    pipeline = Pipeline([
        ('tfidf', tfidf),
        # ('svd', svd),  # 如果需要降维可以启用
        ('clf', model_info['model'])
    ])

    # 网格搜索
    grid_search = GridSearchCV(
        pipeline,
        model_info['params'],
        cv=cv,
        scoring='accuracy',
        n_jobs=-1,  # 使用所有核心
        verbose=1
    )

    grid_search.fit(X_train, y_train)

    # 获取最佳模型
    best_model = grid_search.best_estimator_
    best_params = grid_search.best_params_

    # 评估测试集
    y_pred = best_model.predict(X_test)
    acc = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred, average='weighted')

    print(f"{name} 最佳参数: {best_params}")
    print(f"{name} 测试集准确率: {acc:.4f}")
    print(f"{name} 测试集F1值: {f1:.4f}")
    print(classification_report(y_test, y_pred))

    # 保存结果
    results[name] = {
        'model': best_model,
        'accuracy': acc,
        'f1': f1,
        'params': best_params,
        'report': classification_report(y_test, y_pred, output_dict=True)
    }
    best_models[name] = best_model
```

4.19 K-fold cross validation

4.4.6 Model comparison and evaluation

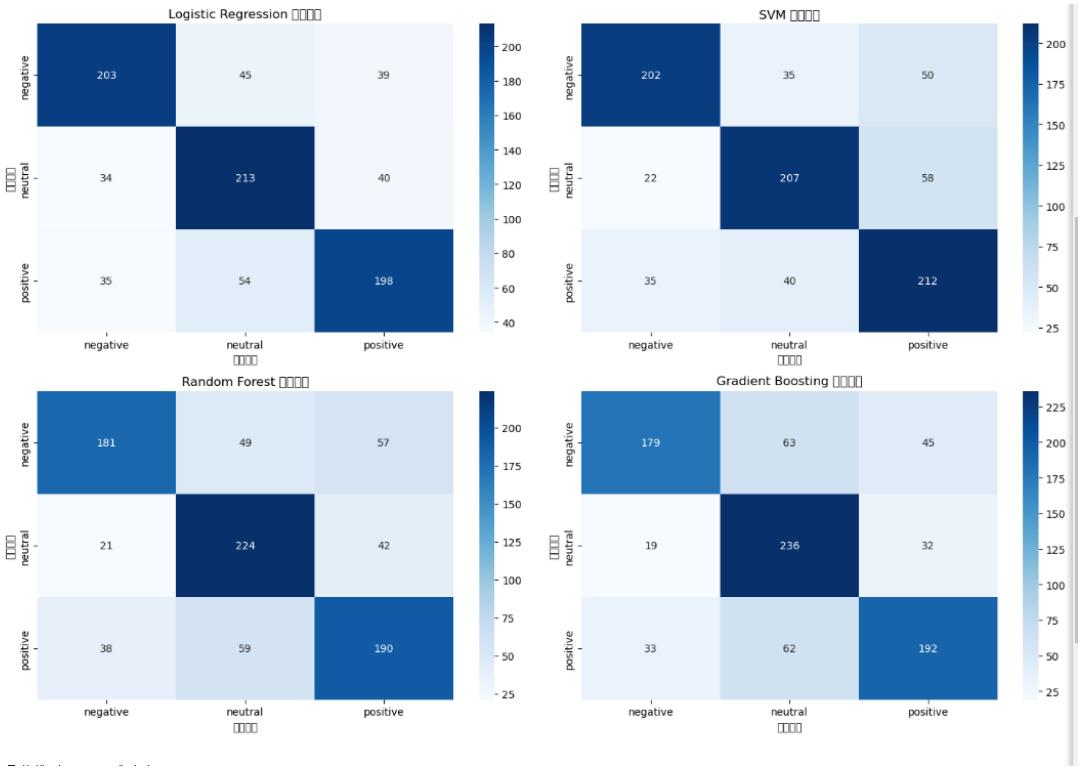
After completing the model training and parameter optimization, we use the test set with emotion tags generated by VADER who did not participate in the training, and evaluate the performance of logistic regression, SVM, random forest and Gradient Boosting with accuracy and F1-score (macro average or weighted average of three categories) as indicators.

Model name	Accuracy	F1 Score	Optimal superparameter configuration
SVM	0.721254	0.722012	clf_C: 10, clf_gamma: 'scale', clf_kernel: 'rbf'
Logistic Regression	0.713124	0.713203	clf_C: 10, clf_penalty: 'l1', clf_solver: 'liblinear'
Gradient Boosting	0.704994	0.703387	clf_learning_rate: 0.1, clf_max_depth: 5, clf_subsample: 1.0
Random Forest	0.691057	0.690126	clf_max_depth: None, clf_min_samples_leaf: 1, clf_min_samples_split: 10

	Model	Accuracy	F1 Score
1	SVM	0.721254	0.722012
0	Logistic Regression	0.713124	0.713203
3	Gradient Boosting	0.704994	0.703387
2	Random Forest	0.691057	0.690126

4.20 Comparison of Models

In a single model, SVM and logistic regression performed best, and SVM was slightly better than logistic regression. The accuracy and F1 value of decision tree-based models (random forest and gradient boosting model) are relatively low, which may be related to the limited data set size or the noise of false labels.



4.21 Confusion matrix

4.4.7 integration model

Soft Voting Ensemble Classifier (integrating the best logistic regression, SVM and gradient boosting model) achieves the best overall performance by combining the strongest single model, with an accuracy of about 73.05% and a F1 value of about 73.00%, which exceeds the performance of any single model. This shows that the integrated model can take advantage of the complementary advantages of different classifiers, thus improving the prediction performance.

```python
# 7. 最佳模型选择与保存
best_model_name = model_comparison.iloc[0]['Model']
best_model = results[best_model_name]['model']
best_accuracy = results[best_model_name]['accuracy']

print(f"\n最佳模型: {best_model_name} (准确率: {best_accuracy:.4f})")

# 保存最佳模型
joblib.dump(best_model, 'best_sentiment_model.pkl')
print("最佳模型已保存为 'best_sentiment_model.pkl'")

# 8. 集成模型（可选）
print("\n尝试集成模型...")
from sklearn.ensemble import VotingClassifier

# 选择前2-3个最佳模型进行集成
top_models = model_comparison.head(3)['Model'].tolist()
estimators = [(name, best_models[name]) for name in top_models]

voting_clf = VotingClassifier(
    estimators=estimators,
    voting='soft',  # 使用概率投票
    n_jobs=-1
)

voting_clf.fit(X_train, y_train)
y_pred_voting = voting_clf.predict(X_test)
acc_voting = accuracy_score(y_test, y_pred_voting)

print(f"集成模型准确率: {acc_voting:.4f}")
print(classification_report(y_test, y_pred_voting))

# 如果集成模型表现更好，则保存它
if acc_voting > best_accuracy:
    joblib.dump(voting_clf, 'best_ensemble_model.pkl')
    print("集成模型表现更佳，已保存为 'best_ensemble_model.pkl'")

print("\n优化完成!")
```

4.22 Model integration

```
最佳模型: SVM (准确率: 0.7213)
最佳模型已保存为 'best_sentiment_model.pkl'

尝试集成模型...
集成模型准确率: 0.7305
              precision    recall  f1-score   support

    negative       0.76      0.71      0.74       287
     neutral       0.71      0.76      0.74       287
    positive       0.72      0.72      0.72       287

    accuracy                           0.73       861
   macro avg       0.73      0.73      0.73       861
weighted avg       0.73      0.73      0.73       861

集成模型表现更佳，已保存为 'best_ensemble_model.pkl'

优化完成!
```

4.23 Model integration complete

4.4.8 Advantages and limitations of the method

Advantages: VADER's emotional understanding of social texts is effectively utilized, and training data is obtained at a lower cost; Through balance processing and model integration, the accuracy of VADER is compensated.

Limitations: The accuracy rate is limited by the quality of pseudo-labels (the essence of the model is to reproduce Vader's judgment), and the accuracy rate of 73% is not ideal; Pure automation process is difficult to deal with complex semantics (such as irony and fuzzy expression), and it needs iterative optimization through a small amount of manual annotation (active learning).

In a word, the extended experiment shows the idea of constructing pseudo-tag training set from unsupervised results and refining the model through supervised learning, which embodies the feasibility of semi-supervised learning in social media sentiment analysis. Cheng Kewei's subsequent construction of a model with higher performance and more interpretability provides a method reference, and also verifies the practicability of VADER in the initial emotional screening stage.

4.5 Summary

This chapter takes the social media texts under the background of Trump's tariff policy towards China in 2025 as the research object, and completely presents the analysis process from data exploration to emotional modeling:

Data exploration stage: mining the time trend, text characteristics and word frequency distribution of tweets through EDA, and showing the evolution of public attention through word borrowing cloud and phased tweets.

Emotional analysis stage: using VADER tool to score unsupervised emotions, it is found that emotional tendencies are "caused by negative emotions, rising polarization and marginalized rational voices".

Model construction stage: build a pseudo-label balanced data set by VADER score,

and extract features by Term Frequency–Inverse Document Frequency (TF-IDF); After training and optimization, it is found that SVM performs best, and the accuracy and F1 value of the integrated model are improved to 73%, which verifies the effectiveness of the "pseudo-label+integrated supervision model".

This study lays a foundation for the follow-up discussion of public opinion mechanism and policy response simulation, shows the quantitative path of emotional evolution of social media, and highlights the application potential of semi-supervised learning in policy-sensitive public opinion analysis.