

**Sentiment Tag-Driven Prediction Model of Weibo Data during Wuhan Epidemic**

Weihua SUN 1155157063 Rui RAO 1155220548



2024-11-10

DOTE5110

Statistical Analysis

## I. Introduction

T

his study constructs a Sentiment Analysis model with Weibo data during the Wuhan epidemic. The core techniques include TF-IDF vectorization, BernoulliNB, LinearSVC, and Logistic Regression. The goal is to automatically categorizing new Weibo post sentiment as Positive or Negative.

Significantly, it provides access to public sentiment, benefiting policymakers and researchers in marketing, social studies, and journalism. Our group is specifically interested in using machine learning to explore the link between the epidemic and sentiment.

## II. Data Description

Our data was extracted from Weibo from 2019.12.30 to 2020.05.30 and contains 201,883 post records after cleaning.

It contains the following 6 fields:

1. **ids:** the id of the post
2. **text:** the text of the post
3. **sentiment\_label:** the polarity of the post (0 = Negative, 1 = positive)
4. **sentiment\_key:** the polarity of the post
5. **positive\_probs:** how much probability the post's polarity is likely to be positive
6. **negative\_prob:** show much probability the post's polarity is likely to be negative

We require only the **sentiment** and **text** fields, so we discard the rest.

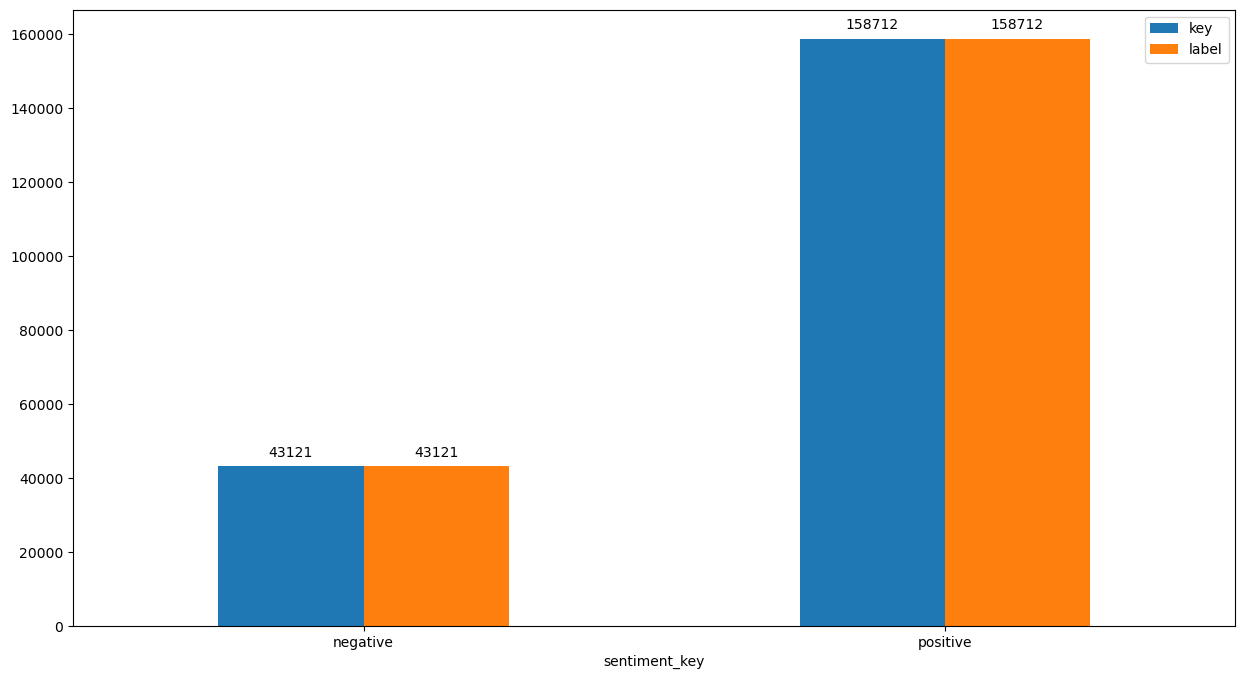
**Head**

Note that we sort the data that negative records are positioned before positive records.

表格, 日历

描述已自动生成

**Distribution of labels**



## III. Data Preprocessing

To render the text in a more digestible format for the machine learning models, the following preprocessing procedures are implemented:

1. **URL Replacement**: Hyperlinks commencing with "http" or "https" or "www" are supplanted with "提到某链接".
2. **Username Replacement**: @Usernames are substituted with the term "某用户".
3. **Stopword Removal**: Stopwords are those that contribute minimally to the semantic essence of a sentence and can be disregarded without impairing the overall meaning. We employ jieba to segment Chinese sentences and eliminate stopwords in accordance with a predefined list.

Now that we have attained the preprocessed dataset, we are afforded a more lucid perspective. We will proceed to generate a word cloud for both positive and negative Weibo posts.

### Word-Cloud for Negative Posts



### Word-Cloud for Positive Posts



### Dealing with skewness

Note that our dataset is skewed, which brings potential risks. An unequal distribution of positive and negative samples may cause models to favor the majority class, leading to inaccurate evaluation and overlooking the minority sentiment, thus distorting the representation of public opinion.

To address this, we used an undersampling method. As the dataset was randomized and sorted by sentiment\_label, we took the first 86,242 entries to balance the classes.

## IV. Model Building and Evaluation

### A. Model Selection

Three different types of models are included in this study:

1. Bernoulli Naive Bayes (BernoulliNB)：Based on Bayes' theorem and the feature independence assumption, it is suitable for discrete data. The calculation is simple and efficient, which make it has an advantage for large, high-dimensional Weibo data.
2. Linear Support Vector Classification (LinearSVC)：It finds a hyperplane for classification and maps to a high-dimensional space to make data linearly separable. It performs excellently in handling linear or approximately linear data, is good at processing high-dimensional data, has strong robustness and excellent generalization.
3. Logistic Regression (LR)：It uses a logistic function to obtain the probability of a category. It is simple, easy to explain and can output probabilities, widely used in many classification tasks. It has stable and accurate performance in classification.

### B. Preparing for Training and Testing Data

The Preprocessed Data is divided into 2 sets of data:

* **Training Data:** The dataset upon which the model would be trained on. Contains 95% data.
* **Test Data:** The dataset upon which the model would be tested against. Contains 5% data.

#### **TF-IDF Vectoriser**

TF-IDF indicates what the importance of the word is in order to understand the document or dataset, it converts a collection of raw documents to a matrix of TF-IDF features.

### C. Evaluation Metrics

Since our dataset is skewed, that is, the number of positive and negative samples is unbalanced, we will reduce the sample quantity of positive records with Undersampling, simply by cutting all records after index 86242. Accuracy, Precision, Recall, and F1 - score are taken as the main evaluation metrics.

Furthermore, we're plotting the **Confusion Matrix** to get an understanding of how our model is performing on both classification types.

### D. Model Training and Evaluation Results

#### **Model 1: Bernoulli Naive Bayes**

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 | 0.92 | 0.80 | 0.85 | 2152 |
| 1 | 0.82 | 0.93 | 0.87 | 2161 |
| accuracy | - | - | 0.86 | 4313 |
| macro avg | 0.87 | 0.86 | 0.86 | 4313 |
| weighted avg | 0.87 | 0.86 | 0.86 | 4313 |

#### **Model 2: Linear Support Vector Classification**

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 | 0.93 | 0.93 | 0.93 | 2152 |
| 1 | 0.93 | 0.93 | 0.93 | 2161 |
| accuracy | - | - | 0.93 | 4313 |
| macro avg | 0.93 | 0.93 | 0.93 | 4313 |
| weighted avg | 0.93 | 0.93 | 0.93 | 4313 |

#### **Model 3: Logistic Regression**

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 | 0.92 | 0.92 | 0.92 | 2152 |
| 1 | 0.92 | 0.92 | 0.92 | 2161 |
| accuracy | - | - | 0.92 | 4313 |
| macro avg | 0.92 | 0.92 | 0.92 | 4313 |
| weighted avg | 0.92 | 0.92 | 0.92 | 4313 |

We can clearly see that the Linear Support Vector Classification performs the best out of all the different models that we tried. It achieves nearly **93% accuracy** while classifying the sentiment of a tweet.

Although it should also be noted that the **BernoulliNB Model** is the fastest to train and predict on. It also achieves 86% accuracy while calssifying.

## V. Model Application

Using AI to generate three news, respectively negative, positive and negative, and apply the pre-trained model to classify them. All three test cases passed successfully.

## VI. Appendix

**Appendix 1**

数据来源:公众号“月小水长”

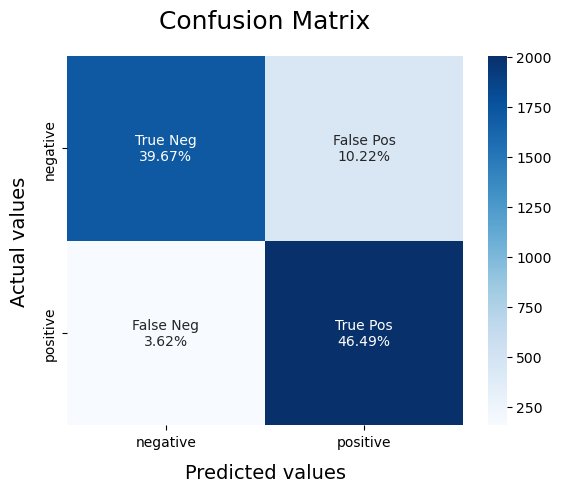
<https://pan.baidu.com/s/10eM4wf5Wqo8jHwANEzIP9g>

**Appendix 2**

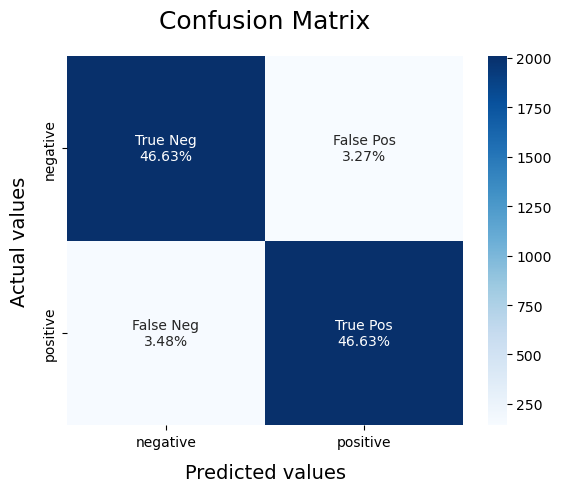
停用词表

<https://github.com/goto456/stopwords>

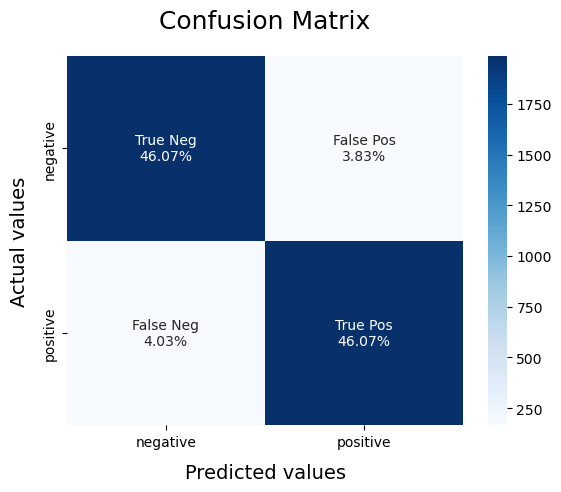
**Appendix 3.1 Bernoulli Naive Bayes**



**Appendix 3.2 Linear Support Vector Classification**



**Appendix 3.3 Logistic Regression**



**Appendix 4** 模型文件

Sentiment-BNB.pickle

Sentiment-LR.pickle

vectoriser-ngram-(1,2).pickle

**Appendix 5** 测试用例

# Using AI to generate three news, respectively negative, positive and negative

text = ["【武汉疫情形势：医疗物资短缺问题凸显】览富财经 2024 年 11 月 9 日讯，武汉部分医疗机构表示，随着新冠疫情的反复，医疗物资尤其是防护用品和特定药品的短缺问题日益严重。尽管各方努力调配资源，但目前仍有不少医院面临物资匮乏的困境。相关部门呼吁社会各界加大对武汉医疗物资的支援力度，以尽快缓解这一紧迫局面。",

"武汉市卫健委关于新冠疫情防控的重大喜讯通报：专家高度赞誉武汉防控成就斐然，全城抗疫防线坚如磐石。目前，武汉市通过全面且高效的大规模疫苗接种、科学精准的防控策略以及全体市民的积极配合，已成功将新冠疫情牢牢控制。全市各个社区不仅实现了长时间的零感染，而且居民生活秩序井然，经济活动也在安全有序的环境中快速复苏。同时，医疗系统的应急响应能力达到了前所未有的高度，随时能够应对各种突发状况，为市民的生命健康提供了坚实保障。武汉，正以昂扬的姿态成为全球抗疫的典范城市。",

"# 武汉新冠疫情反弹 #【武汉市卫健委通报疫情，局部地区疫情形势严峻复杂】据武汉市卫生健康委员会官网通报，近期武汉部分区域出现新冠疫情强烈反弹现象，尽管迅速采取了严格的管控措施，但新增病例仍在持续攀升。多个小区被列为高风险地区，居民被严格限制出行，生活陷入极大的困境。目前已发现疫情有向周边区域扩散的趋势，防控工作面临前所未有的巨大压力，且医疗资源紧张，部分患者无法得到及时有效的救治。到目前为止，疫情传播风险极高且难以控制，整个城市笼罩在疫情的阴霾之下。"]