关于疫苗接种倾向
与疫情发展趋势的研究

第二组
迟熙、潘莹莹、王姿文、吴宇龙、袁大卫
2021 年 6 月

摘要

新冠肺炎疫情对我国经济社会产生了深刻影响，引起经济下滑、失业上升、物价波动、外贸走弱。但在党的坚强领导和人民团结一致的共同努力下，经济社会迅速企稳，各项事业回归正规，人民生活恢复正常。而全球疫情仍不容乐观，需要严防死守和科学应对。为彻底控制疫情，各国将大规模接种疫苗作为首选，中国在这方面也积极努力，并实施全民免费接种。在疫苗普及的过程中，民众对疫苗接种的意愿与疫情发展的趋势成为了备受关注的热点话题。本文运用主题分析、情感分析、实证分析等方法，对民众接种新冠疫苗的意愿、态度和关注度随疫情发展变化的趋势进行了分析。结果显示：疫苗的研发在一定程度上可以控制疫情的蔓延；随着疫情程度加深，对新冠疫苗的关注和讨论度越来越高；公众对疫苗的评论越来越偏中性和积极，对疫苗的接受程度越来越高。因此，本文认为疫苗的研发和接种对于遏制新冠肺炎疫情具有重要意义，而增加疫苗的有效性对于增加公众对疫苗的接受度具有根本性作用。

关键词：新冠肺炎疫情；疫苗接种倾向；文本分析；多项式回归

第 1 章研究背景

1.1 新冠肺炎疫情对中国宏观经济的影响[①]

1.1.1 经济增长

图 1.1 是近十年来，我国国民总收入、国内生产总值以及人均国内生产总值增长率的示意图。从图中所示趋势可以看出，2020 年，虽然增长率均大于 0，但是处在十年以来的最低位，三个国民经济核算指标的主要增长率均处于 2%上下。图 1.2 描绘了近十年来，我国三大产业增加值的增长率趋势。从图中可以看出，第一产业增加值增长率保持稳定，二三产业则遭受较大冲击。这与疫情防控期间的防控措施（如企业延迟复工复产、倡导人员减少出行、实行居家隔离等政策措施）密切相关。

为控制疫情传播，2020 年初，大部分制造业企业延迟开工，企业停止生产，供给能力下降，同时市场需求规模也在缩小，很多企业无法承受成本的持续输出，纷纷倒闭，这对第二产业造成了较大冲击。为减少接触范围、阻断传播链，倡导居民进行居家隔离，务工人员无法外出务工，而第三产业以服务业为主，服务人员是关键要素，所以对第三产业也产生了较强的冲击。第一产业主要是农林牧渔业，农产品是生活必需品，即使居民减少外出，但也不会减少对农产品的消费，所以第一产业增加值增长率保持平稳，并未受到明显的负面冲击。

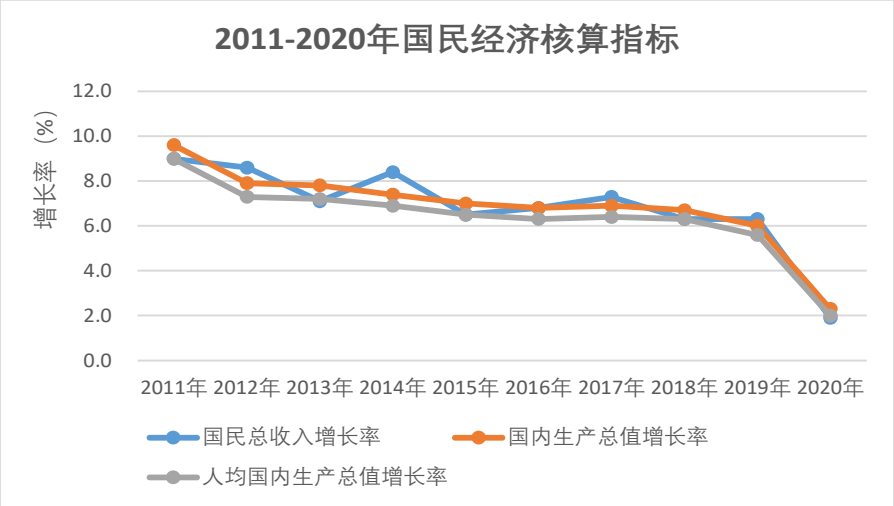

图 1.1 国民经济核算指标

①本章节图表数据来源：国家统计局。

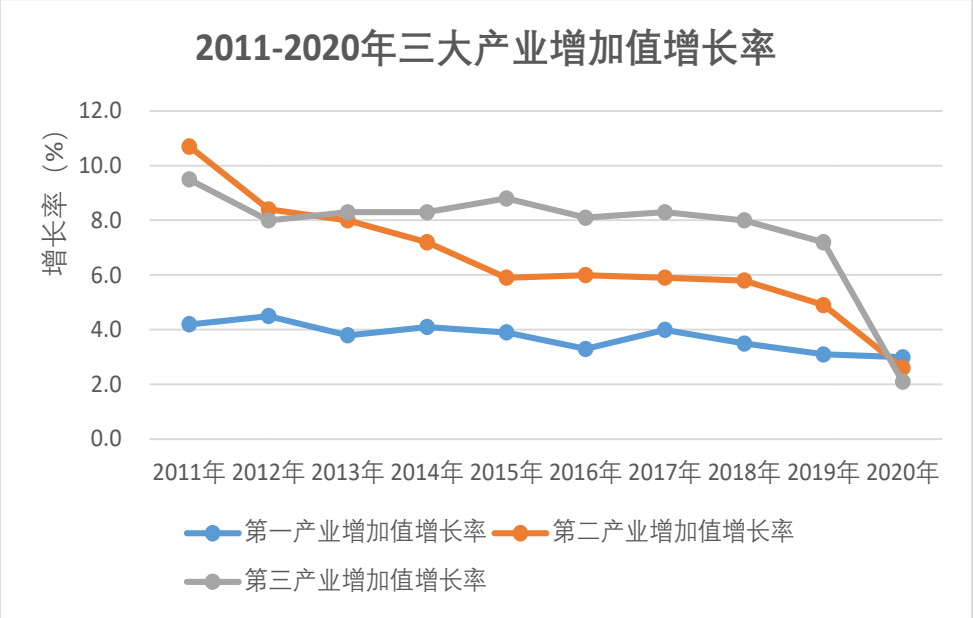

图 1.2 三大产业增加值增长率

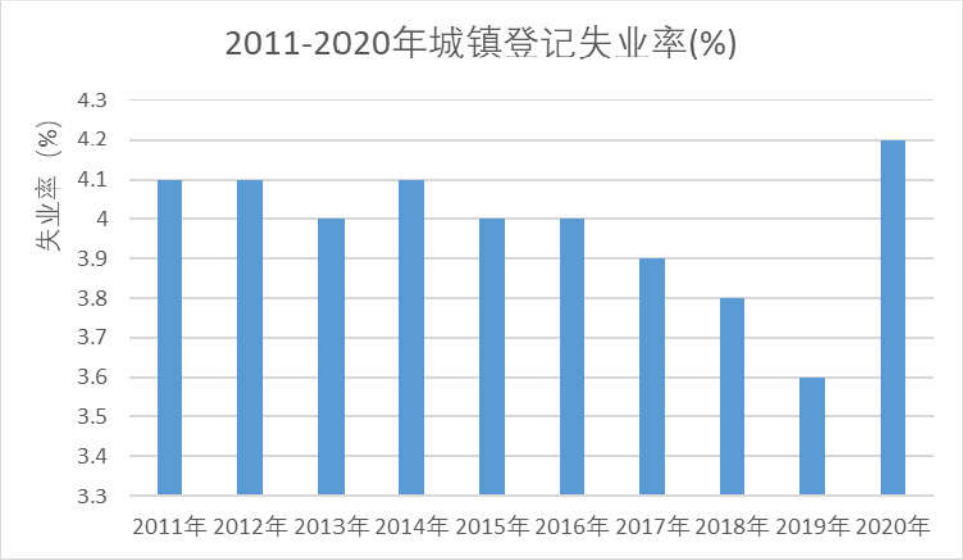

图 1.3 城镇登记失业率（年度）

1.1.2 城镇就业

图 1.3 和 1.4 反映了我国城镇失业率的变化趋势。从年度数据来看，近十年以来，我国城镇登记失业率在 2020 年达到 4.2%的峰值，表明新冠肺炎疫情对我国就业产生了一定负面冲击。从月度数据来看，2020 年初，受到疫情影响，城镇失业率明显上升，3 月份达到峰值，之后逐渐下降，2020 年 11 月，恢复到 2019 年同期水平。

失业率的变化与疫情防控态势密切相关。从劳动力的供给端来看，2020 年初，疫情形势十分严峻，不论是响应国家号召，还是出于主观的自我保护，多数服务业的务工人员都居家隔离，短期内不再外出务工，劳动力的供给下降。从劳动力的需求端来看，受到疫情冲击和延迟复工复产的影响，企业无法生存，纷纷倒闭，对劳动力的需求下降。因此，2020 年的第 1、2 季度，就业率维持在较高位置。2020 年的下半年，我国疫情逐渐得到有效控制，疫苗的研发也取得突破性进展，企业复工复产，逐渐恢复至疫情之前的水平，需要更多劳动力，；同时之前居家隔离的待工人员也开始外出务工，所以城镇失业率逐渐下降，回到疫情之前的正常水平。

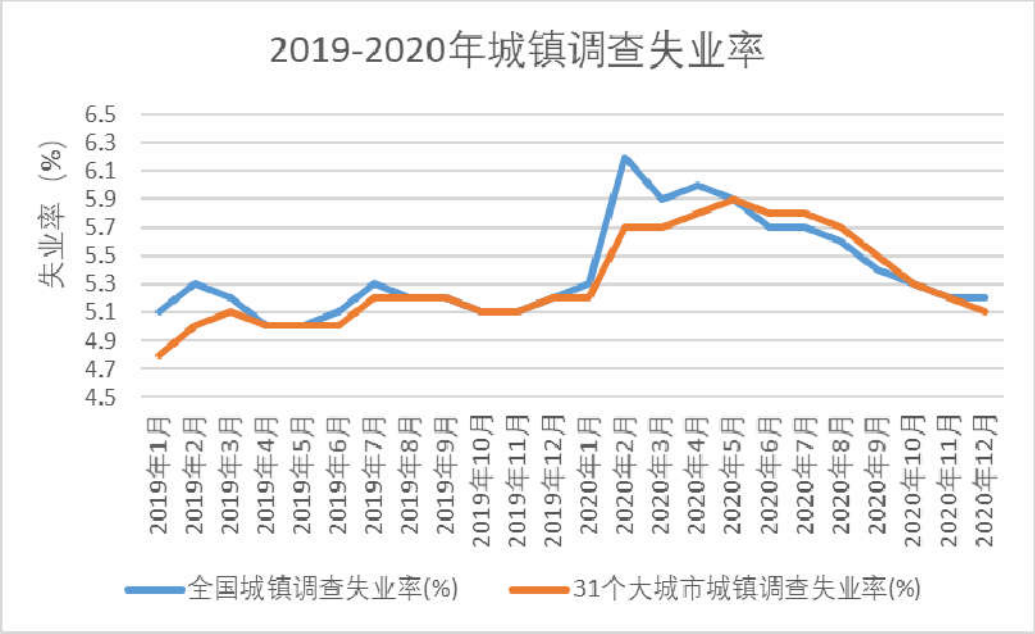

图 1.4 城镇调查失业率（月度）

1.1.3 价格指数

图 1.5 和图 1.6 描绘了我国居民消费价格指数增长率的变化趋势。从年度数据来看，2020 年总体，物价水平保持稳定，与往年物价指数增长率差距不大。从月度数据来看，2020 年初，受到疫情影响，物价水平维持高位，之后逐渐下降，2020 年下半年，基本维持在正常水平。

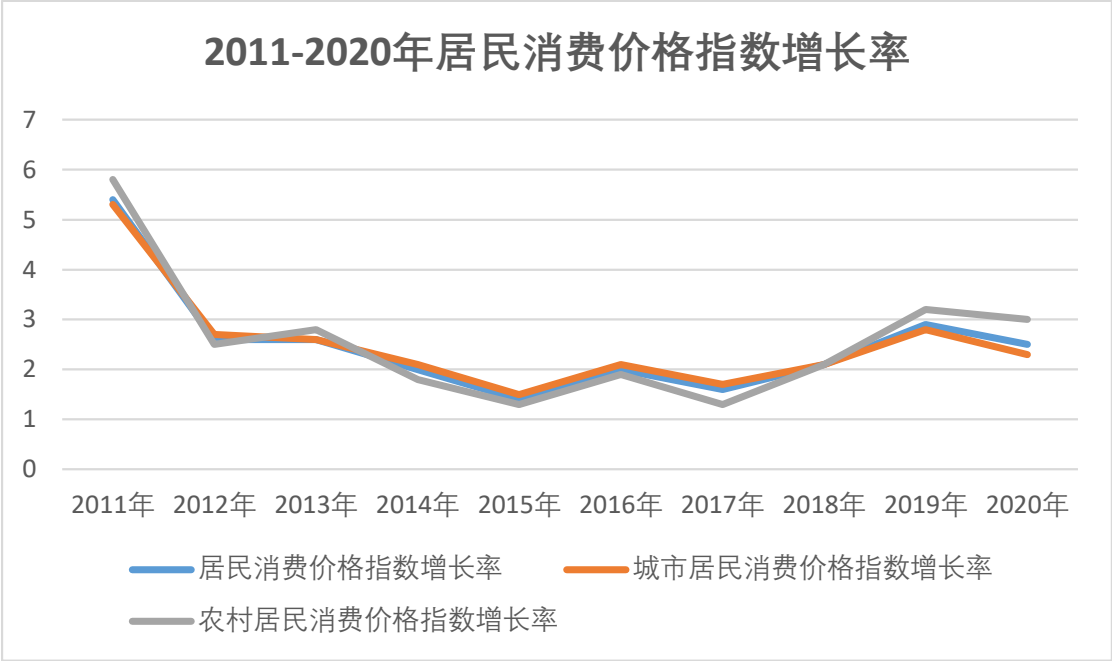

图 1.5 居民消费价格指数增长率（年度）

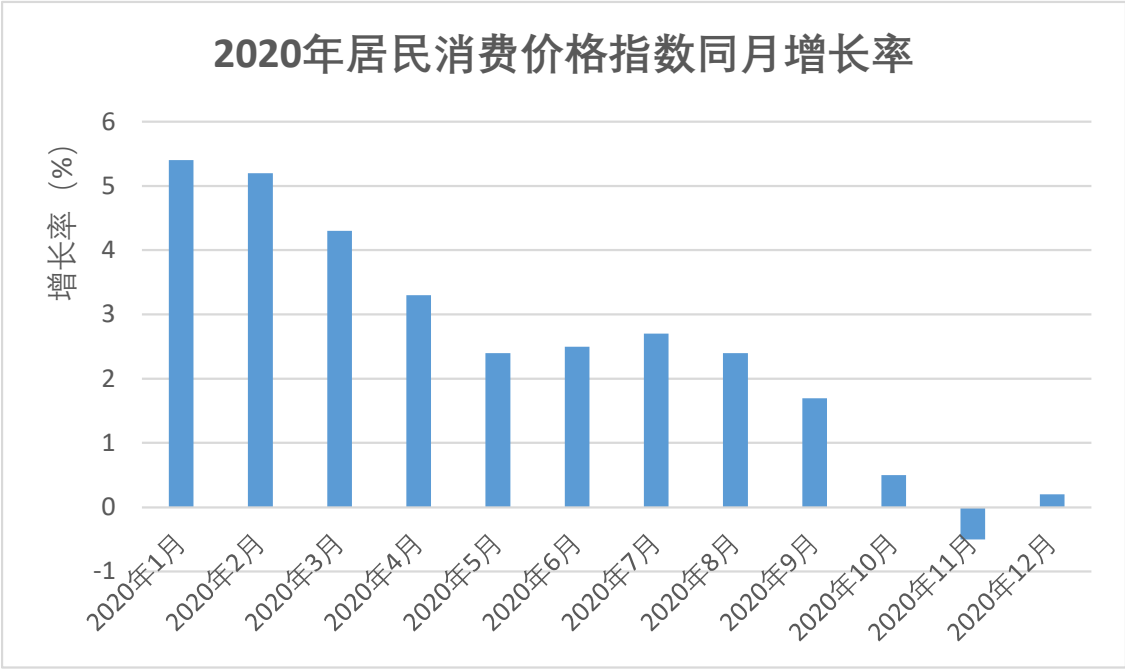

图 1.6 居民消费价格指数增长率（月度）

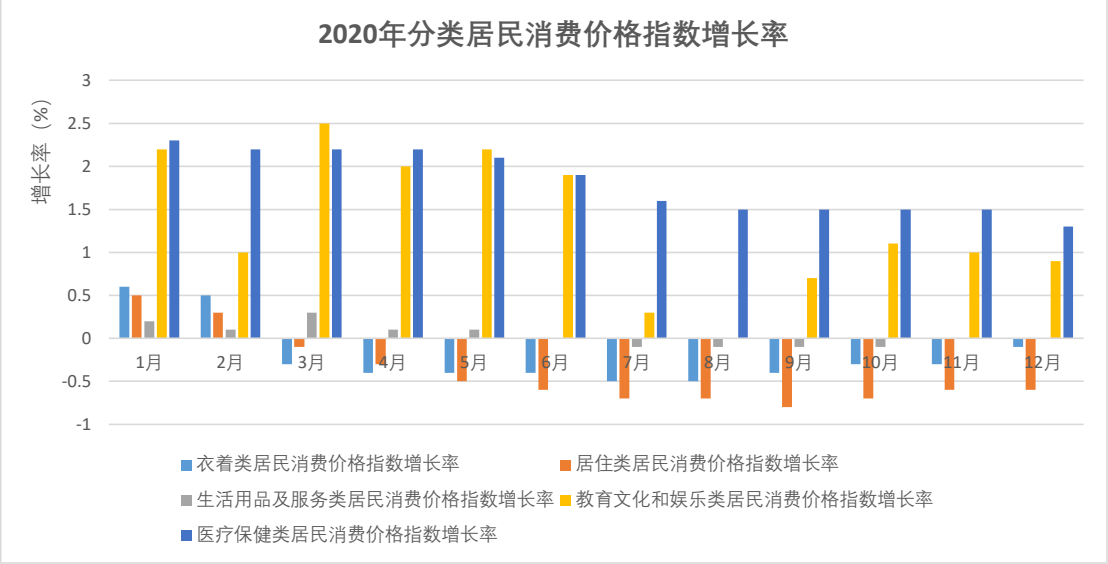

图 1.7 2020 年分类居民消费价格指数增长率（月度）

得利于我国有效的疫情防控措施和效果，2020 年整体，我国基本维持了物价稳定，并未出现长期持续的通胀或通缩。但是分月份来看，在第一季度，我国的物价水平出现了明显上升，这一方面是由于企业未能按时复工复产导致供给能力下降，另一方面也可能因为居民出于对产品供给能力的担忧而大量囤货，供求的双重冲击使得物价上升。从图 1.7，分类别的居民消费价格指数增长率（图中未包含食品烟酒类别，该类别受节假日效应影响较大，不具有可比性）来看，医疗保健类和教育文娱类消费价格指数的增长率均大于其他所列类别。这可能是由于以下原因：第一，疫情发展前期，由于本次疫情属于突发性的公共卫生事件，而且当时正值春节假期，我国多数地区医疗物资出现短缺的情况，这加剧了人们的恐慌心理，越难买越想买，从而将医疗物资价格推升至高点。第二，当时几乎所有学校均不允许线下开课，学生无法返校，只能通过线上进行学习，这难以满足一部分即将参加中考、高考的学生，所以这部分同学可能会在课外教学机构以及学习资料上花费更多。第三，无论是成人还是儿童，待在家中的时间比以往增加很多，大家会寻找一些娱乐方式，如制作美食、观看电影等等，这也会增加文娱消费。

1.1.4 对外贸易

疫情发展初期，为防控新冠肺炎疫情，世界各国相应出台有关贸易管制、国际运输和人员管制的相关政策，国际汇率不稳定，这对中国的对外贸易形成了冲

击，外贸企业也面临着严峻的生存挑战[①]。

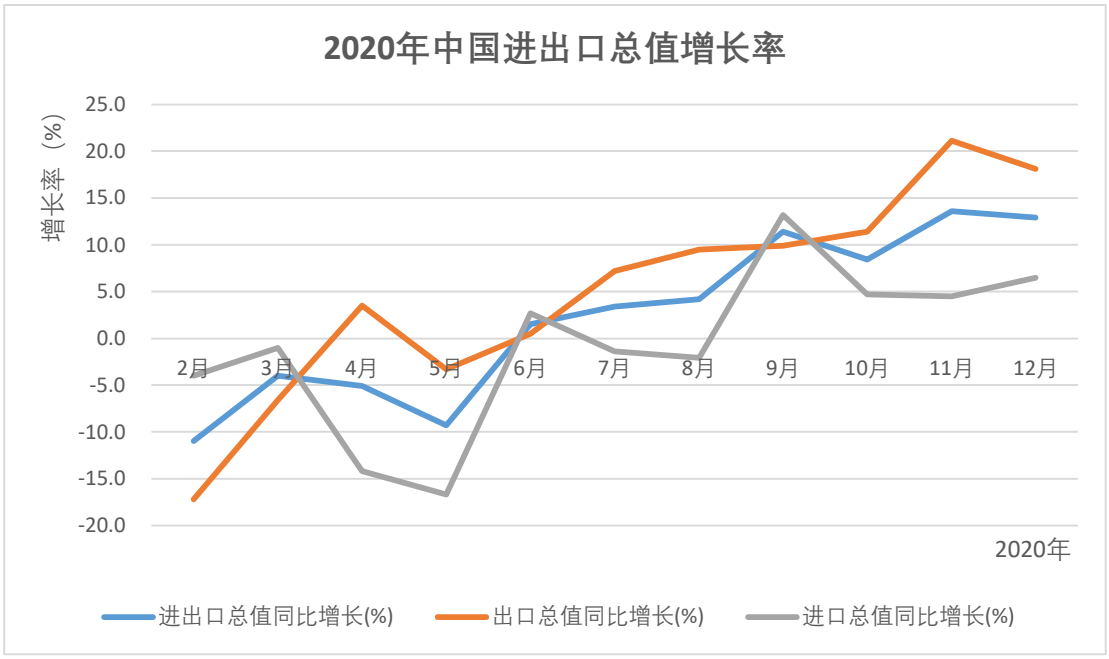

图 1.8 中国进出口总值同比增长率（月度）

中国是最先报道新冠肺炎疫情相关新闻的国家，2020 年初，很多国家对我国实行了进出口管制，增加了贸易壁垒，对我国的国际贸易产生了很大负面影响。在产品贸易方面，就国内的供给能力来看，为遏制疫情发展，我国采取延长春节假期，延后复工复产等一系列措施，这削弱了企业的供给能力，使得生产量大幅下降；就国外的需求规模来看，很多其他国家出台措施对中国出口产品实行管制，比如对货物、邮包、人员实行更加严格的检疫措施等，这也导致国外对中国出口产品的需求规模迅速下降。供给和需求的双重冲击使原本就面临严峻挑战的外贸企业雪上加霜。得益于政府有力的疫情防控政策，我国的疫情得到有效控制，企业加速复工复产，生产能力回升，对外贸易逐渐恢复正常水平。

从 2020 年我国进出口总值同比增长率可以看出，第一季度，我国的进口、出口以及进出口总值与前一年同期相比呈现负增长，经过第二季度的过渡和缓和，2020 年下半年，我国的对外贸易逐渐恢复正增长。

1.2 世界主要经济体新冠肺炎疫情的当下形势

①王铁山,张青.新冠肺炎疫情对我国外贸企业的影响及应对措施[J].经济纵横,2020(03):23-29

图 1.9 显示了截至北京时间 2021 年 5 月 25 日 9 时 55 分的累计确诊数量排名世界前十的国家的疫情形势。从上图可以看到，美国、印度和巴西的确诊人数超过千万，其中美国的累计确诊人数和累计死亡人数最多，累计确诊超过三千万，累计死亡超过五十万，印度的累计确诊人数排在第二位，超过两千六百万，累计死亡超过三十万，而且仍有持续增加的趋势。由此可见，新冠肺炎疫情仍然在世界范围内持续扩散，国内外时刻面临着严峻的抗疫挑战，疫情的防控仍然十分必要。

地区	累计确诊	排序	累计死亡	排序	病死率	排序
美国	33,140,137	1	590,184	1	1.78%	87
印度	26,752,447	2	303,720	3	1.13%	145
巴西	16,120,756	3	449,858	2	2.79%	38
法国	5,917,397	4	108,687	8	1.83%	81
土耳其	5,194,010	5	46,446	19	0.89%	162
俄罗斯	5,009,911	6	118,801	7	2.37%	54
英国	4,464,900	7	151,904	5	3.40%	28
意大利	4,194,672	8	125,335	6	2.98%	36
德国	3,670,423	9	88,039	9	2.39%	51
西班牙	3,647,520	10	79,711	11	2.18%	63

图 1.9 世界累计确诊前十国家的疫情形势[①]

图 1.9 显示了截至北京时间 2021 年 5 月 25 日 9 时 55 分的累计确诊数量排名世界前十的国家的疫情形势。从上图可以看到，美国、印度和巴西的确诊人数超过千万，其中美国的累计确诊人数和累计死亡人数最多，累计确诊超过三千万，累计死亡超过五十万，印度的累计确诊人数排在第二位，超过两千六百万，累计死亡超过三十万，而且仍有持续增加的趋势。由此可见，新冠肺炎疫情仍然在世

①数据来源：https://ncov.dxy.cn/ncovh5/view/pneumonia

界范围内持续扩散，国内外时刻面临着严峻的抗疫挑战，疫情的防控仍然十分必要。

第 2 章研究问题与研究意义

虽然与 2020 年相比，对于新冠病毒，我们已经从一无所知到认识其真正的结构和作用机理，再到研制出多种疫苗。但从上述新冠肺炎疫情对经济的影响以及疫情发展趋势来看，无论是国内还是国外，仍然时刻面临着严峻的防疫抗疫挑战。只有真正地实现世界范围的群体免疫，才能更好地应对新冠病毒的侵扰。

群体免疫(group immunity)是指人群对传染性疾病的抵抗力。群体免疫水平高，表示群体中对传染疾病具有抵抗力的人群百分比高。疾病发生流行的可能性不仅取决于群体中有抵抗力的个体数，而且与群体中个体间接触的频率有关。如果群体中有 70%-80%的人群有抵抗力，就不会发生大规模的爆发流行。根据相关研究表明，人群中新冠肺炎疫苗接种率达到 67%，新冠肺炎感染发病率才会下降[①]。可见，目前实现群体免疫的最好方法就是通过注射疫苗，形成免疫屏障。一旦有局部出现感染个体，能够及时阻断传播链，防止疫情向外扩散。

然而，不论是国内还是国外，一些民众对于注射疫苗存在或多或少的疑虑或抵触情绪。这可能出于以下原因：第一，尽管世界各国科学家已经对新冠病毒的结构、机理展开全面研究，但是毕竟只持续了一年左右，目前是否真正地看清病毒的真面目尚未可知；第二，新冠病毒疫苗是新研制出来的，尽管已经经过临床试验才能上市，但是民众对疫苗的安全性仍然存在顾虑；第三，新冠病毒具有高变异性的特征，接种一种疫苗无法保证对变异病毒免疫，而且无法保证疫苗的实际作用期，所以民众对疫苗的有效性存在疑虑。国内外也有一些针对民众接种疫苗意向的研究，刘晓曦等运用“3Cs”模型通过对全国 34 个省、市、自治区的 18~59 岁居民新冠肺炎疫苗犹豫的调查，评估了人口学特征及“3Cs”模型各维度（信任、自满、便利）对公众新冠肺炎疫苗犹豫的影响。结果发现，“自满是疫苗犹豫的危险因素，信任是疫苗犹豫的保护因素，提示对疫苗安全性、有效性等不信任以及对新冠肺炎严重性和新冠肺炎疫苗必要性自满会影响新冠肺炎疫苗接种意愿”[②]。其研究发现与已有结果基本一致：“感知到的疫苗安全性、有效性以及对提供疫苗的系统的信任越高，接种疫苗越不犹豫；感知到的疫苗接种的必要性以及感染新冠肺炎的严重性越高，接种疫苗越不犹豫；感知到的接种疫苗便利程度越高，接

① RANDOLPH HE, BARREIRO LB. Herd immunity: understanding COVID-19[J]. Immunity, 2020,52（5）:737-741.

②刘晓曦,戴俊明,陈浩,等.基于“3Cs”模型的公众新冠肺炎疫苗犹豫影响因素的横断面调查[J/OL].复旦学报(医学版):1-6[2021-05-16].http://kns.cnki.net/kcms/detail/31.1885.r.20210510.2048. 006.html.

种疫苗越不犹豫”[①②]。

结合已有研究以及疫苗研发与接种效果等现实因素，要消除民众接种疫苗的疑虑，使民众意识到接种疫苗的必要性，提高疫苗接种率，需要从以下角度考虑。其一是真正了解民众对于接种疫苗的想法，针对具有不同想法和不同需求的群众对症下药，最大可能回答民众的问题，打消民众的顾虑；其二是使民众充分了解新冠肺炎已经对经济生活造成的影响以及如果不建立群体免疫屏障，日后可能带来的严重后果，使民众自愿接种疫苗。

在接下来的文章中，本文根据国外 reddit 论坛上网友有关疫苗接种的评论，通过文本分析挖掘随着时间进程民众对于疫苗接种的意愿和倾向变化，并尝试拟合这一变化与客观世界疫情发展趋势的关系，通过实证检验分析疫苗接种倾向对于疫情发展的影响。最后根据得出的结果提出切实可行的抑制疫情发展以及提高民众接种疫苗意愿的建议。由此可见，这一研究不仅具有理论意义，而且可以应用到实际中解决现实问题。

① QUINN SC, JAMISON AM,AN J, et al. Measuring vaccine hesitancy, confidence, trust and flu vaccine uptake: results of a national survey of White and African American adults [J] .Vaccine,2019,37（9）:1168-1173.

② GONZÁLEZ-BLOCK M, ARROYO-LAGUNA J, RODRÍGUEZ-ZEA B, et al. The importance of confidence, complacency, and convenience for influenza vaccination among key risk groups in large urban areas of Peru [J] . Hum Vaccine Immunother,2021,17（2）:465-474.

第 3 章数据分析

本文采用两个网络公开数据集，分别为 Reddit Vaccine Myths 和 Coronavirus (COVID-19) Trend Data[①]和全球疫苗接种情况数据[②]。旨在透过数据分析了解疫苗接种倾向，以及不同国家的疫情发展趋势，最后得出结论以便应用到实际中解决现实问题。

3.1 数据介绍

对疫苗情感态度的数据为 2014 年至 2021 年欧美国家的网络知名论坛—Reddit 上与 Vaccine（疫苗）相关的数据，其数据获取使用到 Python 内的 praw（Reddit API Wrapper）来爬取数据，而数据集内所爬取到的数据不仅为 Reddit Post Data，还有疫苗相关的网络新闻以及疫苗相关的网络图片。本数据集共有 8 个变量，本文透过网络爬取的方式再新增 3 个文本变量，最后共得到 11 个变量用以分析。

3.2 分析方法

探索性数据分析（Exploratory Data Analysis）

以探索性数据分析对于不同变量有更深的了解，获取数据特性帮助数据预处理，以便后续分析。

多项式回归（Polynomial Regression）

一种线性回归模型，通过最小二乘法逼近给定数据的变化趋势。

长短时记忆网络（LSTM，Long Short-Term Memory）

一种时间循环神经网络，通过遗忘门、输入门和输出门构建起门循环玩单元并组建成神经网络，可以学习长期依赖与短期依赖的信息。因此可以用于时序信息的预测，但具有较差的可解释性。

① https://www.kaggle.com/therealcyberlord/coronavirus-covid-19-visualization-prediction/notebook#Coronavirus-(COVID-19)-Visualization-&-Prediction

② https://www.kaggle.com/gpreda/covid-world-vaccination-progress

主题模型（Topic Model）

使用 EM 算法中的主题模型来替数据作聚类，此为非监督式学习模型。旨在比较整体时间（2014 年至 2020 年）和新冠疫情期间（2020 年至 2021 年），人们对于疫苗的主题讨论有什么有趣的变化。

a. 布朗聚类（Brown Clustering）

布朗聚类是一种自底向上的层次聚类算法，基于 n-gram 模型和马尔科夫链模型。布朗聚类是一种硬聚类，每一个词都在且只在唯一的一个类中，使用到二叉树的概念，适合词汇聚类。

b. 隐含狄利克雷分布（Latent Dirichlet Allocation, LDA）

隐含狄利克雷分布（Latent Dirichlet Allocation, LDA）的理论基础是贝叶斯理论。隐含狄利克雷分布根据词的共现信息的分析，拟合出词-文档-主题的分布，进而将词、文本都映射到一个语义空间中。

隐含狄利克雷分布算法假设文档中主题的先验分布和主题中词的先验分布都服从狄利克雷分布，通过对已有数据集的统计，就可以得到每篇文档中主题的多项式分布和每个主题对应词的多项式分布，并据此来推断文档中主题的后验分布，得出聚类结果。

情感分析（Sentiment Analysis）

a. 极性（Polarity）

透过 Polarity 的情感评分，得到连续型数值评分，再以情感评分将连续型数值分为三个类别—正向情感（Positive）、负向情感（Negative）、中性情感（Neutral）。

b.主观性（Subjectivity）

透过 Subjectivity 连续型数值评分，越接近 1 则越主观，越接近 0 则越客观。

Coronavirus (COVID-19) Trend Data

3.3 实证分析

3.3.1 疫情确诊人数时序建模结果

首先，对于自 2020 年 1 月 22 日至 2021 年 5 月 30 日全球每日累计确诊人数和每日新增确诊人数进行了展示，如图 3.1 所示。

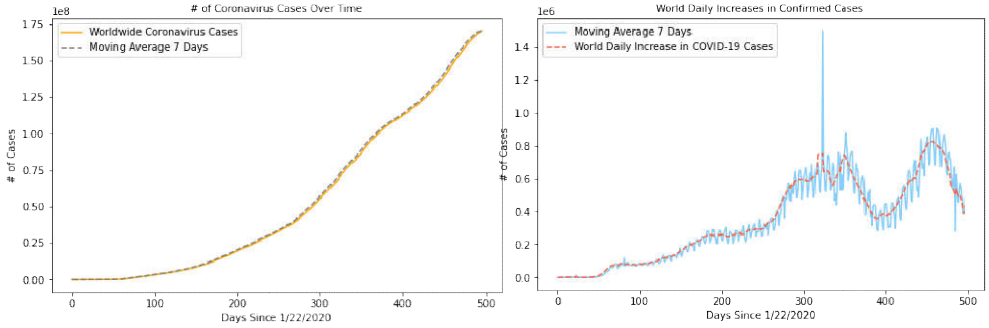

图 3.1

图 3.2 采用四次多项式回归的方法对累计确诊人数数据进行了拟合，模型平均绝对误差为 1.12×10^{7}，得到的零到四次方各阶系数分别为-2.65×10^{6}，1.48×10^{5}，-1.71×10^{3}，9.77，-1.14×10^{-2}。模型中训练集采用了自 2020 年 1 月 22 日起 450 天的数据，其余数据作为测试集，图 3.2 中橙色曲线展示了模型在测试集上的预测表现。

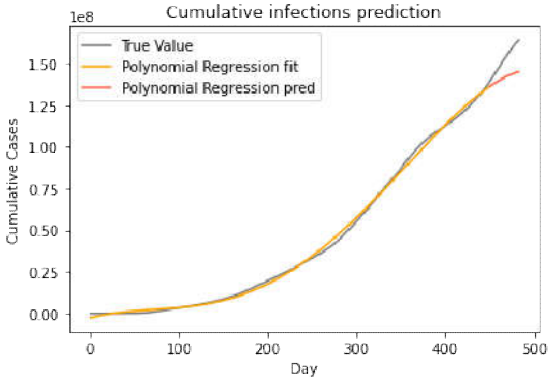

图 3.2

之后对于全球每日总确诊人数的时序数列尝试使用 16 层的长短时记忆网络，设定单位序列长度为 10，其网络架构为：

```
(lstm): LSTM(1, 16, batch_first=True)
(linear): Linear(in_features=160, out_features=1, bias=True)
```

同样采用前 450 天数据为训练集，得到的模型效果如图 3.3 所示，模型平均绝对误差为 1.11×10^{7}，拟合效果优于多项式回归，但预测效果与其类似。

由于每日总确诊人数的时序数据的曲线较平滑，神经网络捕捉的特征信息较少。将相同的网络应用在每日新增确诊人数时序变化中，得到的训练结果大大提升，模型平均绝对误差降到了 4.94×10^{4}，预测效果如图 3.4 所示，除部分异常离群点之外，**模型的表达能力和泛化能力均较强**。

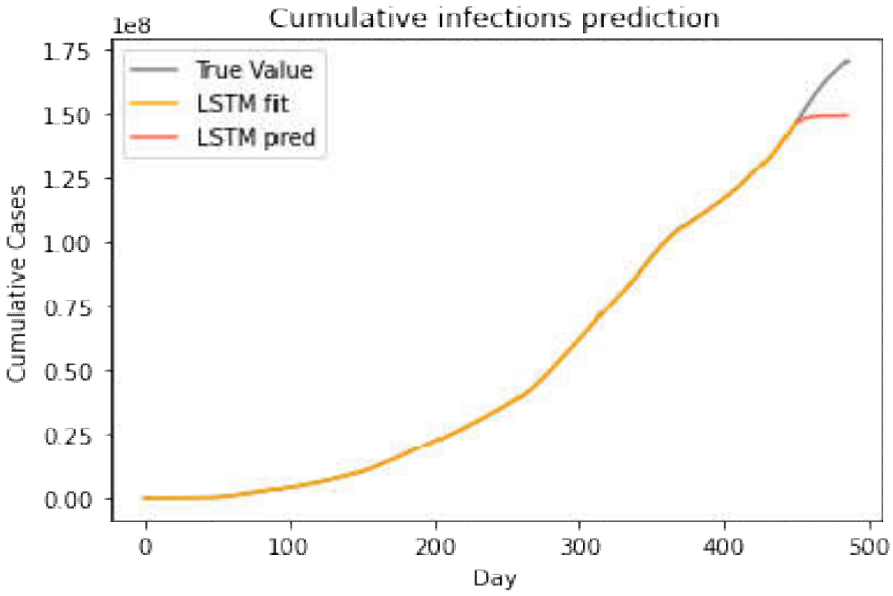

图 3.3

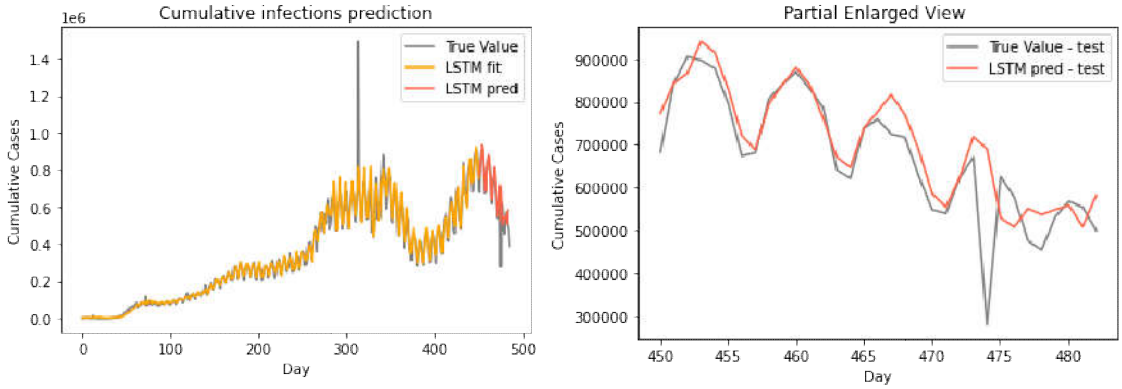

图 3.4

图 3.4 左图：每日新增确诊人数曲线 LSTM 拟合与预测情况；**右图：预测段放大展示**

3.3.2.对于疫苗情感态度数据分析

探索性数据分析（Exploratory Data Analysis）

Reddit 数据集共有 1510 笔数据、8 个变量，经由爬虫扩充量数据后转为 1510 笔数据、11 个变量：title、score、id、url、commns_num、created、body、timestamp、blog_text、ocr、text，以下叙述较为重要的探索性数据分析结果：

网络文章数量趋势：

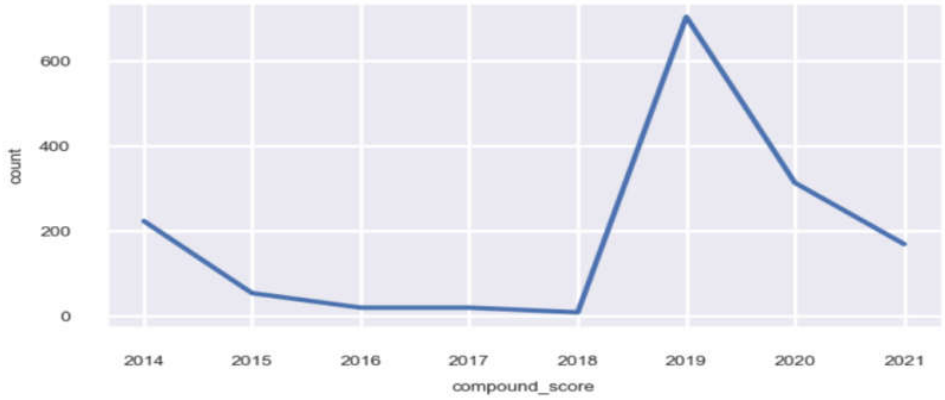

图 3.5 2014 年至 2021 年网络文章数量

created、timestamp 这 2 个变量都是文本创建时间，因此采用一个即可。图 3.5 可以看出文本数据在 2019 因为美国流感达到高峰，而新冠疫情期间（2020 年至 2021 年）虽然讨论量没有 2019 年来得多，但相对 2014 年至 2018 年也明显高出许多。

图 3.5 可以看出新冠疫情期间（2020 年至 2021 年）夏天的讨论数较低，因为新冠疫情在夏天较为趋缓。

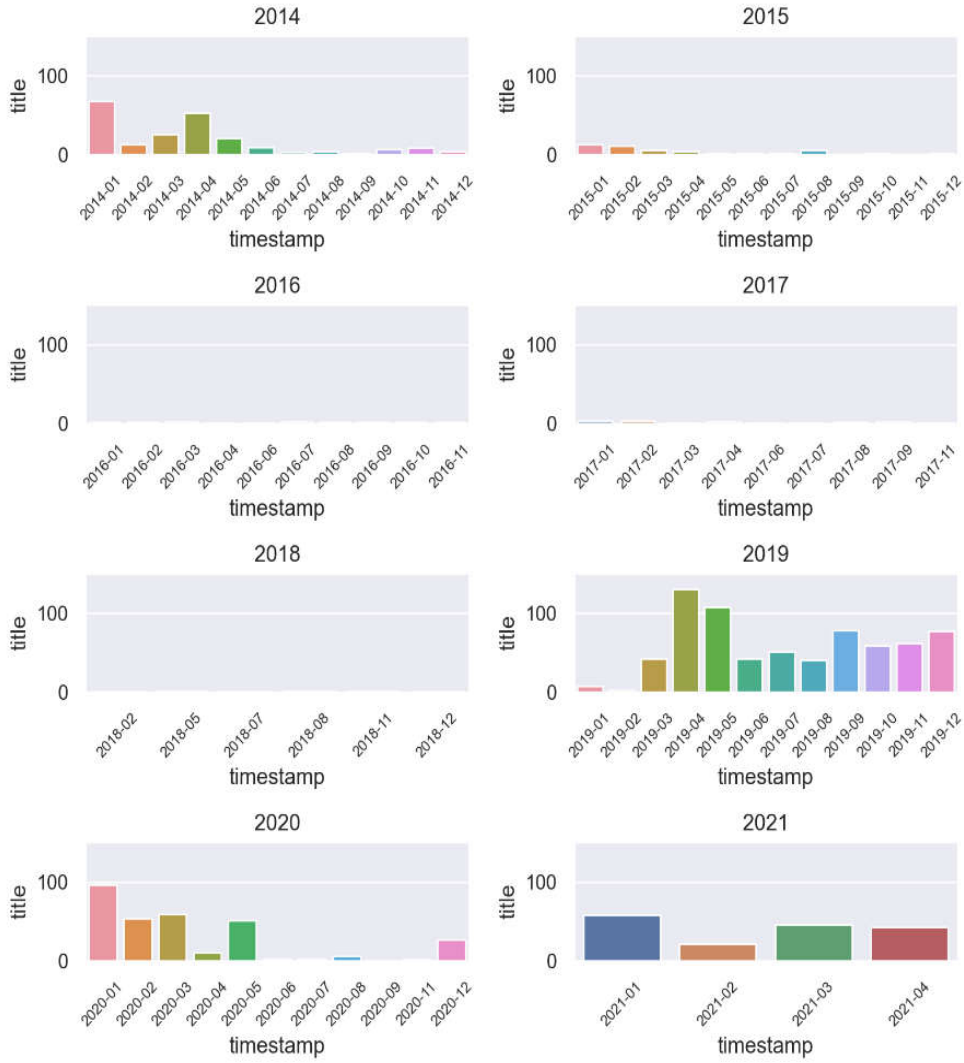

图 3.6 2014 年至 2021 年各月份网络文章数量

以 url 扩充文本数据：

与文本相关的变量中，title 有 1510 笔数据、body 有 1144 笔数据、url 有 452 笔数据。由于文本数据的完整性不足，因此决定以 url 来爬虫，扩充出 blog_text 和 ocr 变量，再将 title、body、blog_text、ocr 结合成为新的文章变量 text，以利后续文本分析使用。

blog_text 由 BeautifulSoup 爬取网络文章而得，而 ocr 由 urllib 和 cv2 爬取网络图片且下载到本地，再以 paddleocr 将图片转换为文字而获得新的文本数据，ocr 的取得如图 3.7（a）和图 3.7（b）所示。

' them. Love Protect them. Never injectthem. There s afe NO are vaccines H Shaken Baby Syndrome Chronic E ar Infections Death SIDS Seizures ADD Allergies Asth ma Autism Diabetes Meningitis and polio are caused b y adverse reactions to vaccine poisons.'

图 3.13（a）ocr 数据

图 3.7（b）ocr 原始爬虫取得图片

词云图（wordcloud）

上述扩充出新的文本变量后，后续主要以 text 来做文本分析，然而还是想确认与文本变量是否与新冠疫情相关，因此分别画出 title、body、blog_text、ocr、text 在新冠疫情期间（2020 年至 2021 年）的词云图，确认可以使用来作为后续分析。

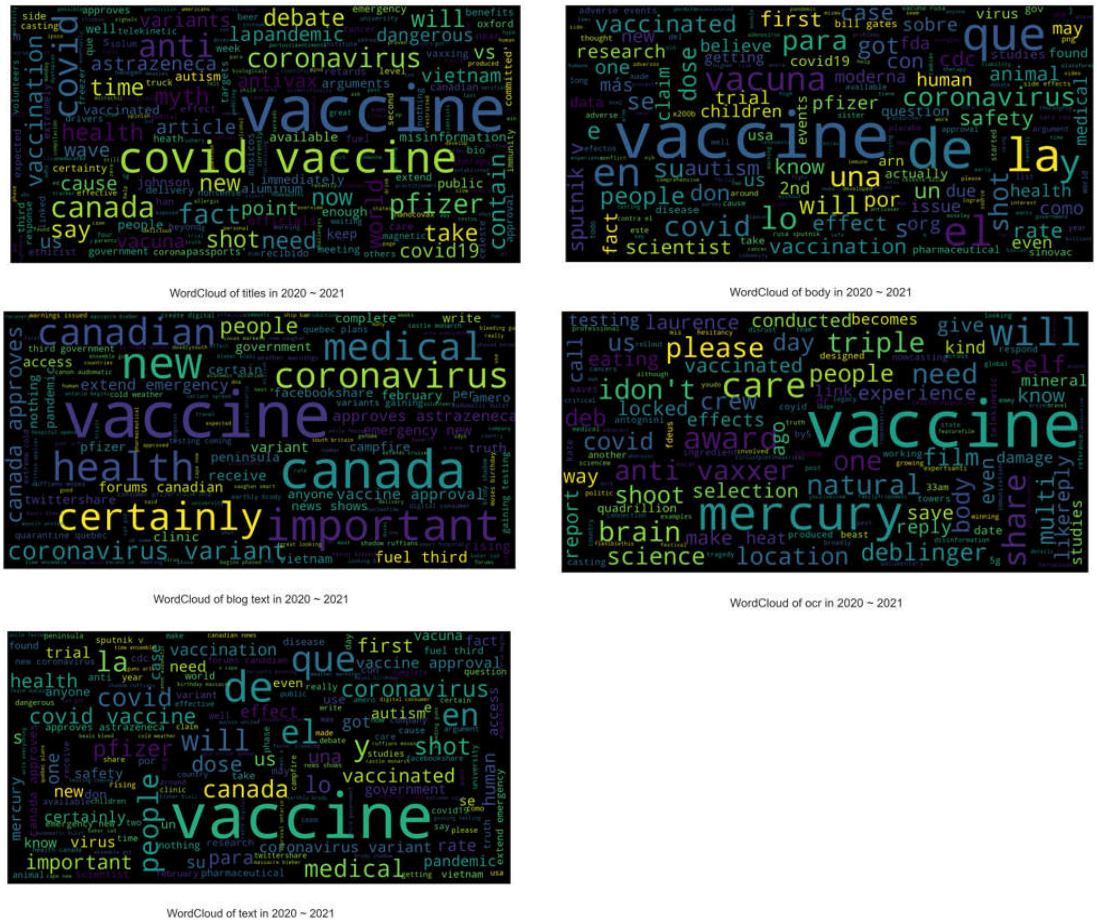

图 3.8 词云图

图 3.8 可以看出 title 的关键字为 covid，也涵盖 anti, fact, myth...等；而 body、blog_text 的关键字都有 pfizer（辉瑞疫苗），因此这个文本数据能供深入文本分析使用。

主题模型（Topic Model）

由此开始文本分析，欲比较整体时间（2014 年至 2020 年）和新冠疫情期间（2020 年至 2021 年），人们对于疫苗的主题讨论有什么有趣的变化。

a. 布朗聚类（Brown Clustering）

图 3.9（a）和图 3.9(b) 为 vaccine 聚为一类的词汇，可看出聚焦到新冠疫情期间（2020 年至 2021 年）的时候，有关疫苗的议题会讨论到 phase, government, pandemic，可以判断是在讨论疫苗研发状况，并且还提及 sputnik（俄罗斯卫星疫苗），能确认新冠疫情期间（2020 年至 2021 年）的主题为疫苗研发。

```
[('medium', 107),
 ('protein', 106),
 ('shoot', 105),
 ('variant', 104),
 ('baby', 103),
 ('paper', 102),
 ('document', 101),
 ('19', 100),
 ('deadly', 99),
 ('certainly', 98)]
```

图 3.9(a)整体时间（2014 年至 2020 年）vaccine 聚类

```
[('variant', 169),
 ('government', 168),
 ('vaccinate', 167),
 ('news', 166),
 ('phase', 165),
 ('warn', 164),
 ('pandemic', 163),
 ('anyone', 162),
 ('sputnik', 161),
 ('ever', 160)]
```

图 3.9(b) 新冠疫情期间（2020 年至 2021 年）vaccine 聚类

b. 隐含狄利克雷分布（Latent Dirichlet Allocation, LDA）

整体时间（2014 年至 2020 年）：

topic 1: 与 covid 相关且也讨论到 pfizer（辉瑞疫苗），故为 covid 主题。

topic 2: 与 autism(自闭症)相关的主题。

topic 3: 与加拿大相关的流感疫苗主题。

topic 4: 因为有 polio（小儿麻痺症），又提到家长和孩童，故可能为小儿麻痺症主题。

由于 topic model 是非监督式学习，因此较无法判断要多增加 topic 数抑或减少 topic 数以得到更清晰的 topic，已经尝试过分割 topic=5 或 topic=3 但也没有较好解释，有趣的是，即便是整体时间（2014 年至 2020 年），也能分出 covid 主题，可见新冠疫情的网络讨论不容小觑。

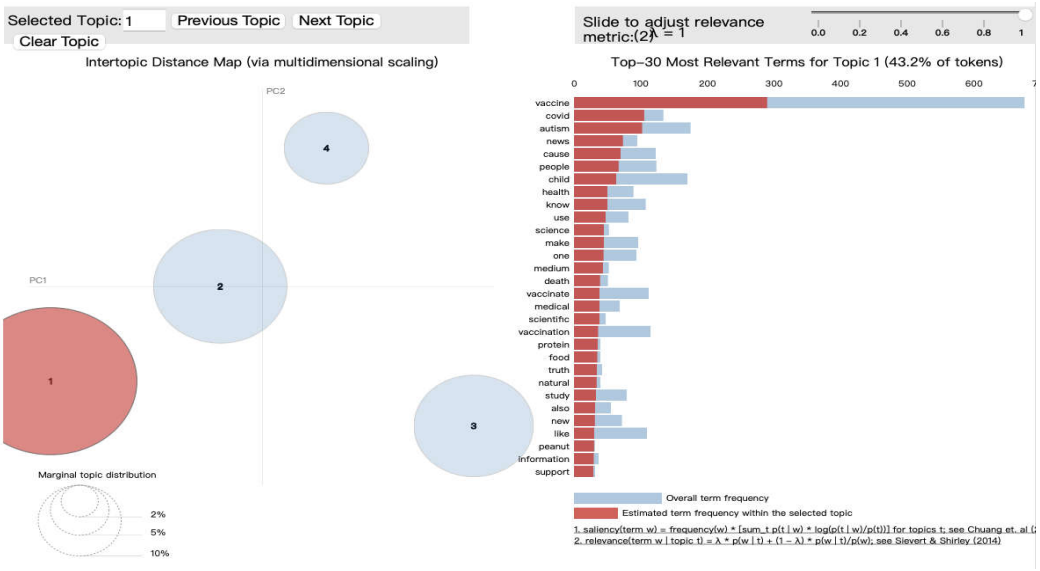

图 3.9（a）整体时间（2014 年至 2020 年）topic 1

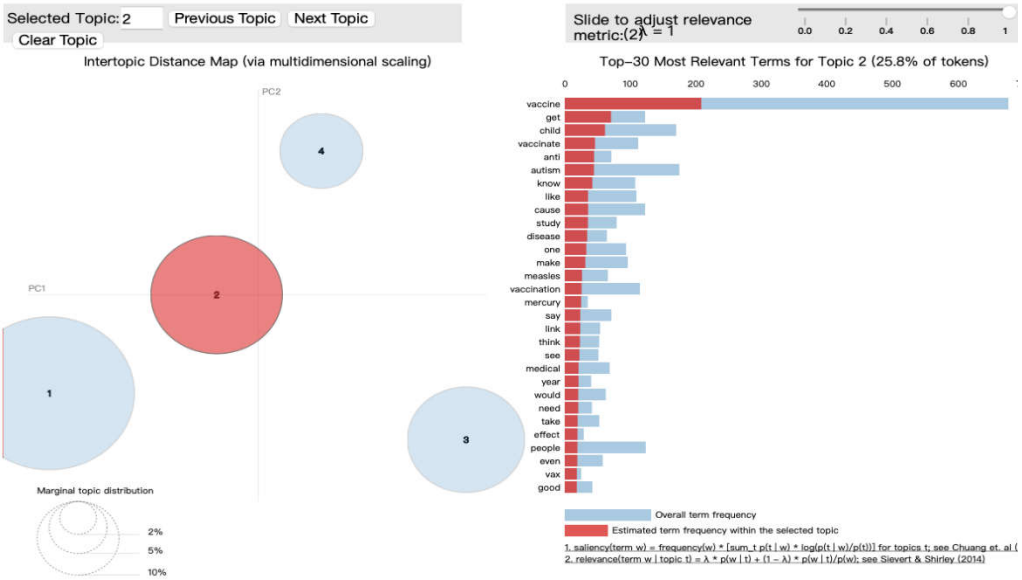

图 3.9（b）整体时间（2014 年至 2020 年）topic 2

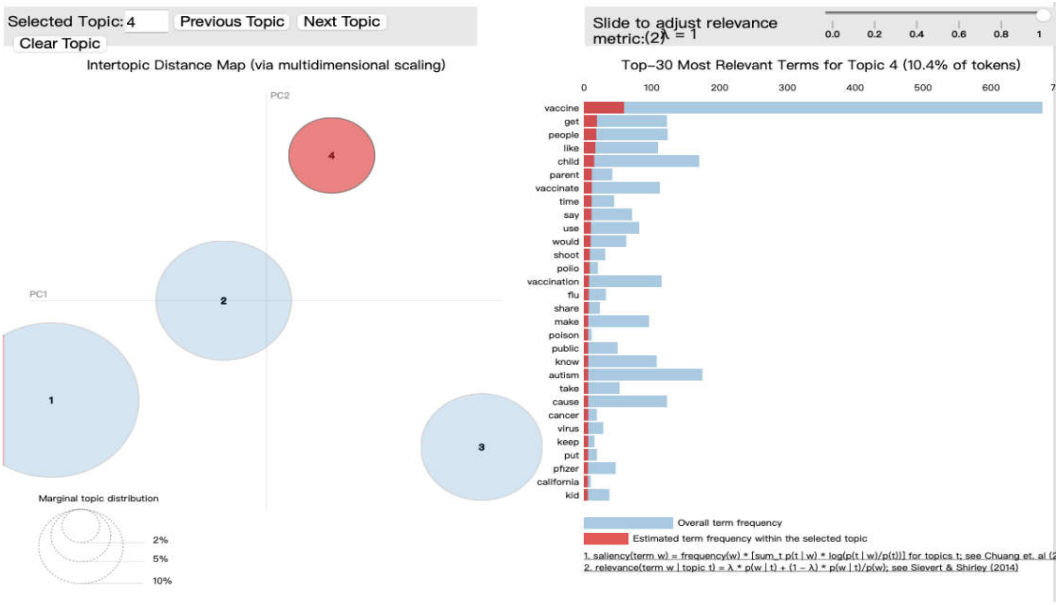

图 3.9（c）整体时间（2014 年至 2020 年）topic 3

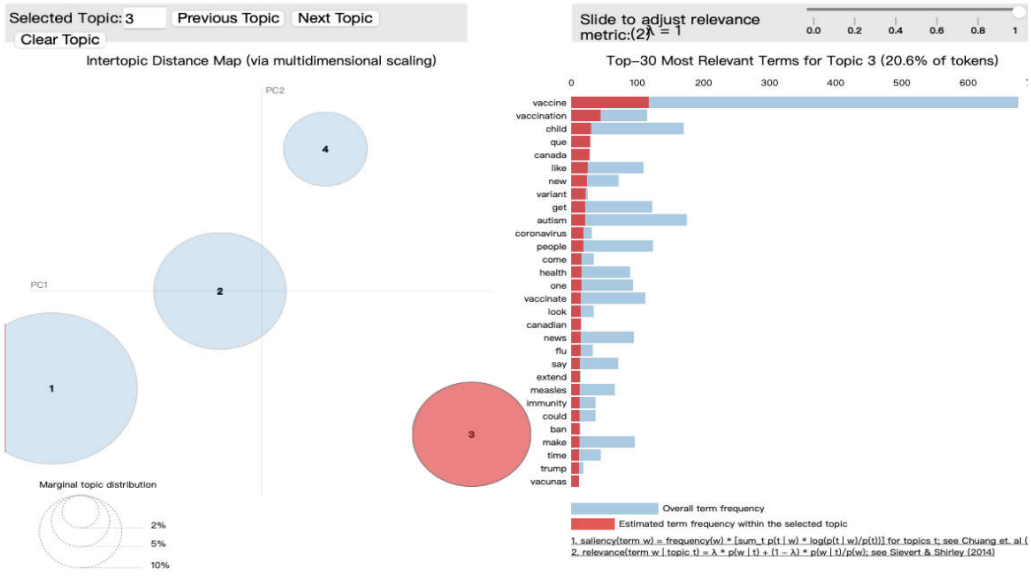

图 3.9（d）整体时间（2014 年至 2020 年）topic 4

- 新冠疫情期间（2020 年至 2021 年）：

topic 1：加拿大相关的 pfizer（辉瑞疫苗）主题

topic 2：讨论到 sputnik（俄罗斯卫星疫苗），pfizer（辉瑞疫苗），故为各种新冠疫苗主题

新冠疫情期间（2020 年至 2021 年）所有疫苗主题都与新冠疫情相关。

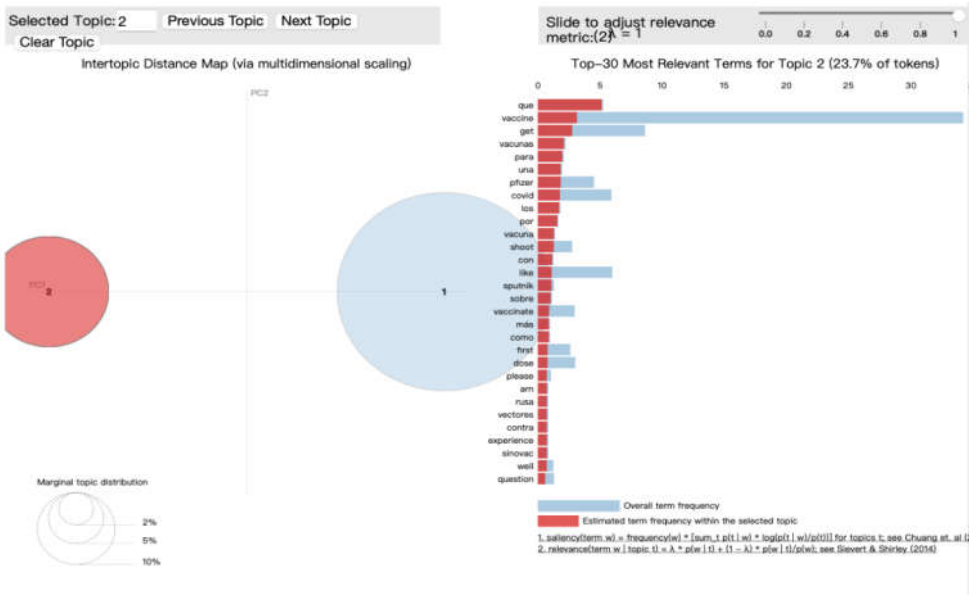

图 3.10（a）新冠疫情期间（2020 年至 2021 年）topic 1

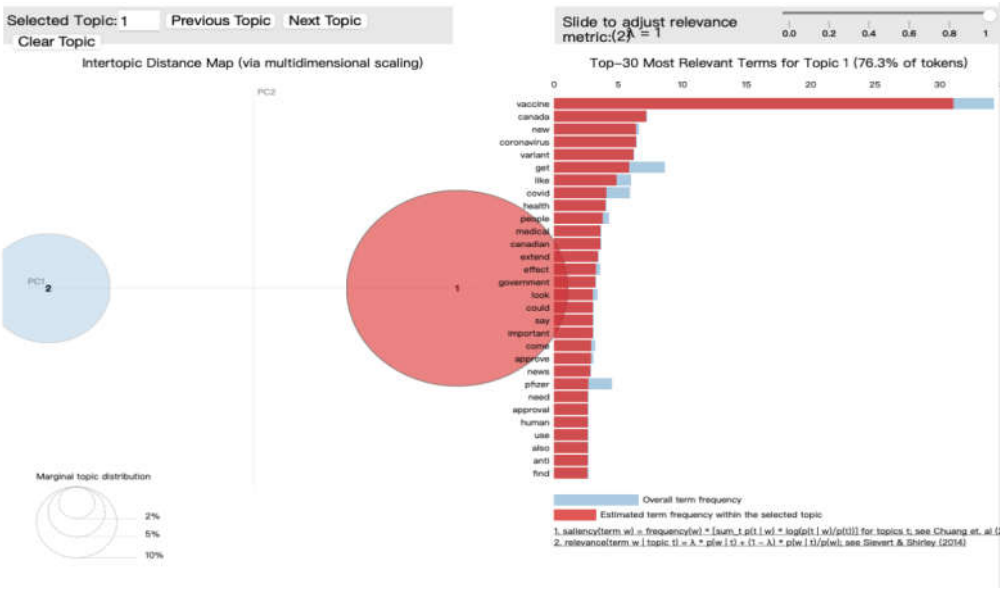

图 3.10(b)新冠疫情期间（2020 年至 2021 年）topic 2

情感分析（Sentiment Analysis）

以词云图和主题模型分析完之后，十分确认新冠疫情期间（2020 年至 2021 年）的文本数据为新冠疫苗文本数据，因此情感分析部分聚焦在新冠疫情期间（2020 年至 2021 年）。

a. 极性（Polarity）

使用 nltk 替 text 做极性评分，设定评分为正值则为正向情感（Positive）；评分

为正值则为负向情感（Negative）；评分为零则为中性情感（Neutral）。由下可以看出正向情感和负向情感在 2021 均相对增加，应是因为疫苗的出现所导致的，而正向情感上升的比例（42%上升至 53%）又比负向情感（30%上升至 33%）多。整体而论，网上对于疫苗的中性观感在新冠疫苗产生后下降（28%下降至 13%），且在新冠疫苗产生后正面观感大于 50%，还是比较有信心的。

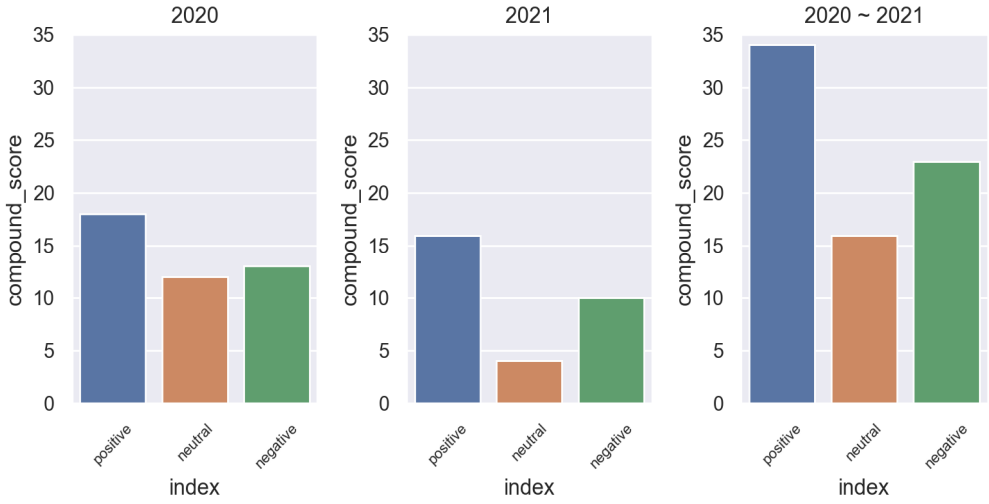

图 3.11 新冠疫情期间（2020 年至 2021 年）情感分析柱状图

在 2021 可以看见正向情感有很大部份提及 pfizer（辉瑞疫苗），另外负向情感部分的部分，2020 很大与 animal 相关，可能是一开始全世界都在讨论是否是哪种动物传染出来的。整体而论，网络文本情感对于 pfizer（辉瑞疫苗）有正向情感。

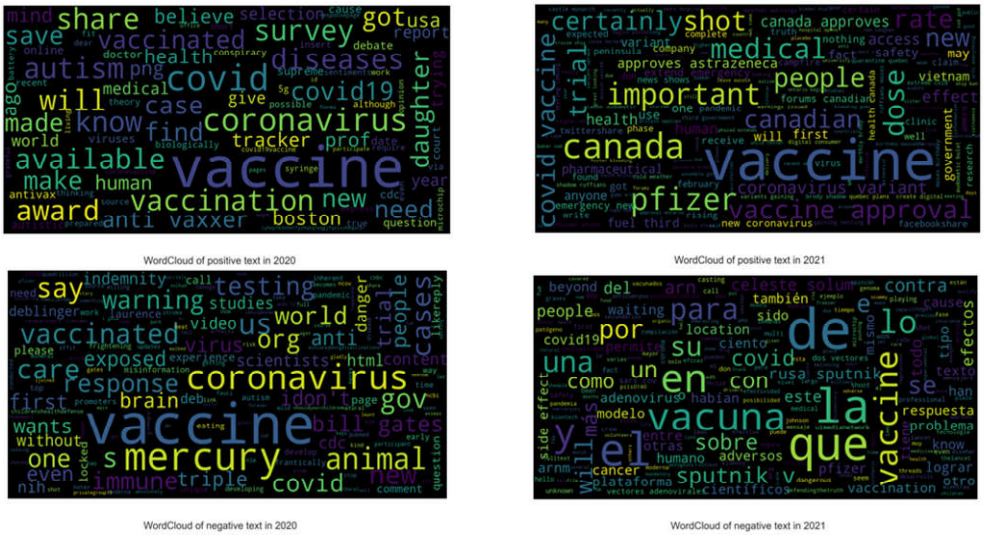

图 3.12 新冠疫情期间（2020 年至 2021 年）情感分析的词云图

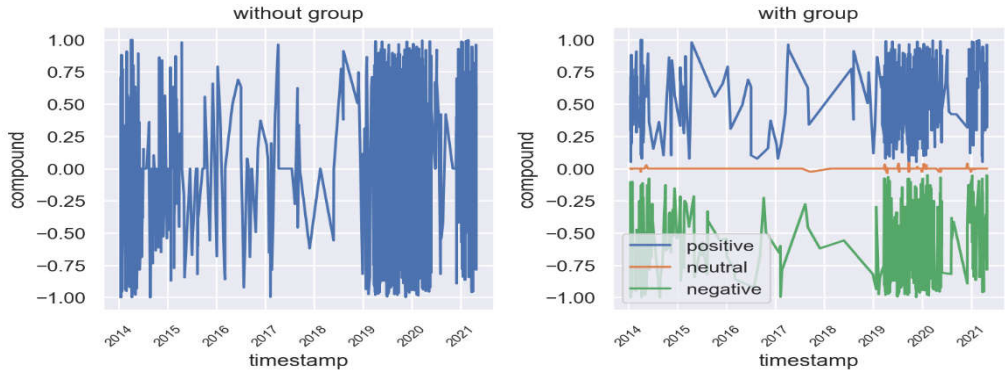

图 3.13 极性评分时间趋势图

最后由 J 可以看出从 2019 年开始，文本的情感波动较为明显且数值的绝对值也较大，可以确认这段时间开始后，网络上对于疫苗相关的文本有着较为明显的情感用词。

b. 主观性（Subjectivity）

主观性数值越接近 0 越客观，越接近 1 越主观，k 明显看出呈现漏斗状，越中性的内容越客观，且大多数的数据集中在主观性为 0.4~0.6 间，可见大部分的内容期没有太大的主观和客观。

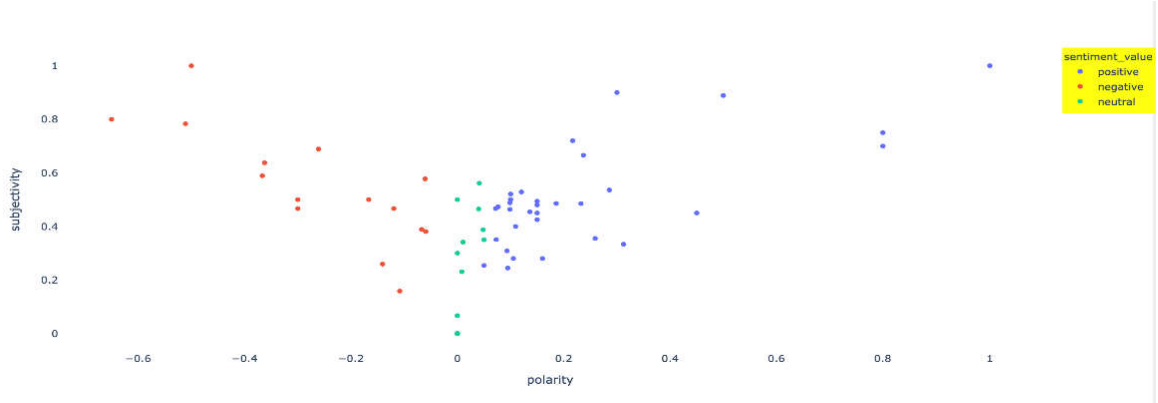

图 3.14 新冠疫情期间（2020 年至 2021 年）极性与主观性分布图

3.3 疫苗接种情况的可视化与数据挖掘

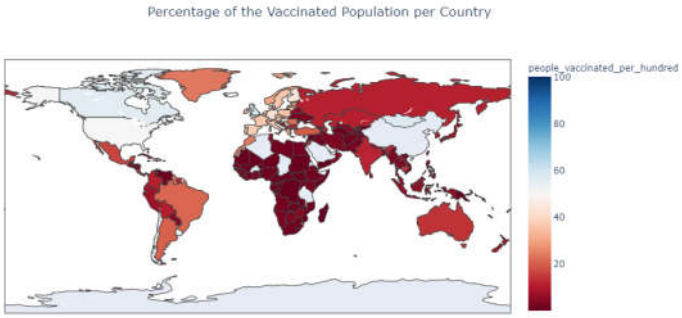

图 3.15

图 3.15 展示了 kaggle 的“COVID-19 World Vaccination Progress”数据集中的各国疫苗接种数目统计（kaggle 数据来源于 WHO 公开数据）。

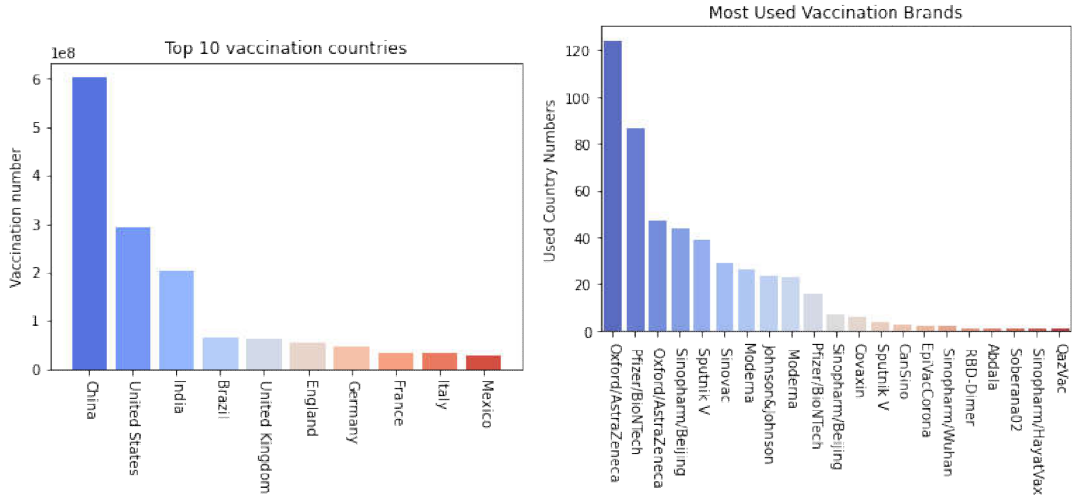

图 3.16

图 3.16 展示了全世界接种人数前十位的国家排行以及应用国家最多的疫苗品牌，其中中国的疫苗接种人数远超其他国家，疫苗品牌中英国的阿斯利康应用国家最为广泛，达到了 124 个国家，**国产疫苗中北京国药的疫苗排名第三**。

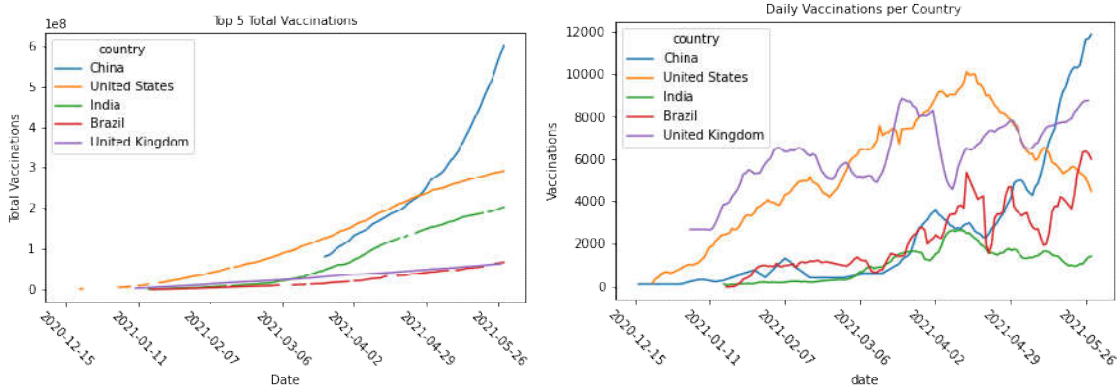

图 3.17

图 3.17 左图为每日统计的总接种人数随时间变化的曲线，由于数据集中有部分日期的数据缺失，因此曲线存在断点。右图展示了每日新增接种人数随日期变化的曲线，**可以看到中国在五月初国内疫情零星散发以来，每日疫苗接种数目显著提升。美国在四月中旬疫苗接种热度达到高潮，随着美国国内的疫情得到控制，美国的每日新增接种人数也随之下降。**

为了进一步探究疫苗接种行为和群众疫苗接种意愿与疫情发展的关系，本文将全球确诊人数增长曲线、全球每日新增疫苗接种曲线和本文文本分析得到的 reddit 论坛网友的每日疫苗评论积极性统计量两两对比，如图 8，9，10 所示。（由于数据集限制，此部分分析仅采用 2020 年 12 月至 2021 年 5 月的数据）（“reddit 论坛网友的每日疫苗评论积极性统计量将每日”论坛评论与内容的积极/消极程度评分后取平均作为改日的积极性标签。

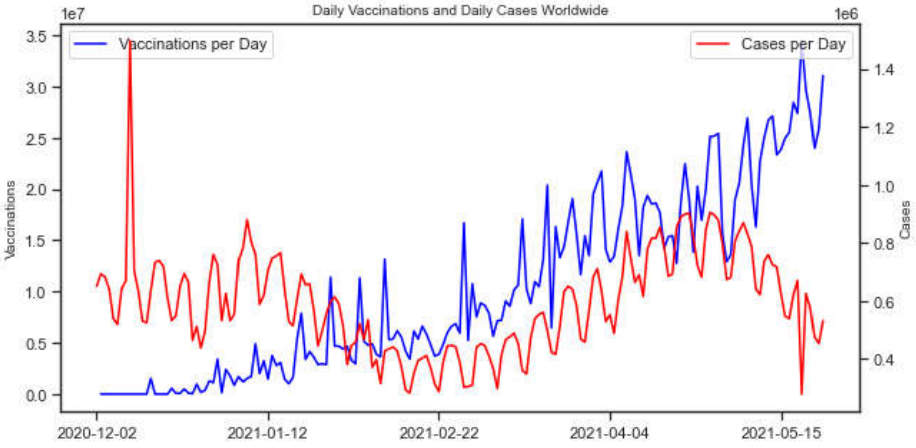

图 3.18

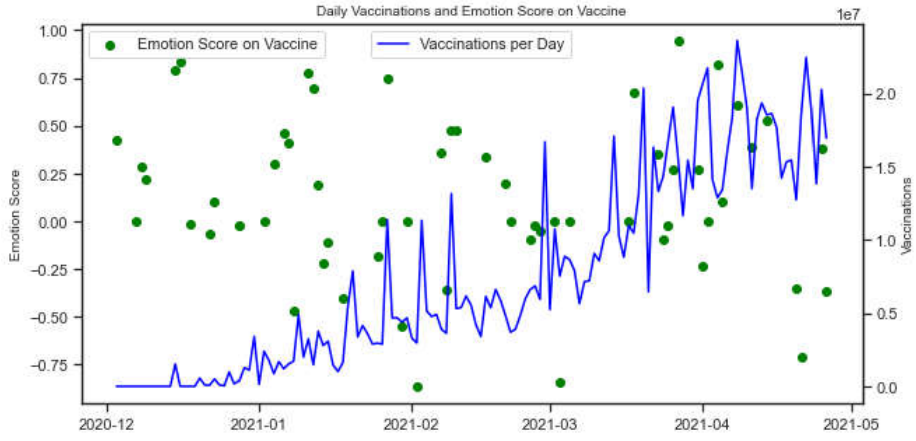

图 3.19

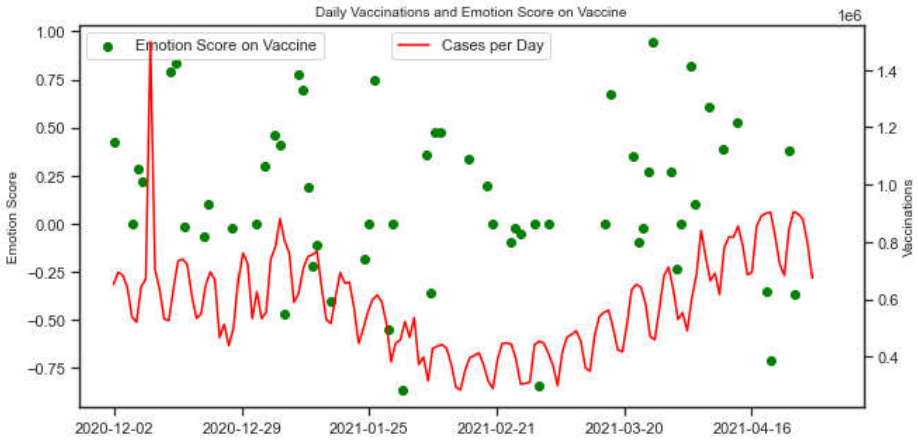

图 3.20

由图 3.18 可知，随着疫苗接种的展开，全球每日确诊人数增长数在 2021 年 2 月末达到了一个局部的最小值。说明疫苗接种对于防控疫情具有较为显著的作用。

图 3.19 并不能直观提供疫苗接种程度与网友评论积极性的关系，原因可能在于网友评论数据较少，而且 reddit 论坛的主要用户为北美用户，全球每日新增接种人数与 reddit 网友评论积极程度的相关性可能较低。

图 3.20 可以大致展现较积极评论主要出现在在疫情较为高发的日期中。（红色曲线的两个峰值附近散点的评分较高）大致估计在疫情较为严重的时间，网友对于疫苗的期待和正面评论较为广泛。

第 4 章总结与建议

新冠肺炎疫情的蔓延对人类社会的生产生活带来了巨大的影响，且在可见的时间内人类社会仍将需要付出巨大的努力与之对抗。而阻断病毒的蔓延，形成群体免疫，最重要的途径便是通过疫苗的大范围接种。基于此，本文希望通过对 kaggle 上关于新冠疫苗接种相关评论的公开数据进行分析以对疫情发展和公众对于新冠疫苗之间的关系进行研究。

首先，本文通过四次多项式拟合，建立了对新冠肺炎感染的预测模型，通过对实际走向相比，可以看到，除一些异常点之外，具有比较好的预测能力；其次，通过对 reddit 上网友对于新冠肺炎相关评论数据的分析，发现随着新冠肺炎在世界范围内的影响越来越大，网友对新冠疫苗的关注度和讨论度越来越高，整个情感态度的也由最初的负面居多到中性偏积极，可见，随着疫情严重程度加深和疫苗的逐渐成熟，人们对疫苗的接受程度越来越高；最后，对有关新冠疫苗在世界范围内的接种和应用范围数据的分析，可以明显看到，我国的疫苗接种量位于世界首位，疫苗技术的成熟也在世界范围内得到了认可。

基于前文的分析，对于之后的新冠肺炎防治来说，公众对疫苗的接种的怀疑情绪已经不是阻碍疫苗接种率的最大的问题，虽然 Reddit 主要是北美用户但从评论分析来看，公众对新冠疫苗的接受度已经越来越高，重要的是面对在不同国家不同品牌疫苗，公众如何选择接种不同疫苗，而更重要的是，在新冠病毒变种不断出现，疫情仍继续蔓延的情况下，各个品牌的疫苗如何证明自己对新冠肺炎防治的有效性，以保持公众的信任度具有重要意义。

因此，建议积极利用 5 月以来国内部分地区疫情零星散发导致民众接种意愿升高的形势加紧疫苗推广接种工作，借此提升疫苗接种普及率。同时，建议针对国产疫苗对于英国、印度等地区变种病毒的预防有效性开展深入研究，为下一阶段疫情防控做好基础数据支撑。最后，建议积极鼓励针对变异病毒的疫苗研发工作，以及吸入式等新型疫苗接种形式的临床试验工作，为更加有效地应对疫情做好准备。

参考文献

[1]RANDOLPH HE, BARREIRO LB. Herd immunity: understanding COVID-19［J］. Immunity,2020,52（5）:737-741.

[2]QUINN SC, JAMISON AM,AN J, et al. Measuring vaccine hesitancy, confidence, trust and flu vaccine uptake: results of a national survey of White and African American adults［J］.Vaccine,2019,37（9）:1168-1173.

[3]GONZÁLEZ-BLOCK M, ARROYO-LAGUNA J, RODRÍGUEZ-ZEA B, et al. The importance of confidence, complacency, and convenience for influenza vaccination among key risk groups in large urban areas of Peru［J］. Hum Vaccine Immunother,2021,17（2）:465-474.

[4]王铁山,张青.新冠肺炎疫情对我国外贸企业的影响及应对措施[J].经济纵横,2020(03):23-29.

[5]刘晓曦,戴俊明,陈浩,等.基于“3Cs”模型的公众新冠肺炎疫苗犹豫影响因素的横断面调查[J/OL].复旦学报(医学版):1-6[2021-05-16].http://kns.cnki.net/kcms/detail/31.1885.r.20210510.2048. 006.html.