西安電子科技大學
XIDIAN UNIVERSITY

人工智能学院

智能数据挖掘课程作业报告

对 Twitter 上推文进行情感分析

姓名：杨文韬

学号：18020100245

班级：1920012

2022 年 1 月 21 日

目录

对 Twitter 上推文进行情感分析

1 绪论

本章节是关于新冠肺炎变种 Omicron 在 2021 年 12 月份时 Twitter 推文的情感分析 (需要科学上网)，目的在于利用数据挖掘了解 Twitter 用户的感觉。用 `df.info()` 查看数据集各列的数据类型，空值和内存占用情况如下：

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 48168 entries, 167 to 45168
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                48168 non-null  int64
 1   user_name         48168 non-null  object
 2   user_location     37356 non-null  object
 3   user_description  45357 non-null  object
 4   user_created      48168 non-null  datetime64[ns]
 5   user_followers    48168 non-null  int64
 6   user_friends      48168 non-null  int64
 7   user_favourites   48168 non-null  int64
 8   user_verified     48168 non-null  bool
 9   date              48168 non-null  datetime64[ns]
 10  text              48168 non-null  object
 11  hashtags          35291 non-null  object
 12  source            48168 non-null  object
 13  retweets          48168 non-null  int64
 14  favorites         48168 non-null  int64
 15  is_retweet        48168 non-null  bool
dtypes: bool(2), datetime64[ns](2), int64(6), object(6)
memory usage: 5.6+ MB
```

2 推文分析

omicron.csv 数据集中有 15168 条推文，在 user_description, user_location 和 hashtags 条目有一些空值，分别表示一些用户没有简介、一些用户没有位置信息以及一些推文没有哈希标签。用 `df[df.duplicated()]` 可以发现没有重复推文。

2.1 不同日期推文分析

在数据中有两列 ('date', 'user_created') 表示时间，将其转化为 pandas datatime 以便于分析。统计得到不同日期的推文数量柱状图如图 1 所示。

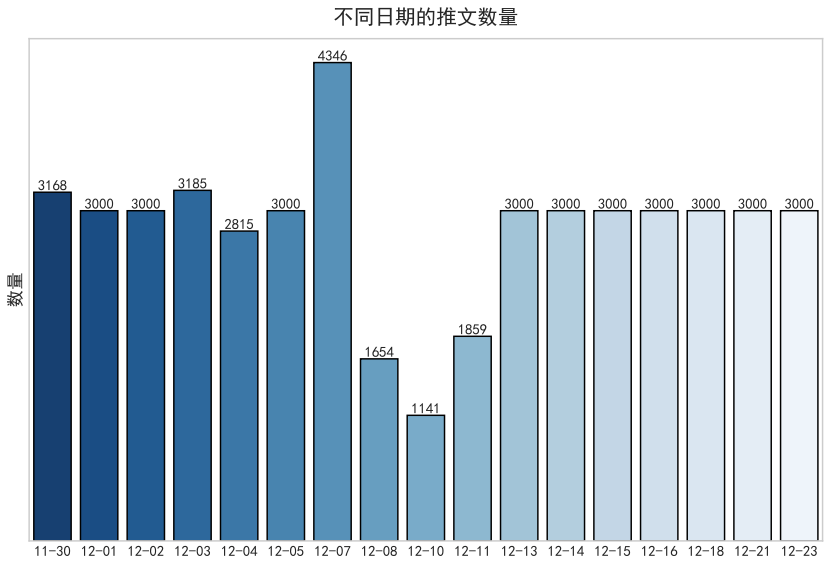

图 1: 不同日期的推文数量柱状图

可以看出关于 Omicron 推文数量最高的在 12 月 7 日，值得注意的是，12 月 13 日和 12 月 23 日之间的推文数量是恒定的，都是 3000 条，这可能是由于每天的推文爬取限制。

2.2 24 小时内不同时间的推文分析

同理可以统计得到 24 小时内不同时间的推文数量柱状图如图 2 所示。

可以看到从 00:00 到 6:00 的有上升趋势，从 6:00 到 9:00 有下降趋势，以及从 9:00 到 12:00 的僵硬上升。

2.3 不同国家推文分析

通过代码 `df['user\_location'].value\_counts()` 可以发现用户位置不总是国家，我们可以通过 pycountry 这个库提取出国家。创建新列表属性“国家”，找不到国家则为空值。接着统计推文数大于 30 的国家推文数，不同国家的推文数量柱状图如图 3 所示。

2.4 推文点赞分析

通过 `fav\_tweets = df[['text','favorites']].sort\_values('favorites', ascending=False)` 可以将推文按点赞数降序排序，然后通过 `fav\_tweets['text'].iloc[:10].values` 找出点

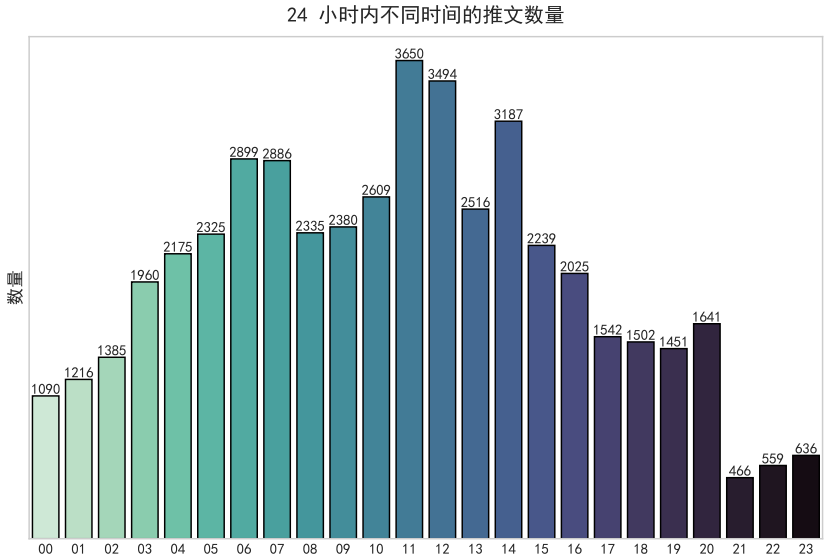

图 2: 24 小时内不同时间的推文数量柱状图

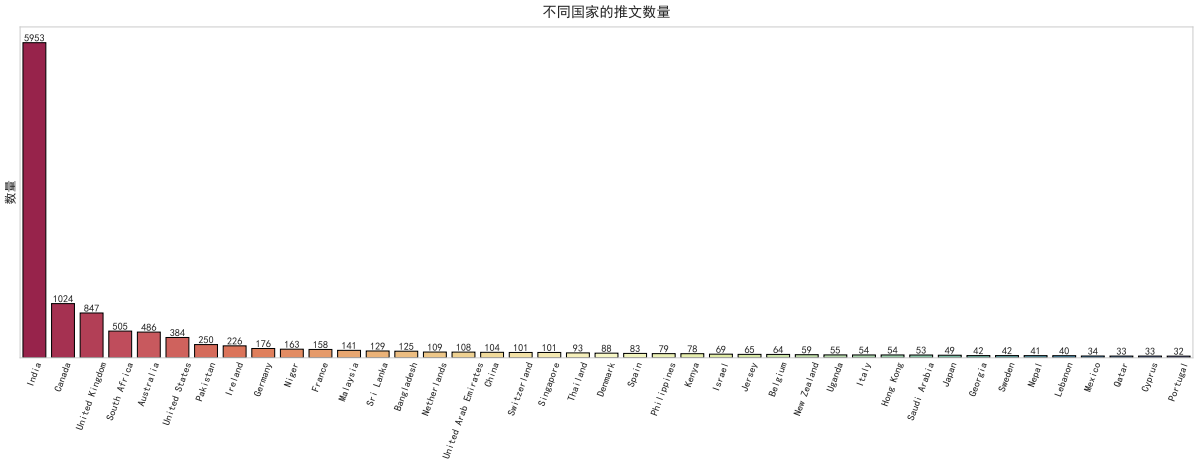

图 3: 不同国家的推文数量柱状图

赞数前 10 的推文内容如图 4 所示。

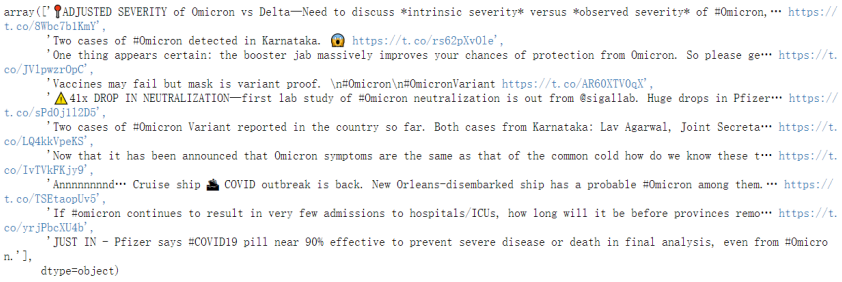

图 4: 点赞数前 10 的推文

2.5 推文转推分析

通过 `retweets = df[['text','retweets']].sort\_values('retweets', ascending=False)` 可以将推文按转推数降序排序，然后通过 `retweets['text'].iloc[:10].values` 找出转推数前 10 的推文内容如图 5 所示。

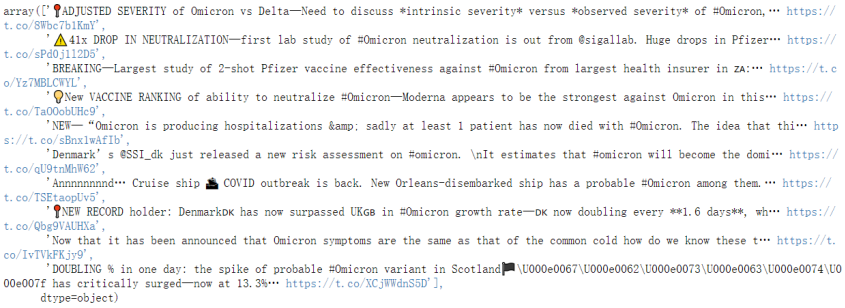

图 5: 转推数前 10 的推文

可以发现点赞数和转推数最多的推文都是由美籍华裔健康经济学家 Eric Feigl-Ding 发布的关于 Omicron 病毒的评价，这与他庞大的粉丝数量是密切相关的。

3 推文情感分析

首先对推文数据用自定义函数进行数据清洗，主要是去掉表情和一些特殊字符并改成小写。清洗后文本长度如图 6 所示 •，观察到一些推文单词数很少，少于 10 个单词的文本柱状图如图 7 所示。在后续分析中将去掉少于 5 个单词的推文。

接下来用 NLTK 和 TextBlob 分别进行情感分析，我将绘制每种方法的词云，首先需要添加一些 stopwords，指所有类别推文都共享的词，比如’covid’ 和’omicron’ 等等。

3.1 正面推文

正面推文词云如图 8 所示。综合两种方法得到的综合正面推文词云如图 9 所示。综合正面推文如下：

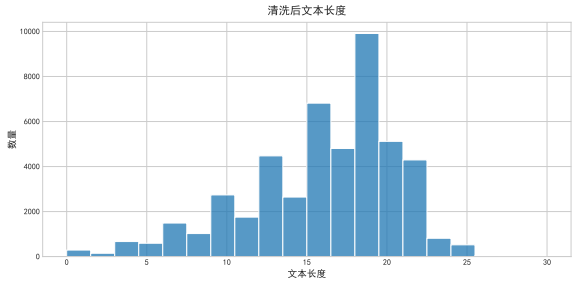

图 6: 清洗后文本长度

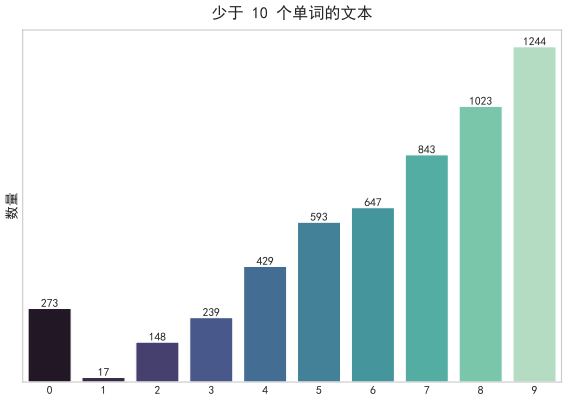

图 7: 少于 10 个单词的文本柱状图

```
array(['wonderful project sir this project looks very interesting i am
    interested and i will support this',
       'outstanding thread from one of the best cc omicron covid19',
       'good join the club happily membership is growing omicron',
       'good evening dear friend be careful with omicron',
       'nice looks like you are better protected against omicron then me',
       'omicron alert level4 help raise awareness bettermasks good fit amp
          filtration is key ffp2 cloth mask good fit over',
       'happy holidays mani xmas omicron',
       'this is nuts do better health care do better testingtestingtesting
          covid19 omicron',
       'discover and share your best business content download the best app
           gt theeconomist',
       'wise words from a wise man omicron'], dtype=object)
```

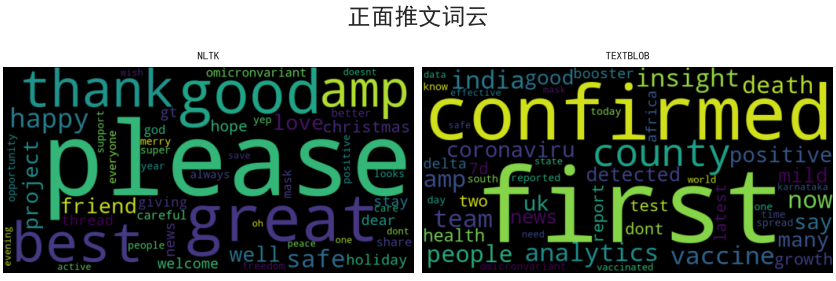

图 8: 正面推文词云

3.2 负面推文

负面推文词云如图 10 所示。综合两种方法得到的综合负面推文词云如图 11 所示，可以看出这些词确实都非常负面。综合负面推文如下：

```
array(['omicron is less severe buy risk',
       'omicron you crazy son of a bitch',
```

综合正面推文词云

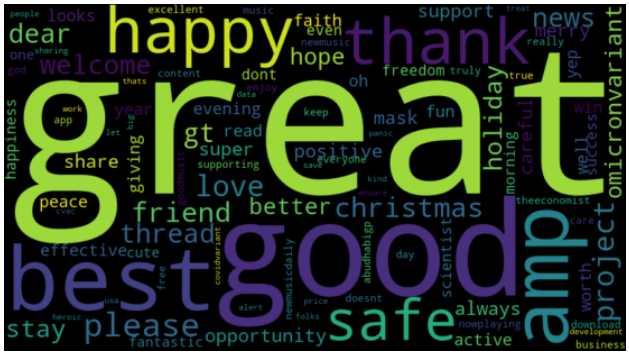

图 9: 综合正面推文词云

```
3       'hospitalizations lowest death rate lowest severity lowest ippudanna
            aa fear mongering aapesi lock down l',
4       'flu omicron serious chance of serious illness',
5       'omicron sorry omicon covid scam',
6       'they mad because no one is scared of omicron',
7       'omicron bad could be worse get boosted good night',
8       'open door fucker via f1finale idiot omicron',
9       'america ohio americasgottalent dead another death jayjayphillips
           was discovered dead',
10      'shocking racism omicron africa europe'], dtype=object)
```

负面推文词云

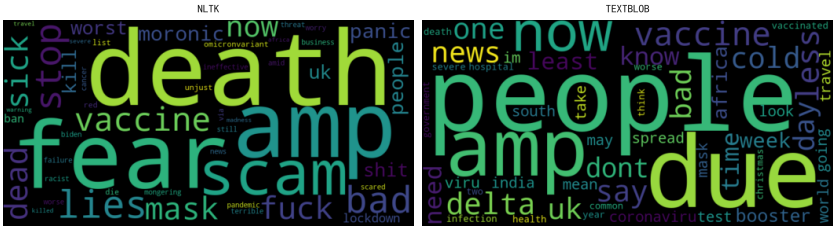

图 10: 负面推文词云

综合负面推文词云

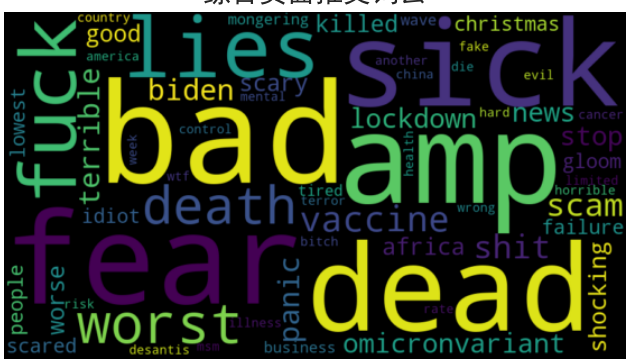

图 11: 综合负面推文词云