MAPPING NEOLOGISM AND COLLECTIVE HUMAN BEHAVIORAL CHANGES: A STUDY OF COVID-19 RELATED EMERGENT NEOLOGISMS USING BIG DATA

SIYU LEI

PhD

The Hong Kong Polytechnic University

This program is jointly offered by The Hong Kong Polytechnic University and Xi'an Jiaotong University

2023

The Hong Kong Polytechnic University

Department of Chinese and Bilingual Studies

Xi'an Jiaotong University

School of Foreign Studies

Mapping Neologism and Collective Human Behavioral Changes: A Study of COVID-19 Related Emergent Neologisms Using Big Data

LEI Siyu

A thesis submitted in partial fulfillment of the

requirements for the degree of Doctor of

Philosophy

June 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____

LEI Siyu

_____

Abstract

Neologisms have been widely recognized as an extremely sensitive linguistic indicator of a new social event. Prior studies have found that the preference for selecting a gain or loss framing strategy to represent a new social event such as a pandemic can respond to relevant policies, pandemic status, and speakers' emotional change during the pandemic. However, existing studies on disease-related neologisms largely ignored the important role of neologisms to respond to human behavioral changes during the epidemic. As an emerging pandemic, COVID-19 is not only a medical event but also a significant social event that generates a large number of online discussions. Under the big data era, a diachronic tracking study on the use of COVID-19 emergent neologisms on the internet can provide a critical lens for us to better understand collective human behavior during different stages of the pandemic. In this study, collective human behavior is operationalized as relevant policy announcement, pandemic development, and public emotional changes under the ongoing disease.

This thesis employs a mixed research design. The qualitative part tracks the developmental patterns of COVID-19 emergent neologisms from the Baidu Index, from the end of 2019 to the end of March 2021, and analyzes that their fluctuations according to important policies. The most important part of the present thesis lies at the prediction and N-gram co-occurrence. The prediction involves two steps: (i) correlating the internet searches of COVID-19 emergent neologisms with the pandemic cases, and then (ii) training/validating/testing their mathematical relationship based on multiple (non)linear and fine-tuned regression models. To highlight a more important role of emergent neologisms in associating with the collective human behaviour, the current thesis compares the predictability of COVID-19 emergent neologisms

with buzzwords motivated by the COVID-19 pandemic (i.e., vector names and Personal Protection Equipment (PPE) names). The other quantitative part explores if emergent neologisms can be good indicators to monitor the change of public attention. This part also involves two steps: (i) hypothesizing the public emotional change at different stages of the pandemic from the general development of COVID-19 emergent neologisms, and then (ii) verifying the hypothesis by N-gram co-occurrence (N = 2 - 6 Chinese words) of the crawled Sina Microblog posts.

The qualitative result showed that the development of COVID-19 emergent neologisms corresponded with the important policies over the fifteen months after the pandemic outbreak. The prediction result showed that, compared with buzzwords, COVID-19 emergent neologisms are better to predict pandemic cases by using binomial main effects. A better fitting curve is a Least Angle Regression model. For the monitoring effect of COVID-19 emergent neologisms on the public emotional change, the public emotion was hypothesized and verified to experience the three stages from fear to relaxation together with caution based on observing the general development of COVID-19 emergent neologisms over the fifteen months.

The contributions of this thesis are twofold. For real life application, emergent neologisms, rather than buzzwords, are better indicators for predicting an emerging public health event and for monitoring the change of public emotion. For theoretical explication, the current thesis proposes an interactive network associating emergent neologisms, relevant policies, pandemic development, and public attention to highlight the important role of emergent neologisms in responding to the collective human behavior and to provide valuable insights to the role of neologisms in language change.

Publication based on the Thesis

The pilot study for the present thesis was published on international prestige journal, *Lingua*, and awarded **Editors' Choice**.

Lei, S., Yang, R., & Huang, C.-R. (2021). Emergent neologism: A study of an emerging meaning with competing forms based on the first six months of COVID-19. *Lingua, 258*, 103095.

The pilot study used the first six-month data of nomenclatures to the COVID-19 pandemic on the Baidu Index to echo the important policy announcement of that period and to model the development of pandemic by simple linear, logistic, and binomial regressions.

The significant contributions of this pilot study are threefold:

Firstly, it proposes a new terminology 'emergent neologism', which refers to the non-replacement change of neologisms compared with most existing studies on replacement change.

Secondly, it highlights the importance of using internet searches as an early warning indicator to model the development of an emerging pandemic.

Thirdly, it explains the preference of selecting different variants to refer to the disease at different stages of the pandemic based on the strategies of categorization, avoidance, and synthesis.

Notes

1. The current thesis uses Leipzig Glossing Rules to present the Chinese languages.
2. The decimals of all the results are kept at three.

Acknowledgements

Throughout the writing of the current dissertation, I am very grateful to receive a great deal of support and assistance from various parties.

I would first like to extend my sincere gratitude to my brilliant, patient, supportive, and inspiring chief supervisor at The Hong Kong Polytechnic University, Chair Professor Chu-Ren Huang. His insightful and constructive comments and suggestions pushed me to sharpen my thinking and brought my work to a higher level. Meanwhile, his patience and encouragement make me convinced to continue on the academic road and aspire to become such a knowledgeable and courteous academic master as him in the future. The positive influences and support from Professor Huang are also reflected in many aspects. I clearly realize that words cannot express my gratitude to Professor Huang.

My deep gratitude also goes to my chief supervisor, Professor Ruiying Yang at the Partner Institution, Xi'an Jiaotong University. In my long research endeavor, she has guided me meticulously in thinking and writing, encouraged and helped me kindly when I had difficulties in research and life. Without her support and encouragement, I might not have had the opportunity to follow Professor Chu-Ren Huang to do research at The Hong Kong Polytechnic University.

Meanwhile, Doctor Sing Bik Cindy Ngai from Chinese and Bilingual Studies, The Hong Kong Polytechnic University has given detailed suggestions for the coherence of different sections in my dissertation at my confirmation of registration. Professor Katheleen Ahrens and Professor Dennis Tay from the Department of English and Communication, The Hong Kong Polytechnic University have also provided valuable comments on my pilot study on the neologism evolution study published on *Lingua*. Doctor Joe Ching from English Language

Center who offered a great deal of guidance to the thesis writing is also a respected teacher I would like to thank. I am also indebted to Professor Xiaowen Zhu and Doctor Yawei Yang from Xi'an Jiaotong University for giving guidance on statistical tests and visualization work, respectively, to my pilot study.

I would also like to thank my group members such as Professor Annie Xiaowen Wang from Guangdong University of Foreign Studies, Doctor Yuyin Hsu and Doctor Mingyu Wan from The Hong Kong Polytechnic University, and Doctor Qi Chen and Mr. Liang Xu from Xi'an Jiaotong University.

My family has consistently supported me in pursuing a bitter-and-sweet academic career. I am greatly indebted to my grandmother, grandfather, mother, and father for their unconditional love and complete support.

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

COVID-19: Corona Virus Disease 2019

L2: Second Language Acquisition

LAR: Least Angle Regression

NLP: Natural Language Processing

PPE: Personal Protective Equipment

SARS: Severe Acute Respiratory Syndrome

SBC: Schwarz Bayesian Criterion

VIF: Variance Inflation Factor

WHO: World Health Organization

CHAPTER ONE INTRODUCTION

1.1 Overview

Neologisms have been widely accepted and considered to be one of the sensitive linguistic indicators of new social events (Jing-Schmidt & Hsieh, 2019; Jiang *et al.*, 2021), which thus has received much attention from scholars interested in studying the relationship between language use and societal development (Huang & Hsieh, 2015; Labov, 1966). Corona Virus Disease 2019 (henceforth COVID-19), an emerging pandemic since the end of 2019, has been spreading worldwide and causing tremendous harm to society and the people. Given that everyone has easy access to the internet in the current big-data era via typing target terms in the search box, or uploading posts on social media, it will be convenient to uncover and capture how important neologisms play a role in responding to the change in human behavior over time through tracking COVID-19-related internet searches.

In the linguistic field, there are various definitions for neologisms. They can only refer to the newly coined form-meaning pairs (e.g., Plag, 2002). They can also refer to both existing linguistic forms with new references and brand-new form-meaning pairs (Gove *et al.*, 1993, Webster's Third New International Dictionary; Liu & Liu, 2014). The present thesis adopted the latter definition for neologisms for two considerations. For one thing, it is a broader definition that has received much attention among linguistic researchers (e.g., Oxford Dictionary, 2003; Asif *et al.*, 2021). For another, the inclusion of both existing linguistic forms with new references and brand-new form-meaning pairs is more fit with the 'non-replacement' language change in the internet era. In the current big data era, the internet can record all linguistic variants of a neologism from its emergence to the fixation to the language system. The selection of a certain

linguistic variant is significant to echo the societal development (e.g., Lei, Yang, & Huang, 2021) and human psychological states (e.g., Kahneman & Tversky, 1979) at different stages of an event. For example, in Lei *et al.*'s (2021) research, the linguistic variant 'stigmatizing names' (e.g., *Wuhan bìngdú* 'Wuhan disease') was found to be widely used by the Chinese people when the pandemic caused thousands of deaths, while the linguistic variant 'official names' was found to be more stably used by the Chinese people when the pandemic was weakened to a great extent.

In addition, Kahneman and Tversky (1979) found that there is a close relationship between the public's preference for a specific linguistic variant of a neologism and the cognitive bias among the public. When describing the outcomes, particularly for life-related social events, the public commonly has two options for framing strategies, i.e., positive (or gain) or negative (or loss) framing strategies. Under the pandemic, the positive framing can largely reflect the psychological states of reducing risk probability by the public, while the negative framing may reflect the intensive self-protection awareness away from the disease. On the basis of the observation, the Framing Effect under the Prospect Theory in psychology was proposed (Kahneman & Tversky, 1979). However, prior studies on evolutional linguistics of neologisms have been most focusing on the 'replacement change', i.e., the old form and the finally lexicalized form, which ignores the exploration of competing and co-existing variants before lexicalization (Holubnycha *et al.*, 2020; Klosa-Kuckelhars & Wolfer, 2020). Hence, the inclusion of both existing linguistic forms with new references and brand-new form-meaning pairs is used to examine a very important theoretical issue, i.e., the relationship between different framing strategies and the public's psychological states during the pandemic.

The existing linguistic forms with new references and brand-new form-meaning pairs are called 'emergent neologisms', which was proposed by Lei *et al.* (2021). Under the guidance of

the notion of 'emergent neologisms', the present thesis is conducted on the 'non-replacement' change of language data in order to make full use of the search history recorded by the internet and elaborate on the important role of emergent neologism to the collective human behavior. Specifically, this thesis is designed based on both qualitative and quantitative methods. Through tracking the proportions of searching the COVID-19 emergent neologisms through the internet over the fifteen months since the outbreak of the pandemic (December 21st, 2019 - March 30th, 2021), we analyzed the developmental patterns of internet searches to refer to the COVID-19 pandemic according to significant social events, pandemic cases, and the change of public attention. To establish a robust mapping relation between emergent neologisms and pandemic cases, we also considered buzzwords (i.e., vector names and Personal Protective Equipment, PPE names) that are being highly discussed by the Chinese netizens during the pandemic to compare the model using buzzwords as predictors with the model based on COVID-19 emergent neologisms.

Overall, the thesis is going to address six research questions, including two qualitative ones and four quantitative ones.

Qualitative Research Questions:

(1) What is the pattern of development of COVID-19 emergent neologisms on the Baidu Index during the first 15 months (December 21st, 2019 - March 30th, 2021) of the COVID-19 pandemic?

(2) Whether the developmental pattern of the COVID-19 emergent neologisms can correspond to the important social events over the 15 months?

Quantitative Research Questions:

(3) Do the frequencies of COVID-19 emergent neologisms statistically correlate with pandemic cases, i.e., newly confirmed, newly suspected, new deaths, and currently suspected?

(4) Can the pandemic growth be modeled by the COVID-19 emergent neologisms based on multiple regression models? Which regression model better interprets the mapping relation? Whether COVID-19 emergent neologism better predicts pandemic development than buzzwords?

(5) What type of change in public attention can be reflected based on the observation of the general development in COVID-19 emergent neologisms?

(6) Whether the deduction to the change of public attention by COVID-19 emergent neologisms can be verified by Sina Microblog posts about *dài kǒuzhào* 'wear masks' at selected time points?

Two main hypotheses are generated according to the six research questions shown above. For the qualitative part, the development of COVID-19 emergent neologisms over the fifteen months after the pandemic outbreak can reflect the significant social events. For the quantitative part, COVID-19 emergent neologisms show statistically significant correlation with the pandemic cases and they may demonstrate a better prediction result to the pandemic development based on a specific regression model than buzzwords. Meanwhile, COVID-19 emergent neologisms correspond to the changes in public attention by the verification of Sina Microblog posts.

1.2 COVID-19 Emergent Neologisms and Buzzwords

As has been reviewed in Section 1.1, we used 'emergent neologisms' to represent the nomenclatures that refer to the COVID-19 pandemic. Following the categories of emergent

neologisms in Lei *et al.*'s (2021) research, this thesis also includes the linguistic variants, i.e., under-specifications (e.g., *yìqíng* 'pandemic'), pre-official names (e.g., *bùmíng yuányīn fèiyán* 'pneumonia with unknown reasons'), stigmatizing names (e.g., *wǔhàn bìngdú* 'Wuhan virus'), official names (e.g., *xīn xíng guān zhuàng bìngdú* novel type corona shape virus 'COVID-19'), and English abbreviations (e.g., COVID-19). Definitions and more examples of the emergent neologisms are listed in Chapter Three.

Additionally, the present thesis also considers two types of buzzwords that have been widely mentioned and discussed in the communication during the COVID-19 pandemic, i.e., vector names and PPE names. The consideration of the buzzwords in the present thesis derives from two aspects. For one thing, based on humans' fundamental perception of a new object or event, especially when a new pandemic emerges, in addition to what the disease is (that is, COVID-19 emergent neologisms), there is no doubt that how it transmits (that is, vector names) and how it can be avoided (that is, PPE names) should be also the most critical issues that the public are concerned with an enormous extent. For another, the inclusion of two types of buzzwords can be also compared with the COVID-19 emergent neologisms in predicting the pandemic development, thereby highlighting the important role of emergent neologisms to emerging public health event.

1.3 Goals and Objectives of the Thesis

The goal of the thesis is to employ a mixed research design to present a comprehensive picture of how emergent neologisms interact with related social events and human behavior during the pandemic.

Specifically, the qualitative analysis aims to associate the important policy announcement with the proportional change of searching the COVID-19 emergent neologisms on the internet. The most important part of the current thesis lies at two quantitative analyses. One of the quantitative analyses aims to examine the mapping relationship between COVID-19 emergent neologisms and pandemic cases by multiple regression models to determine a better model interpreter and a better predictor for the future emergent pandemic prediction at the early stages when the pandemic system has not been well established. The other part of the quantitative analyses aims to deduce the attentional changes experienced by the public at different times of the pandemic. Based on the crawled Sina Microblog posts at selected time points, we are then to verify the acceptability of the deduced attentional change by the development of COVID-19 emergent neologisms.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows.

Chapter Two first of all reviews the Framing Effect under the Prospect Theory, which functions as a theoretical framework. Meanwhile, this chapter also reviews the themes and methods of previous neologism-related studies, the applicability of the traditional S-curve to the development of neologisms, and the necessity of tracking the change of public attention during the pandemic situation. Generally, the literature review chapter explains why COVID-19 emergent neologisms and buzzwords are worth studying by the present thesis and how they will be analyzed based on prior literature.

Chapter Three describes the methodology used in this thesis, including data collection, data analyses in terms of the qualitative part, the quantitative part of regression, and the

quantitative part of N-gram co-occurrence (N = 2-6 Chinese words). This chapter also elaborates on the mathematical principles and functions for multiple regression models and the procedures and algorithms behind the co-occurrence calculation because the two quantitative parts are more important in the present thesis.

Chapter Four reports the qualitative result of how the percentages of the COVID-19 emergent neologisms and important policy announcements are associated over the fifteen months after the pandemic outbreak and the regression result of how the COVID-19 emergent neologisms can predict pandemic cases by comparing with buzzwords.

Chapter Five reports the deduction for the change of public attention by the general development of COVID-19 emergent neologisms and the verified result for N-gram co-occurrence calculation based on Sina Microblog data.

Chapter Six discusses qualitative and quantitative results by comparing the findings of the present thesis with previous relevant research. Besides, this chapter proposes an interactive network combining emergent neologisms, policy announcements, pandemic development, and public attention in order to provide some implications for future work in health linguistics.

Chapter Seven concludes the thesis by recapitulating the main findings and pointing out the limitations and future direction.

CHAPTER TWO LITERATURE REVIEW

This chapter deals with how previous literature motivates the present thesis. Section 2.1 introduces the theoretical framework used by the current thesis, i.e., the Framing Effect under Prospect Theory and links it to our research. Section 2.2 reviews the themes and methods in the previous neologism-related linguistic studies in order to point out the necessity of focusing on the disease neologisms and of adopting the quantitative methods. The S-curve has been shown to be a traditional model to describe the evolutions of neologisms in many contexts, so Section 2.3 reviews it and regression models that were found good in describing the evolutions of neologisms. This section is going to provide available mathematical models that can be used for modeling and predicting the mapping relationship between COVID-19 emergent neologisms and pandemic cases. Section 2.4 reviews the methods by previous literature retrieving information on public attention based on social media texts. This section aims to provide an effective method to verify the acceptability of using emergent neologisms to monitor the change of public attention. Section 2.5 reviews the research necessity of studying buzzwords in the neologism studies. Section 2.6 summarizes this chapter.

2.1 Framing Effect Under Prospect Theory as Theoretical Framework

Prospect Theory, a popular concept in behavioral economics, was initially proposed to describe the psychological tendency when people make decisions for risk investment (Kahneman & Tversky, 1979). The term ‘prospect’ refers to the potential result of a risky investment such as a lottery. Based on Kahneman and Tversky’s (1979) and Tversky and Kahneman’s (1992) controlled studies, individuals tend to feel much more pain when losing $100 than the pleasure

they gain from earning the same amount. Such asymmetric psychology to losses and gains is called 'loss aversion' (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981). Even if the potential outcome of losing $100 and later earning $100 does not make the person have any losses, people's dissatisfaction remains due to the preliminary losses. In the marketing occasions, sellers thus try to present initial gains to buyers for more significant transaction success rates.

Such asymmetric psychology of gains and losses will be more readily understood by Figure 1 visualization, where losses are on the left side of the value axis (Area 3), and gains are on the right side (Area 1).

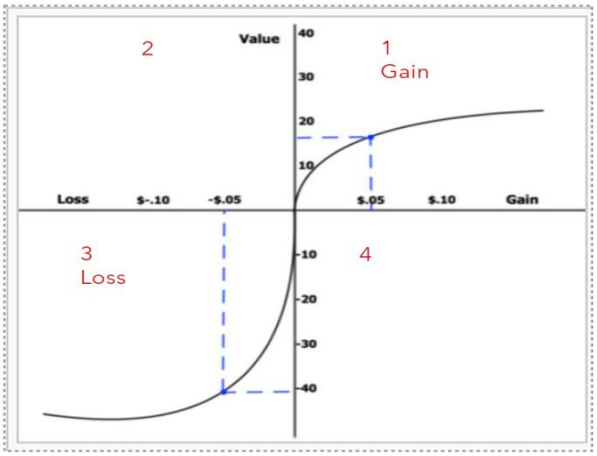

Figure 1 Properties of the value function (https://en.wikipedia.org/wiki/Prospect_theory)

Figure 1 shows the asymmetric psychology by demonstrating a steeper gradient on the losing side than on the gain side.

Based on the close association with people's psychology, the loss aversion has been widely applied in interpreting the psychological status of humans. In the psychological field, the loss aversion is called the Framing Effect (Plous, 1993), which indicates how the different framing strategies influence the change of human's psychology. For instance, Tversky and Kahneman (1981) explored how positive framing (i.e., how many people would live) and negative framing

(i.e., how many people would die) affected participants' responses to a choice in a hypothetical life and death situation. The experiment background is that there are 600 passengers in total on the ship who were affected by a deadly disease. In Treatment A, 400 passengers were assumed to die. For such a scenario, participants were asked to choose either a positive frame (e.g., "saves 200 lives") or a negative frame (e.g., "400 people will die") to describe 400 deaths. Treatment B announced a 33% chance of no death, but there was a 66% chance of all 600 passengers' deaths. Experimenters prepared some positive frames (e.g., "A 33% chance of saving all 600 people, 66% possibility of saving no one.") and some negative frames (e.g., "A 33% chance that no people will die, 66% probability that all 600 will die.") for participants to choose one out of the two expressions. Results indicated that 72% of participants chose the positive framing, but only 22% chose negative framing in Treatment A. And, more than 80% of participants chose positive frames and less than 20% of participants chose negative frames in Treatment B. The results also received successful replication by Druckman (2001b). The close link between framing strategies and the public's final decision is elaborated more in the field of policy support and implementation. According to Druckman's (2001a) survey result, a greater number of people would support an economic policy highlighting the employment expansion than that highlighting unemployment reduction. In addition to more gain, the positive framing also means less loss. Plous (1993) found that almost 62% of participants opposed the opinion of "allowing public condemnation of democracy", but the opposition votes decreased to 46% when the opinion framing changed to "forbid[ing] public condemnation of democracy". Similarly, in Gächter *et al.*'s study (2009), more than 93% of registration would be completed earlier when a penalty fee was announced to be paid for late registration. By contrast, the registration rate would drop to 67% when a discount was announced for early registration.

Recent work confirmed the interpretability of the Framing Effect in public decision-making. Barry, Sherman, and McGinty's (2018) study focused on an event of a drug overdose in the U.S. The status quo of the time is that drug overdose caused thousands of deaths in the U.S., which provokes the abomination of the public. The U.S. government decided to address this opioid epidemic. One of the measures was to establish a drug-related website to deliver the harm of the drug overdose to the Americans. The survey result showed that different framing used to name the website won different support rates: the term "overdose prevention website" obtained more than a 45% support rate, which was attributed to the potential to expand the population who are willing to view the overdose prevention policy as acceptable; the term "safe consumption website" seems to make this life-threatening activity safer and more legal for the addicts, which thus causes a lower support rate (around 29%). The Framing Effect has been also found to play a role in the vaccination rate of the COVID-19 pandemic in China. In Peng, Guo, and Hu's (2021) online survey, the loss framing messages[1] show a stronger persuasion effect than the gain framing messages[2].because what people can gain from the vaccination is more direct and preferable than what they will avoid by the vaccination.

On the basis of the above results, the exploration of different framing strategies can respond to the public's psychological change and decisions. From a macro perspective, different linguistic expressions of the same event, for example, the COVID-19 pandemic, would be a useful indicator to capture the public behavior at different stages of the event. Such an

[1] The example of the loss framing message: Nanshan Zhong, an academician of the Chinese Academy of Engineering and head of a high-level expert group at the National Health Commission, said: "The longer it takes to get vaccinated, the more likely there will be more mutant strains (Peng *et al.*, 2021, p. 4).

[2] The example of the gain framing message: China's biological COVID-19 inactivated vaccine protection rate is nearly 80%. At present, more than 1 million people have been vaccinated, and no serious adverse reactions have occurred to any of them. None of the tens of thousands of people who have gone overseas to high-risk countries and regions have been infected, which fully proves the safety and effectiveness of the vaccine (Peng *et al.*, 2021, p. 4).

assumption has received preliminary confirmation by a recent work (Lei *et al.*, 2021). According to one of their main findings, Chinese netizens tended to use under-specifications (which do not specify the disease, e.g., *bìngdú* 'virus' and *yìqíng* 'pandemic') whenever mentioning the pandemic. They explained the wide use of under-specifications by the avoidance strategies. Facing the fatal disease, the fear among the public would be unprecedented. Under-specifications which are a type of nomenclatures with vague mentions of the disease can to some extent lower the fear. This in turn reflects the public selection of avoidance in communicating the pandemic. When the pandemic was relaxed to a more considerable extent, official names (which synthesize the nature of viruses in the references, e.g., *xīn xíng guān zhuāng fèiyán* novel type corona shape pneumonia 'COVID-19') were frequently used by the Chinese netizens. They explained the frequent use of official names at the relaxed stage by the acceptance strategies. With the precise examination of the virus in the emerging pandemic, the understanding of the virus has become clearer and clearer, thereby making the public's fear to the unknown and fatal disease to a greater extent reduced. The frequent use of official names including the virus and the major symptom responds to the public psychological states and social development of the time.

Based on Lei *et al.*'s (2021) pilot study, this thesis is going to continue tracking the use of different linguistic variants of COVID-19 emergent neologisms under the guidance of the Framing Effect of Prospect Theory. However, different from the pilot work, the present thesis is going to conduct a comprehensive and systematic study on the relationship between COVID-19 emergent neologisms and collective human behavior. Firstly, in addition to COVID-19 emergent neologisms that have been examined by Lei *et al.*'s (2021), we added buzzwords, i.e., vector names and PPE names to compare the COVID-19 emergent neologisms in the prediction of pandemic development. Secondly, to verify the finding based on six-month internet searching

data in Lei *et al.* (2021), we expanded the time period to fifteen months. Thirdly, the target variable does not only stay at the significant social events. The present thesis also considers how COVID-19 emergent neologisms interact with policy announcements, pandemic development, and public attention, which are included into collective human behavior. These three subdimensions of collective human behavior correspond to three research designs, respectively. In examining how COVID-19 emergent neologisms reflect important policy announcements, we analyzed the increases and decreases of COVID-19 emergent neologisms by the announcement of important policies. In examining how COVID-19 emergent neologisms predict the pandemic development, we tried multiple regression models to train, validate, and test their mathematical relationship to find a better algorithm and to examine the predictability of COVID-19 emergent neologism compared to buzzwords to the pandemic development. In examining how COVID-19 emergent neologisms monitor the public's psychology, we used N-gram co-occurrence based on Sina Microblog data at eight important time points.

2.2 Themes and Methods in Neologism-Related Linguistic Studies

A neologism can be coined by a variety of reasons. The emergence of social events could be one of the most critical factors, so most neologism-related studies in the linguistic field focused on the association between the use of neologisms and related social events. For example, Jing-Schmidt and Hiseh (2019) noticed that the growing mentions of the term *gāotiě* 'high-speed rail' around 2006 were influenced by the policy implementation of building high-speed rails in China in late 2005. Based on Google Books Ngram Viewer, Li *et al.* (2020) found that the developmental pattern of the two terms 'gaming' and 'gambling' was closely associated with different social events around 1990: the growth of the term 'gaming' followed the rise of video

games, while the term ‘gambling’ showed a closer association with sports betting. In addition to English neologisms, many scholars also examined the relationship between neologisms in Mandarin Chinese and social events (e.g., Chen, 1999; Chou & Hsieh, 2013; Dong *et al.*, 2020). For example, Chou and Hsieh (2013) noticed the emergence of a new lexical lend led by *wēi* (e.g., *wēi diànyǐng* ‘short film’, *wēi mànhuà* ‘four-page comics’) because of the coming of the *wēi shídài* ‘micro age’ in the 2010s. Dong *et al.* (2020) found the close relationship between weather expressions and weather-related events in Sinitic languages.

Comparatively, evolutionary research on the theme of disease-related neologisms received less attention. Till the completion of thesis writing, there have been three publications tracking the use of different framing strategies of naming a disease over time and exploring the relation between framing strategies and related social events and public psychology. For example, during the 2009 H1N1 period, the term ‘H1N1’ was used more frequently than the term ‘swine flu’ in the tweet posts (Chew & Eysenbach, 2010) and was finally selected as the official nomenclature by WHO because this term can make the disease opaquer to the purpose of reducing public fear. The selection of weakening the severity of the H1N1 corresponds to the gain framing strategy. By contrast, supposing that the condition is to raise the public’s awareness of self-prevention to the disease, the loss framing may be preferred. By analyzing how two frames to refer to Ebola, i.e., ‘epidemic’ and ‘outbreak’ were distributed in the New York Times, Daily Mail, and Ynet during the Ebola period, Gesser-Edelsburg *et al.* (2016) found that these two frames were interchangeably used on the three newspapers at the emergence of the disease. Nevertheless, the term ‘epidemic’ showed a sudden increase after the former President Barack Obama employed it in his speech. Based on a great request that the public should pay much attention to the

infectiveness and fatality of Ebola, Barack Obama selected the loss framing strategy (i.e., 'epidemic') to reflect the seriousness and a long duration of Ebola.

Gain- and loss-framing strategies are not always trade-off selection during one disease. Instead, they may be interchangeably used during different periods of the disease. The ongoing COVID-19 pandemic is a typical example where gain and loss framing strategies were used interchangeably. In one of the Lei *et al.*'s (2021) findings, under-specifications such as *bìngdú* 'virus' and *fèiyán* 'pneumonia' have been the most dominant Chinese expressions to refer to this disease since the pandemic outbreak. Their research attributed the wide use of under-specifications by the avoidance strategies in which the public's fear can to some extent be reduced. However, during the most severe period of the COVID-19 pandemic, some stigmatizing names (e.g., *wǔhàn bìngdú* 'Wuhan virus') were frequently used and they even overtook the proportion of under-specifications on the Baidu Index. The overwhelming use of stigmatizing names at the most severe period corresponds to the use of loss-framed strategies in which the public's attention to self-protection can be raised. Though not including all salient symptoms of this pandemic, pre-official names (e.g., *bùmíng yuányīn fèiyán* 'pneumonia with unknown reasons') were also frequently used by Chinese netizens to raise the self-protection awareness. This can be also considered to be the use of loss-framed strategies. When the disease received much relaxation, official names (e.g., *xīn guān fèiyàn* 'novel corona pneumonia') coined by WHO jumped to the second widely used term by Chinese netizens to refer to the pandemic. This is the return to the gain framing strategy, which reflects that the public was not at the denial stage of the pandemic. Instead, they were dared to confront it.

Following the research paradigm in the the tracking studies of disease-related neologisms, the first two research questions are:

RQ 1. What is the pattern of development of COVID-19 emergent neologisms on the Baidu Index during the first 15 months (December 21st, 2019 - March 30th, 2021) of the COVID-19 pandemic?

RQ 2. Whether the developmental pattern of the COVID-19 emergent neologisms can correspond to the important social events over the 15 months?

In addition, we noticed that there is still a big space to deepen the exploration on the relationship between the emergent neologisms and collective human behavior. One possible reason that restricts the depth of the research is the method of qualitative analysis, i.e., tracking the use of neologisms over time and attributing the use of neologisms to possible social events. Hence, the following sections will review and find the quantitative methods that could be applied to interpret the mapping relation between emergent neologisms and collective human behavior.

2.3 Mathematical Expressions between Neologisms and Social Events

Language change often does not proceed linearly but nonlinearly (Bailey, 1973; Kroch, 1989; Labov, 1994, 2001). The most typical algorithm of nonlinearity is the sigmoidal change like a S-curve. The S-curve is the most classical mathematical formulation in describing the language change (Kroch, 1989; Chambers & Trudgill, 1992; Denison, 2002). It includes three stages. The first stage is the innovation stage where a new word or phrase emerged to name a new social event, but few people used it. As an increasing number of people accepted this neologism and would like to use it to describe the social event, exponential growth occurred at the phase of selection and propagation. When the whole community accepted the neologism, the S-curve enters the fixation stage which means that the neologism becomes one part of the language system.

The applicability of the S-curve has been confirmed in many sound change studies. For example, Dras and Harrison (2003) and Harrison, Dras, and Kapicioglu (2002) both reported simulations of the rise of backness harmony in the Turkish lexicon. They created a model of how Turkish speakers interact with each other during the daily life. Their simulation began with a 1,000-word glossary, with 50% of the words harmonic concerning backness. When encountering a word, each speaker will choose to either harmonize or disharmonize the word with some probability, i.e., changing the vowels of the lexical entry of the word. In the simulations, the researchers remained the probability of disharmonizing a word at zero and modeled cases where words were harmonized with non-zero probabilities. They found that the harmonization of a word has a sigmoidal relationship with the number of harmonic entries in one's lexicon. Wang, Ke, and Minett (2004) analyzed the gradual diffusion of a sound change through the lexicon. In this model, they did not restrict themselves to a small number of independent parameters; instead, they considered the interaction of many words in the lexicon. Whereas a phonetic symbol adopts a changed form because of a social bias over a simulation increment, the system propagates through the population in a sigmoidal fashion. Considering the factor of speakers' background, the sigmoidal expression still has interpretability. Baker (2008) found that a sigmoidal progression can predict the interrelationship between a speaker's probability of adopting the sound change and their popularity in the speech community. In this study, he noted that the time when a new sound variant witnesses a decrease in the intonation is critical to understanding language change because it means that a social event that may influence language change occurred. This finding reminds the present thesis to notice important time points when tracking the development of emergent neologisms.

The S-curve can be also applied in interpreting the network structure in simulations of language change (e.g., Ke, Gong, & Wang, 2008; Fagyal *et al.*, 2010; Eisenstein *et al.*, 2014). Ke *et al.* (2008) found that a language change adopted by aligned conversations between two individuals propagates through the community in a sigmoidal fashion. Fagyal *et al.* (2010) successfully replicated the three stages of the S-curve in the generic path of language change by agent-based computer simulations. The result also showed the applicability of the S-curve. Based on a latent vector autoregressive model of Twitter data, Eisenstein *et al.* (2014) modeled the mathematical relationship among linguistic uses on social media, geographical location, and population numbers in the U. S. from 2009 to 2012 and confirmed the applicability of the sigmoidal function to describe the close association between neologisms and social events.

According to the above review of how the S-curve is applied to the evolution of neologisms, we found that the neologisms included in the prior studies are replacement change (Blythe & Croft, 2012). So, it is unclear how different linguistic variants of a neologism compete and develop under the pandemic situations. A vivid example that can show the inapplicability of the S-curve to word change is Li *et al.* (2020)'s study. It tracked the development of two near-synonyms, 'gambling' and 'gaming', over the past 300 years. When removing the factor of related social events, the S-curve can interpret how the term 'gambling' competed with and replaced the term 'gaming'; however, when considering social and economic reasons, the S-curve cannot be applied to the development of two terms. As discussed above, COVID-19 emergent neologisms belong to the non-replacement change of language, so it is reasonable to assume that the S-curve model might not be suitable to explain the development of different COVID-19 emergent neologisms when considering the factor of the COVID-19 pandemic.

Lei *et al.*'s (2021) pilot study partly proved our assumption. They found that the COVID-19 emergent neologisms experienced an exponential increase at the early stage and a gentle decrease at the later stage. Based on this observation, they proposed a binomial mathematical expression to model the mapping relation between the development of the COVID-19 emergent neologisms and pandemic cases during the first six months after the outbreak of the pandemic. To verify the applicability of the S-curve and bionomial regression model, the present thesis considered the buzzwords in the model prediction to compare with the COVID-19-related emergent neologisms, extended the time ranges, and tried multiple regression models. The quantitative part of regression modeling in this thesis aims to respond to an essential theoretical issue: whether the S-curve can be applied in explaining the development of COVID-19 emergent neologisms over fifteen months after the pandemic outbreak. Accordingly, the sigmoid function of log-linear regression is regarded as the baseline. Binomial regressions are also examined in the regression analysis of the present thesis for verification.

So, we asked RQ 3 and RQ 4 as listed below.

RQ 3. Do the frequencies of COVID-19 emergent neologisms statistically correlate with pandemic cases, i.e., newly confirmed, newly suspected, new deaths, and currently suspected?

RQ 4. Can the pandemic growth be modeled by the COVID-19 emergent neologisms based on multiple regression models? Which regression model better interprets the mapping relation? Whether COVID-19 emergent neologism better predicts pandemic development than buzzwords?

Since the prediction part is the most important of the present thesis, there are multiple regression models in addition to log-linear regression and binomial regression models. More details can be found in Chapter Three.

Before moving on to the next part, we want to highlight our contribution compared with two recent works that considered various factors as the early warning indicators for the pandemic cases. Kogan *et al.* (2021) proposed an early warning system established by six sources, namely, Google Trends data, Twitter data, COVID-19-related doctors searches from UpToDate, predictions by the global epidemic and mobility model, human mobility data, thermometer measurements in the United States from March to September 2020. Similarly, Liu *et al.* (2020) created a new machine learning model to forecast the COVID-19 activity in Mainland China ahead of 2 days based on normalized data from Baidu keyword searches, the number of news articles, the cumulative cases of the COVID-19 pandemic, and daily forecasts of COVID-19 activity from a mechanistic metapopulation model in the previous two days of all 32 provinces in mainland China from February 3 to 21, 2020.

These two studies focused on the prediction accuracy of the COVID-19 pandemic cases. In contrast, our thesis regarded COVID-19 emergent neologisms as a useful indicator for reflecting, modeling, and monitoring collective human behavior, i.e., important policy announcement, pandemic development, and the change of public psychology under different pandemic stages. Hence, no merely staying at the prediction as to the focuses by Kogan *et al.* (2021) and Liu *et al.* (2020), our thesis has wider-ranging social contributions via depicting more comprehensive pictures of emergent neologisms mapped with collective human behavior in the context of an epidemic.

Note that the quantitative part of using the internet searches to model the pandemic development and to monitor the public's psychological change do not mean that the development of emergent neologisms determines the development of the pandemic and the change of public attention. Instead, the exploration on gain and loss framing strategies of emergent neologisms at

the different epidemic stages is an important measurement to reflect epidemic development and public psychology, as the Framing Effect of Prospect Theory indicates. From this perspective, the two quantitative parts of this thesis can point out an enormous potential of using internet searches for early prediction of the pandemic and public psychological change, especially when the pandemic system and after-pandemic psychological services are not well established at the early stage of the pandemic. Our thesis could reflect Barry *et al.*'s (2018) opinion that "language matters in combating the ... epidemic" (p. 1157).

2.4 Buzzwords on Collective Human Behavior

Undoubtedly, disease neologisms are comparatively most investigated in the neologism-related studies under the disease situation. However, with the emergence of a new pandemic, the buzzwords about the transmission path of the pandemic and the protection measurement also received much attention by the public. The transmission path expressed by the vector names can provide information on what living organisms transmit infectious pathogens between humans or from animals to humans. There have been a variety of studies that are investigated on what animals or the humans with what symptoms will have bigger probability to transmit the virus (e.g., Arnold, 2020; Corbett *et al.*, 2020; Dinnon *et al.*, 2020; Lam *et al.*, 2020; Wu *et al.*, 2020; Zhou *et al.*, 2020; Shou *et al.*, 2021). For example, Corbett *et al.* (2020) and Shou *et al.* (2021) found that the causative agent of the COVID-19 is the small animal. Furthermore, Dinnon *et al.* (2020) developed a recombinant virus, mouse-adapted SARS-CoV-2 MA, which can interact well with mouse ACE2 and cause more severe disease in aged mice. In addition to mice, snakes and bats were possible vectors that led to the COVID-19 disease in the early biomedical publications. For example, Latinne et al. (2020) found that bats are the inducement for the

COVID-19 pandemic. Zhou *et al.* (2020) thought the causal agent of SARS-CoV-2 is closely related to a bat coronavirus (RaTG13), while its receptor-binding domain is more like that of pangolin coronaviruses (Lam *et al.*, 2020). When the pandemic developed, human-to-human transmission phenomena has also confirmed (e.g., Wu *et al.*, 2020), making people scarier. Since people have easy access to the website where these scientific findings have been recorded and published, the public will search the related information to avoid the vectors' influence on their health.

The protection measurement expressed by PPE names can provide information on what equipment is better used to protect the person against hazards. People cared much about the information on PPE on the website. As Artenstein (2020) stated, the supply chain for secure gowns, gloves, face masks, goggles, face shields, and N95 respirators has been in shortage worldwide for a long time after the outbreak of coronavirus. Therefore, people searched PPE on the overseas website to provide sufficient supplies. Therefore, the extended inclusion of buzzwords, i.e., vector names and PPE names, can be used to be a comparison to indicate the interpretability of COVID-19 emergent neologisms in reflecting the policy announcement and predicting the pandemic cases. More details for what type of vector names and PPE names are selected in the present thesis are included in Chapter Three.

In addition to the pandemic prediction by COVID-19 emergent neologisms, as the Framing Effect of Prospect Theory argues, the selection of different framing in the COVID-19 emergent neologisms can reflect the public's psychological change. The other quantitative part of the thesis is thus to verify whether COVID-19 emergent neologisms can monitor the change of public attention and whether emergent neologisms can be better indicator than buzzwords in

monitoring the change of public attention. The following part will review the social media data and the methods used in the verification.

2.5 Emergent Neologisms and Public Attention

As the Framing Effect of Prospect Theory guiding the present thesis has pointed out the reflective role of gain and loss framing strategies in public psychology, we hypothesized that the public's psychological change can be monitored by the development of COVID-19 emergent neologisms. However, this hypothesis needs verification, so we conducted N-gram co-occurrence analysis based on Sina Microblog data to examine whether the general development of COVID-19 emergent neologisms will monitor the change of public psychology.

Social media has developed rapidly in recent years (Zhao *et al.*, 2020) and has become an important channel for promoting risk communication (Househ, 2016; Gui *et al.*, 2017). Hence, it has been applied in many studies related to public attention during infectious diseases such as H7N9 (Gu *et al.*, 2014; Chen *et al.*, 2019), Ebola (Seltzer *et al.*, 2015; Househ, 2016; Fung *et al.*, 2016), Zika virus (e.g., Seltzer *et al.*, 2017), and Dengue fever (e.g., Wang *et al.*, 2020). Sina Microblog is one of China's leading social media platforms (Zhao *et al.*, 2020), so it could provide much information about public sentiment during the pandemic.

However, previous studies on public attention have frequently used online and offline surveys (e.g., Ahmad & Murad, 2020; Huynh, 2020; Shen *et al.*, 2021; Peng *et al.*, 2022). Ahmad and Murad (2020) examined how people were affected during the COVID-19 pandemic by designing an online Likert-scale survey on Facebook. Huynh (2020) conducted a Likert-scale survey to understand COVID-19 risk perception from socio-economic and media attention perspectives. Shen *et al.* (2021) used support-or-not-support questions to conduct a survey study

to collect public health professionals' attitudes to public health education in China after the COVID-19 pandemic. Also, by a Likert-scale survey, Peng *et al.* (2022) investigated the public attitude to mask-wearing during the COVID-19 pandemic by using different framing strategies. With the advances in AI technologies, emerging Natural Language Processing (NLP) methods have also been used to conduct binary (positive and negative) or triple (positive, negative, and neutral) polarity sentiment and emotion analysis (Sakti, Mohamad, & Azlan, 2021; Tan *et al.*, 2021). Sakti *et al.* (2021) conducted sentiment, and emotion analysis with a supervised machine learning approach by COVID-19-related post data in Indonesia collected from Twitter, Facebook, Instagram, and YouTube from March 31st to May 31st, 2020. The result shows that the general sentiment among the public is positive, and the emotion of trust takes the dominant position compared with other emotions. A recent study by Tan *et al.* (2021) collected the posts from Sina Microblog for the whole of 2020 and analyzed the change of general public sentiment to the pandemic by Tencent NLP. They found that the pandemic had a short- and long-term negative impact on public emotion.

However, the survey method and sentiment and emotion analysis given by NLP stay at the coarse level due to the limited emotion categories, which seems to simplify complex human emotions during different stages of the emerging pandemic. Besides, the automatic extraction sometimes may give an opposite or one-sided result, leading to a necessity to elaborating on the sentiment analysis by other methods. The present thesis adopted the calculation of the N-gram co-occurrence (N = 2-6 Chinese words) to obtain richer psychological information from the public. Unlike prior relevant studies that track the general change of sentiment and emotions to the pandemic itself, the present thesis conducts the emotion analysis through tracking their attitude to *dài kǒuzhào* 'wear masks', an inevitable behavior for self-protection. The exploration

of public emotion to this important self-protection behavior can not only verify the monitoring effect of COVID-19 emergent neologisms on the public attention, but also provide an available approach to the government and stakeholders for monitoring the change of public attitude to mask-wearing during the pandemic.

In addition, exploring how the change of public attention during the COVID-19 pandemic based on Sina Microblog data can also be used to verify the applicability of the Issue-Attention Cycle theory. Downs (1996) proposed the theory of Issue-Attention Cycle, which represents an attention trend line an environmental issue could receive from the public or media. Generally, there are five stages: at the early stage, the issue just occurred but did not receive the public's attention, though some experts or interest groups had begun to find solutions; as the issue became more serious, it won much attention among the public. The social ability to provide more suggestions to address the issue also has been moved to the agenda. The third stage finds the necessary cost for sacrificing a large population, though every problem behind the issue has been addressed. The fourth stage witnessed the decreased public interest step by step. The reduced public interests are reflected in discouragement, fear, or bore. Regarding the post-problem stage, the public attention on this issue was finally replaced by the other social issue. The theory receives much support in political areas (e.g., Petersen, 2009; Steffen & Cheng, 2021) and in some epidemic outbreaks (e.g., Arendt & Scherr, 2019).

For another, posts from social media can, to a large extent, reflect the public attention and psychological change at different times of pandemic, thereby helping the government and stakeholders better to monitor the change of public attitudes to mask-wearing during the pandemic, to take actions accordingly, and to learn lessons for possible pandemic prevention in the future. Similar to previous infectious disease outbreaks such as severe acute respiratory

syndrome (SARS) in 2003 and Ebola virus disease in 2014, COVID-19 has not only threatened the physical health of the public but also imposed a wide range of negative emotions, including fear, depression, and panic disorder (Hawryluck *et al.*, 2004; Mak *et al.*, 2009; Thompson *et al.*, 2017). Unlike previous diseases, the main ways of knowing the COVID-19-related information are from the internet (Tran *et al.*, 2021). However, the age of "infodemic" (Zarocostas, 2020) tends to aggravate negative emotions. Such negative emotions could harm public mental health and even trigger social unrest (Brooks *et al.*, 2020). Therefore, understanding how public psychology has been affected by the pandemic can provide valuable information for policy-makers, government administrators, and mental health service providers (Tan *et al.*, 2021).

Recent work by Lei *et al.* (2021) has demonstrated that the disease nomenclatures witnessed competition among stigmatizing names (e.g., *wǔhàn bìngdú* 'Wuhan virus'), under-specifications (e.g., *bìngdú* 'virus') and pre-official names (e.g., *xīn xíng bìngdú* 'novel type virus') at the early stage (around the whole of January and the middle of February in 2020) of coronavirus over the first six months after the outbreak of the COVID-19 pandemic. Nevertheless, from late February to the end of June, especially after WHO suggested not to use region/country-related words to name the disease, the nomenclatures searched by Baidu netizens became stable after the coinage of official names (e.g., *xīn xíng guān zhuàng bìngdú* 'COVID-19'). According to the theoretical framework of the Framing Effect under Prospect Theory adopted by the present thesis, the search activities might be influenced by social-emotional factors. In addition to the governmental-level policy implementation, the individual's emotional change might probably affect the change in search activities. We thus have reason to believe that the change from the fierce competition among a variety of disease nomenclatures at early stages to the regular use of disease

nomenclatures at later stages can correspond to the change of public emotion from fear to relaxation, respectively.

Previous studies on public emotions, such as Lei *et al.*'s (2021) study that mentioned the strategy of avoidance, have not yet discussed the possibility of using COVID-19 emergent neologisms to predict the psychological change among the public, let alone the further verification of the assumption. Hence, our thesis is going to respond to the gap by extracting the phrase *dài kǒuzhào* 'wearing face masks' from Sina Microblog in order to verify the monitoring effect predicted by the development of COVID-19 emergent neologisms. On this basis, we asked the last two research questions:

RQ 5. What type of change in public attention can be reflected based on the observation of the general development in COVID-19 emergent neologisms?

RQ 6. Whether the deduction to the change of public attention by COVID-19 emergent neologisms can be verified by Sina Microblog posts about *dài kǒuzhào* 'wear masks' at selected time points?

2.6 Chapter Summary

This chapter introduced the theoretical framework of the thesis, i.e., Framing Effect under Prospect Theory. The theory points out the importance of exploring the relationship of COVID-19 emergent neologisms with collective human behavior. To give a more comprehensive picture of how humans behave during the pandemic situation, the thesis reviewed the prior studies on exploring the social events and the change of neologisms searched on the internet, the pandemic development and neologisms, and the public's psychology and neologisms based on the social media posts. To validate the predictability of COVID-19 emergent neologisms on the pandemic

development, the present thesis also considers buzzwords, i.e., vector names and PPE names. The next chapter presents more details of the methodology employed by the present thesis.

CHAPTER THREE METHODOLOGY

Section 3.1 explains why the current thesis focuses on the study setting in China. Then this chapter moves on to Section 3.2 to introduce four main data sources, i.e., Baidu Website and Baidu Index, Baidu News, the Official COVID-19 Pandemic Website, and Sina Microblog. Since this thesis uses a mixed research design, i.e., qualitative and quantitative way, to explore how COVID-19 emergent neologisms interact with collective human behavior under the pandemic, data analysis parts were reported accordingly. Section 3.3 reports how data was analyzed qualitatively. In this section, we listed the categories of COVID-19 emergent neologisms and buzzwords (i.e., vector names, and PPE names) focused on by this thesis and the time for important policy implementation by WHO and the Chinese government over the fifteen-month period after the COVID-19 pandemic outbreak. Section 3.4 reports data analysis in the prediction. This section introduces the mathematical mechanisms of the sigmoidal function of log-linear regression (baseline), simple linear and multiple linear regression models, and fine-tuned regression models with binomial and trinomial expressions. Section 3.5 reports how to use Web Scraper to crawl the Sina Microblog data with the search phrase *dài kǒuzhào* 'wear masks' and how to analyze the crawled blog data through *SnowNLP* package in Python and how to visualize the association strength of N-grams by *VOSviewer*. Section 3.6 summarizes this chapter.

3.1 Study Setting in China

The Chinese government is a well-deserved example in the COVID-19 pandemic response over the world. To respond to the pandemic rapidly, the Chinese government has implemented prevention and control measures, including establishing a Central Leadership Group for

Epidemic Response and the Joint Prevention and Control Mechanism (Report of the WHO-China Joint Mission on COVID-19, 2020), National Infectious Disease Information System (IDIS)[3], as well as providing pandemic-against suggestions to peoples. Given that China is the most experienced country in responding to the ongoing disease, sufficient and accurate data would be available for the present thesis. More importantly, our exploration of the known Chinese pandemic data and the Chinese public's emotional change may guide China to address the unknown virus variants of ongoing disease.

On the other hand, it is essential to note that Chinese morphemes are mainly monosyllable. Hence, each syllable corresponds to a different character according to its meaning. Since the collection of Chinese characters has remained relatively steady for nearly two thousand years, it is scarce to introduce new forms at the character/morpheme level. Moreover, since the semantics are related to orthography, it is infrequent for the Chinese language to completely deviate the new meaning from the particular character/morpheme's conventional meaning. Therefore, a new referent such as *xīn guān fèiyán* novel corona pneumonia 'COVID-19' should be more or less closely linked to the fundamental meanings of competing for linguistic forms, thereby providing rich linguistic contexts to observe the development of different linguistic variants at different pandemic stages.

3.2 Data Sources and Collection

Four types of data sources are used in the thesis, i.e., the Baidu website and Baidu Index, the Baidu News, the COVID-19 official website of China's CDC (Chinese Center for Disease Control and Prevention), and Sina Microblog.

[3] This system reports the diagnosed cases electronically by the responsible doctor (Report of the WHO-China Joint Mission on COVID-19, 2020)

3.2.1 Baidu Website and Baidu Index

The Baidu website is one of the most important search websites to the Chinese people. It provides a primary source for the candidate of emergent neologisms to refer to the COVID-19. The Baidu Index (http://index.baidu.com) is a data analysis platform that records searches and use by the Baidu netizens. It provides the frequencies for the emergent neologisms.

Like Google Trends, the Baidu Index is one of the most comprehensive platforms providing statistics on internet usage and search data worldwide. Choosing Baidu Index rather than Google Trends is that Baidu Index is the most appropriate platform in recording Chinese people's searching activities. Chinese mainlanders, an enormous population compared with Hong Kong, Macau, and Taiwan in Greater China, have easier access to Baidu, which naturally makes the Baidu Index more accurate and accessible in terms of the data records of internet usage in China. There are also search frequency recorders for Hong Kong, Macau, and Taiwan users. Baidu Index is good at offering the frequencies of a keyword searched by Baidu users at a specific date, making it possible for our focus on the longitudinal development of COVID-19 emergent neologism over time. In addition, it can provide metadata such as the proportions of users' gender, age, profession, and location issues of the searches. After registering an account, Baidu Index is an open-access source, so our thesis based on Baidu Index data is easy to reproduce. There is no personal information violation concerning the data ethics because only summarized statistics are available on Baidu Index.

We used Baidu Index to collect the usage frequencies of COVID-19 emergent neologisms and buzzwords including vector names, and PPE names. Since the Wuhan government identified the first date of confirmed COVID-19 cases on December 21st, 2019, our data collection for emergent neologisms was also set from December 21st, 2019. Regarding the time intervals, there

are two reasons for our choice of every five days. One is derived from the result of our pilot study that five-day intervals presented the apparent trends of usage change and, simultaneously, took account of subtle shifts by comparison with ten-day intervals. On the other hand, Millar (2009) argued not to choose everyday intervals in longitudinal corpus studies of their potential inaccuracies. The Baidu Index data is updated daily, and the time intervals for circulating the news might take at least two days, thereby causing the inappropriateness of data collection of everyday intervals. Our data collection started on December 21st, 2019, followed by December 26th, 2019, *etc*. The end date for our data collection is March 30th, 2021 (the date for completing the thesis), which takes 15-month data for our analysis. The search areas in our thesis are set in Greater China, including Hong Kong, Macau, and Taiwan.

One thing that needs attention is that the search frequencies for each type of COVID-19 emergent neologisms and buzzwords are unnecessary for all the use related to COVID-19. For example, the search frequencies (100, 000 occurrences) of *biānfú* 'bat' on the Baidu Index do not mean that there are 100, 000 times searched by Chinese netizens to know how *biānfú* 'bat' caused COVID-19. The reasons why the current thesis still uses the internet searching frequencies on the Baidu Index to represent the searches by the Chinese public to mention words related to the pandemic are twofold. For one thing, it is challenging and almost impossible to separate the unrelated words linked to the pandemic from the related words linked to the pandemic. For another, the COVID-19 pandemic has been a heatedly discussed topic among the Chinese public since the pandemic outbreak till the current time. So, internet searches on the words related to the ongoing disease should dominate their words unrelated to the ongoing disease. Even if the search frequencies such as *biānfú* 'bat' collected in the present thesis also include the other references like the term *biānfú* 'bat' in the context of introducing animals, the

dominating frequencies related to the COVID-19 pandemic can capture the general development of pandemic-related emergent neologisms.

3.2.2 Baidu News

Baidu News (https://news.baidu.com/) is an important platform for recording millions of Chinese news at home and abroad. It includes varieties of columns, e.g, hot news, regional news, domestic news, international news, entertainment news, sports news, financial news, scientific and technological news, army news, *etc*. Besides, users can also access the news by the trending news searches based on key words. Baidu News provides the information for important policies.

3.2.3 The Official COVID-19 Pandemic Website

The data of the official COVID-19 pandemic website is given by the National Hygiene and Health Committee and the Hygiene and Health Committee of all provinces (http://2019ncov.chinacdc.cn/). This website updates the number of pandemic cases every day for Greater China. The trends of emergent neologism searches on the Baidu Index are not developed linearly over time, but the cumulative cases can only show constant increases. We only selected newly confirmed cases, new suspected cases, new deaths, and currently suspected cases for further analysis. There is one issue that should be given attention: The official COVID-19 pandemic website provides the statistics from January 16th, 2020, which is a little later than when all three types of emergent neologisms appeared on the Baidu Index. However, the two dates are both critical for this thesis.

For one thing, the qualitative part needs to link the possible reasons behind the competition and development of COVID-19 emergent neologisms related to governmental-level policy

implementation, so the data of the internet searches beginning on December 21st, 2019 needs to remain for the qualitative analysis. The most crucial part of our thesis is the quantitative part, which models the mapping relationship between COVID-19 emergent neologisms (i.e., internet searches on the Baidu Index) and pandemic development (i.e., new cases of the COVID-19 pandemic) by correlation and regression analyses. To guarantee the comparability of the data of emergent neologisms and pandemic cases, we executed every five-day searches on the COVID-19 pandemic cases and, at the same time, extracted the internet searching data since January 16th, 2020. In this way, two data used for the quantitative part of regression began on January 16th, 2020, and ended on March 25th, 2021, with intervals of five days.

3.2.4 Sina Microblog

The other quantitative part of our thesis is to prove the acceptability of our assumption of how the change of public attention to face mask-wearing over time. Hence, the fourth data source derives from Sina Microblog. As we introduced in Chapter 2, Sina Microblog is one of China's leading social media platforms. Posting comments and feelings has become very popular on Sina Microblog among Chinese people. To extract the public attitudes to wearing face masks, we set advanced searches based on the four columns, i.e., keywords, types, contents, and periods for users. The column of "types" includes famous, original, individual, Very Important Persons, media, and opinions; the column of "contents" includes contents with pictures, contents with videos, contents with music, and contents with short links. The column of "periods" can customize at any time according to users' purposes.

The keywords are *dài kǒuzhào* 'wear masks' because of two aspects. On the one hand, *dài kǒuzhào* 'wear masks' has already become an inevitable living condition for Chinese people

under the ongoing pandemic. Hence, tracking public attitudes to mask-wearing can help the government to monitor the public's psychological change at different stages of the pandemic, policy announcement, and vaccination injection. This motivates me to capture richer contextual information to explore *dài kǒuzhào* 'wear masks'. We did not include other variants of *dài kǒuzhào* 'wear masks' in the Sina Microblog because this part is just to verify the acceptability of using the general development of COVID-19 emergent neologisms to predict the change of public emotions.

Since we focused on the general public's sentiment rather than the attitude of government departments and official accounts to *dài kǒuzhào* 'wear masks', we narrowed the setting down to "individuals" on the Sina Microblog web page. We set the beginning time as them set in the internet searches, i.e., December 21st, 2019, for the first date's data collection. Posts have already provided sufficient information about how the public deals with *dài kǒuzhào* 'wear masks', so only posts were scraped without comments. However, the time point selected from the Sina Microblog is different from the continuous time intervals in the data collection of the Baidu Index. Given that Sina Microblog data is used to verify the psychological states predicted by the use of COVID-19 emergent neologisms and the main purpose of the thesis is not to explore the public emotional change over time, so only critical time points (i.e., discrete data) were selected and used for the verification rather than continuous data. The time points chosen for this part were involved in three years, i.e., December 21st at 2019, four-time points in 2020 (i.e., January 25th, 2020, May 4th, 2020, August 7th, 2020, and December 31st, 2020), and three-time points at 2021 (i.e., January 4th, 2021, May 29th, 2021, and August 2nd, 2021).

Except for considering the time point used in the tracking of COVID-19 emergent neologisms, the central principles of the above time point selection also lie at the factors of

seasonality and population aggregation. Research has shown that the low temperature might be one of the most predisposing factors of aggravating the pandemic. By contrast, summer might make the pandemic more soothing to some extent in the Northern Hemisphere (Callaway *et al.*, 2020). However, hot summer seems 'contradictory' with the behavior of wearing masks, so our interest is how the season would influence the change of public's emotion. Holidays are another salient factor that may probably influence the pandemic development because they are likely to cause congestion. Specifically, the day of officially announcing the first COVID-19 pandemic case (December 21st, 2019) should undoubtedly be the first date for data crawling of the Sina Microblog data collection because the pandemic issue has not been widely spread, which might become a baseline for comparing the public's attention on mask-wearing at the following time point. Among the four critical time points in 2020, we tried to cover the most severe and relatively soothing pandemic period to correspond to the noticeable change in the development of COVID-19 emergent neologisms. January 25th, 2020 is the Chinese Lunar New Year and also the first day of announcing the complete lock-down within the Chinese mainland (Regan *et al.*, 2020), which we thought might be one of the most serious time points. Then, according to the author's first-hand experience, May 4th, 2020 is the date that most universities allowed students and school staff to return to school due to the effective control of the pandemic, which causes crowd assembling. August 7th, 2020 is the deep summer and the summer vacation for universities and schools, so people in many southern provinces of China may generate negative emotions about mask wearing. The public's attitude to mask wearing may starkly contrast with it under the most severe stage. We also collected blog posts by individuals on December 31st, 2020 when people may emerge with complex emotions on the last day of the year 2020, including sighing,

wishing, *etc*. In addition, at the end of December, the first batch of vaccination went to the Chinese market, possibly making the public's emotion change to some extent.

The above time point selection is generally consistent with the period extracted by recent work on the public's sentimental change to the COVID-19 pandemic development. For example, Tan *et al.* (2021) focused on the whole of 2020 and found a generally long-term negative attitude among Chinese mainlanders. Tran *et al.*'s (2021) study from December 2019 to November 2020 showed that Vietnam people's sentiment changed from negative to positive polarity. The change supported the Issue-Attention Cycle: little attention at the first beginning of the issue, increasing and even chaotic attention in the middle of the issue, and a steady drop at the post-issue stage and the final stage of attention replacement.

However, at the end of 2020 and the beginning of 2021, the vaccination was introduced in China. The Chinese government fully encouraged people to inject first and second doses to increase protection rates until August 2021. On the other hand, constant news reported that there is still a probability of being infected after getting vaccinated because of the mutated virus. Whether these two important issues may mediate the applicability of the Cycle to the public's attention under different waves of the disease attracts us to have a further exploration. Hence, our Sina Microblog data was extended to the other three time points in the year 2021, so that the applicability of the Issue-Attention Cycle can be re-examined based on the COVID-19 pandemic data on Sina Microblog.

The extended time points are January 4th, 2021 (the first day after New Year Holiday), May 29th, 2021 (Corresponding to May 4th, 2020), as well as August 2nd, 2021 (the second Summer Term since the outbreak of the pandemic). One reason for selecting these three time points as an extension is to compare the public attitude before and after the vaccination. The other

consideration is based on the author's first-hand experience to mask-wearing after the vaccination. The vaccination could prevent the public from being infected by the virus to a significant extent. On the other hand, the injection of vaccination might not provide once-for-all protection rates, so mask-wearing would become an inevitable behavior for every person in the future. Hence, it is also interesting to examine if the general public has been gradually accustomed to such long-term behavior of mask-wearing. More details of the pre-processing of Sina Microblog data and the algorithm for extracting the N-gram co-occurrence with the search phrase *dài kǒuzhào* 'wear masks' were given in Section 3.5.

3.3 Data Analyses in the Qualitative Part

The qualitative part related to RQs 1-2 aims to explore whether COVID-19 emergent neologisms could reflect important policy announcement during the past one and half year period. Before the attribution, this part first employed the Baidu website and Baidu Index to collect the search frequencies of COVID-19 emergent neologisms and buzzwords. The frequency distribution of COVID-19 emergent neologisms is used to answer RQ 1. Since RQs 1-2 also involves the comparison between emergent neologisms and buzzwords in reflecting the policy announcement, the frequency distribution of buzzwords will be also presented.

On the Baidu website, we searched for the most commonly accepted and used word/phrase in each category of COVID-19 emergent neologisms. For example, *xīn xíng guān zhuàng bìngdú* 'COVID-19' in the COVID-19 emergent neologisms, *biānfú* 'bat' in the vector names, and *yī yòng wàikē kǒuzhào* 'surgical mask' in the PPE names, as the first search word/phrase. With the above prototypical word/phrase as the search terms, similar phrases occurred accordingly on the Baidu web page, such as the abbreviated name *xīn guān* 'COVID-19', the unspecified name

yìqíng 'pandemic', or the full name *xīn xíng guān zhuàng bìngdú fèiyán novel type corona shape virus pneumonia* 'COVID-19' in the COVID-19 emergent neologisms; the name containing human-to-human transmission *chāojí chuánbōzhě* 'superspreaders' in the vector names; and the name about hands protection equipment *jiǔjīng xiāodúyè* 'alcohol disinfectant' in the PPE names. Meanwhile, the WHO website (Chinese version, https://www.who.int/zh/home) also provides terms about the disease, vectors, and PPE, allowing further access to collect more nomenclatures of these three types of terms. Through such a bottom-up method, we could collect an initial word/phrase list for each type of COVID-19 emergent neologisms and buzzwords.

Secondly, we double-checked the phrases selected from the above bottom-up method with the words/phrases recorded in Baidu Index because only the words/phrases with frequency records in Baidu Index could be used for further analysis. The word/phrase lists used for further analysis in the COVID-19 emergent neologisms and buzzwords including vector names and PPE names are shown in Table 1- a, 1- b, and 1-c, respectively.

Table 1 Categorization of COVID-19 emergent neologisms and buzzwords

Table 1- a Emergent neologisms' categories and specific terms (adapted from Lei *et al.*, 2021)

Under-specification	Pre-official names	Stigmatizing names	Official names	English abbreviations
yìqíng 'pandemic'	*bùmíng yuányīn fèiyán* unknown reason pneumonia 'pneumonia of unknown reasons'	*wǔhàn fèiyán* 'Wuhan pneumonia'	*xīn xíng guān zhuàng bìngdú fèiyán* novel type crown shape virus pneumonia 'COVID-19'	COVID-19
fèiyán 'pneumonia'	*bìngdú xíng fèiyán* virus type pneumonia 'viral pneumonia'	*wǔhàn bìngdú xíng fèiyán* Wuhan virus type pneumonia 'Wuhan viral pneumonia'	*xīn guān fèiyán* novel crown pneumonia 'COVID-19'	2019-nCov
bìngdú 'virus'	*xīn xíng bìngdú* novel type virus 'new type of virus'	*zhōngguó bìngdú* 'Chinese virus'	*xīn guān yìqíng* novel crown pandemic 'COVID-19'	Coronavirus

Table 1-a Continued

Under-specification	Pre-official names	Stigmatizing names	Official names	English abbreviations
	xīn xíng fèiyán novel type pneumonia 'new type pneumonia *guān zhuàng bìngdú* crown shape virus 'corona type virus'	*wǔhàn xīn xíng fèiyán* Wuhan novel type pneumonia 'Wuhan new type pneumonia' *wǔhàn bìngdú* 'Wuhan virus'	*xīn guān bìngdú* novel crown virus 'COVID-19' *xīn guān* novel crown 'COVID-19' *xīn xíng guān zhuàng bìngdú* novel type crown shape virus 'COVID-19' *2019 xīn xíng guān zhuàng bìngdú* 2019 novel type crown shape virus 'COVID-19'	SARS-CoV-2

Table 1- b Vector names' categories and specific terms

Animals	Humans
yěwèi 'venison'	*yísì bìnglì* 'suspected patients'
guǒzilǐ 'civet'	*chāojí chuánbōzhě* 'super spreaders'
biānfú 'bat'	*mìqiè jiēchùzhě* close contactee 'persons with close contact'
zhúshǔ 'bamboo rat'	*gǎnrǎnzhě* 'infector'
huān 'badger'	*bìngdú xiédàizhě* 'virus carriers'
sùzhǔ 'host'	
shé 'snake'	
dàxingxing 'gorilla'	
shǔ 'mouse'	
yěshēng dòngwù wild animal 'wildlife'	
shuǐdiāo 'mink'	
chuānshānjiǎ 'pangolin'	

Table 1- c PPE names' categories and specific terms

Hand protection	Eye protection	Face protection	Body protection
shǒutào 'glove' *yī yòng shǒutào* medical purpose glove 'medical glove' *xiāodúyè* 'disinfectant' *kàngjūn xǐshǒuyè* 'anti-bacterial hand sanitizer' *xǐshǒuyè* 'hand sanitizer' *zhǐjīn* 'tissue' *jiǔjīng* 'alcohol'	*hùmùjìng* 'goggle'	*N95kǒuzhào* 'N95 face mask' *kǒuzhào* 'face mask' *yī yòng wàikē kǒuzhào* medical purpose surgery face mask 'medical surgical mask' *wàikē kǒuzhào* 'surgical mask' *yīcìxìng kǒuzhào* 'one-off mask' *fángdú miànjù* antivirus mask 'gas mask'	*fánghùfú 'protective suit'*

As Table 1-a, 1-b, and 1-c indicate, we listed the terms and their categories for further data analysis. For Table 1a, we borrowed the COVID-19 referring terms and the five categorizations proposed by Lei *et al.* (2021). Under-specifications do not specify the disease, such as *yìqíng* 'pandemic', *fèiyán* 'pneumonia,' and *bìngdú* 'virus'. Pre-official names are those used before WHO gave the disease an official name (February 11st, 202). Since the virus was still unknown to the public, pre-official names contain unknown and uncertain information in the names such as *bùmíng yuányīn fèiyán* 'pneumonia of unknown reason'. Even if the term of *bìngdú xíng fèiyán* 'viral pneumonia' representing its similar symptom to pneumonia still does not point out the nature of this novel coronavirus, which is also included in the pre-official names. Before the birth of official names, another category took the biggest proportion, exceeding the usage of under-specifications, which are labeled as stigmatizing names because they include specific country or region names in the disease reference with the stigmatized bias (Ghebreyesus, 2020). The stigmatizing names include the terms like *wǔhàn bìngdú* 'Wuhan virus' and z*hōngguó bìngdú* 'Chinese virus'. Based on the concerted efforts by all walks of life, human beings have a deeper understanding of the pandemic. For example, researchers have proven that this novel disease has a corona shape (e.g., Mousavizadeh & Ghasemi, 2021), so the emergent neologisms reflecting these two unique characteristics by using *xīn* 'novel' and *guān* 'corona' are all included in official names such as the full name used by the government *xīn xíng guān zhuàng bìngdú* novel type crown shape virus 'COVID-19' or abbreviated one *xīn guān* novel crown 'COVID-19'. With the emergence of Chinese official names, English versions also came out. They are primarily abbreviations such as COVID-19. They are called English abbreviations.

The classification for vectors and PPE names adopts the categories by WHO, CDC (Centers for Disease Control and Prevention, https://www.cdc.gov/coronavirus/2019-ncov/daily-life-

coping/animals.html), Wikipedia, Occupational Safety and Health Council (2021). They divided vector names by humanness. At the beginning of the pandemic, as we have already reviewed in Chapter Two, a variety of animals have been discovered successively to cause and transmit the disease. As scientific findings were further reported, human-to-human transmission has also been found and confirmed to be a considerable risk for disease transmission. Hence, the classification of vector names is naturally led by animals and humans.

As for PPE names, they have the previous classification recorded by Wikipedia (https://en.wikipedia.org/wiki/Personal_protective_equipment) and Occupational Safety and Health Council (2021). It has already been categorized into different equipment to correspond to certain protective functions for different body parts. Hence, our thesis categorizes PPE terms into hand PPE terms, eyes PPE terms, face PPE terms, and body PPE terms.

After clarifying the categories of the COVID-19 emergent neologisms and buzzwords, we were then expected to list the important dates for popular social events in order to answer RQ 2. Table 2 lists the critical dates that could be used to explain the development and competition of COVID-19 emergent neologisms during the pandemic. The main reason for selecting these dates derives from the possible impacts of the first case of infection, the entire suspension on companies and institutions, the announcement for the official nomenclature of COVID-19, and the vaccination marketing. Specifically, the date (December 21st, 2019) reporting the first case of being infected by the pandemic should be a starting time for data analysis, so there is no doubt listing it as the first time point in Table 2. January 21st, 2020 is also an important date because it showed the significant research findings on how the virus survives and transmits. This date also anticipated that there would be an unprecedented outbreak by the Institute Pasteur of Shanghai

and the Institute of Plant Physiology and Ecology, Chinese Academy of Sciences. This date may directly lead to the complete suspension of companies and institutions in mainland China.

The recent work conducted by Mallapaty (2022) published on *Nature* indicated that horseshoe bats as the probable infectious agent might promote the process of the official terminology to the disease by WHO on February 3rd, 2020. February 8th and February 11st, 2020 are important because it is the time for proposing the Chinese official name to refer to the COVID-19 and its English version by National Health Commission of the People's Republic of China and WHO. May (1st - 5th May) and August (1st - 20th August) are two critical time periods for full recovery of work and study and summer term for all school staff and students, respectively, which we consider might also stimulate the change of the general public's emotion due to population aggregation and hot weather. Another significant event during the pandemic is vaccination. Sinovac went to the Chinese market. Meanwhile, the Chinese government encouraged Chinese mainlanders to get the injection, which influences the discussion of pandemic-related topics and the emotional change in response to the pandemic.

After clarifying the categories of the emergent neologisms and buzzwords used in the present study, I listed the important dates for significant social events collected from Baidu News, as Table 2 shown. The significant social events are used to respond to the proportion of emergent neologisms over time.

Table 2 Important dates for significant events during the COVID-19 pandemic

Time	Events
2019.12.21	The first case of being infected by COVID-19
2020.1.21	*SCIENCE CHINA Life Sciences*: The virus belongs to Betacoronavirus, similar but different from SARS. It can become a parasite on humans and other advanced animals.
2020.1.23	Complete suspension of work and study
2020.2.3	*Nature*: Conavirus of RaTG13 was found in the horseshoe bats
2020.2.8	National Health Commission of the People's Republic of China: *xīn xíng guān zhuàng bìngdú fèiyán* (full name), *xīn guān fèiyán* (short name)
2020.2.11	International Committee on Taxonomy of Viruses (ICTV): SARS-CoV-2; WHO: COVID-19, do not include region/country in the disease nomenclature
2020.5.1-5	Almost all work and schools have recovered to the normal working mode in mainland China.
2020.8.1-20	Summer Term for school staff and students
2020.12.31	Sinovac went on the market.
2021.1.1-2	Chinese New Year, Chinese Lunar New Year, Peak to return home

Table 2 also adds the New Year in 2021. In the Chinese New Year of 2020, the Chinese mainlanders experienced complete home quarantine on the mainland; by contrast, in New Year in 2021, the government allowed a small range of gathering and travel. So, this time point added here was to compare the development of emergent neologisms at two consecutive New Year with different social distancing policies. In addition, New Year's time is also the cold weather, so whether the gathering and low temperature provided a potential for the outbreak of a new wave is worth further exploration. We ended the list of important dates on January 2021 rather than the end of March (the ending date in emergent neologism data) because there was no significant social event related to the COVID-19 happening in March. Meanwhile, emergent neologisms from January to March 2021 showed similar developmental patterns, so we stopped listing important dates on the Chinese New Year as shown in Table 2.

3.4 Data Analyses in the Regression

The part of regression modeling involving RQs 3-4 aims to answer whether COVID-19 emergent neologisms could correlate and predict the pandemic development over the past one year and

half and whether COVID-19 emergent neologisms can be more predictable than buzzwords to the pandemic prediction.

3.4.1 Pearson Correlation

RQ 3 investigates the correlation between emergent neologisms and pandemic cases, and buzzwords and pandemic cases. Pearson Correlation is used to explore the correlated relationship. Among many types of the correlation coefficient, Pearson Correlation is the most common one (David, 2014), an acronym for Pearson Product Moment Correlation. It shows the linear relationship between independent variables and dependent variables. Equation [1] indicates the formula of Pearson Correlation.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2}} \qquad [1]$$

The correlation coefficient uses the absolute value of *r* to express the correlation strength and the positive/negative to represent the direction. The value of *r* ranges between -1 and 1: -1 means the strongest negative correlation between inputs and outputs, whereas 1 means the strongest positive correlation. 0 represents no association between inputs and outputs. The other two correlation coefficients are also critical to represent the different extent of correlation strength. \|*r*\| = 0.3 means the boundary of moderate correlation, while \|*r*\| = 0.7 demonstrates the boundary of solid correlation. Hence, the absolute value above 0.7 demonstrates a strong correlation between inputs and outputs. By contrast, the absolute value below 0.3 stands for a very weak correlation between inputs and outputs. The absolute values of correlation coefficients between 0.3 and 0.7

indicate moderate correlation. The PROC CORR[4] function computes Pearson Correlation on the SAS Studio (https://www.sas.com/en_hk/home.html), one free and open-access programming platform.

Before conducting the regression modeling, there are four premises to satisfy. First, emergent neologisms and pandemic cases should have linear relationship. This premise is checked by *linearity_test()* in Python. Second, errors should not have autocorrelation with independent variables and their expectation is around 0. This premise is checked by *lin_reg.resid.mean()* in Python. Third, independent variables should not have full multicollinearity issue. This premise is checked by *variance_inflation_factor()* in Python. Fourth, the errors should be of homoscedasticity, which is checked by *homoscedasticity_test ()* in Python. On the basis of four-premise check, the next section will explore the predictability of COVID-19 emergent neologisms to the pandemic development compared to buzzwords based on multiple regression models.

3.4.2 Sigmoidal Function of Log-Linear Regression

RQ 4 examines whether COVID-19 emergent neologisms could predict the pandemic development. The prediction is a very important part of the present thesis, though the second goal. In addition to directly sketch the contour of the development of emergent neologisms at the pandemic period (which will be reported in Chapter Six), the theoretical issue of whether the S-curve can be applied to the development of COVID-19 emergent neologisms at the pandemic period can also be indirectly answered by finding a better curve that can better fit the mathematical relationship between emergent neologisms and pandemic cases. If the fitting curve

[4] The functions of Pearson Correlation used in SAS Studio are case insensitive.

is not as the S-curve develops, it indicates that the S-curve is not applied and vice versa. To pinpoint a "better" model, it is necessary to compare the model performance based on varieties of regression models.

Log-linear regression is a nonlinear regression by the sigmoidal function. It is one of the activation functions, an essential part of any neural network in deep learning. Its working mechanisms convert the inputs into outputs with a range of 0 to 1 to effectively save computation time. It has been widely applied in the automatic diagnosis of diseases by researchers such as Bogard *et al.* (2019) and Arbab *et al.* (2021). Its efficiency lies at its imitation of human brain processing at maximum. When an input enters human brains, it would not be processed linearly, namely, copy-paste. Instead, the input would be simplified based on the human's background knowledge to make the outputs easy to remember and understand. The equation of the sigmoidal function is as [2] shown.

$$y = \frac{1}{1+e^{-z}} \qquad [2]$$

The visualization of Equation 2 is similar to a letter S, but being displaced. When the value of z infinitely approximates a positive number, the predicted value of y will become 1. By contrast, when infinitely approximating a negative number, the predicted value of y will become 0. If the outcome of the sigmoidal function is above 0.5, then we can label it as the positive axis; otherwise, it will go to the negative axis. To computer the sigmoidal expression of log-linear regression, we used PROC CATMOD.

3.4.3 Simple Linear Regression

The 'linear regression' here refers to the simple linear regression where the relationship between one independent variable and one observed dependent variable is assumed to be linear. The equation of simple linear regression is as [3] shown.

$$y = \beta_0 + \beta_1 x + \varepsilon \quad [3]$$

where y is the predicted data, x is the independent variable, ε is the error term, β_0 is the constant, and β_1 represents the regression coefficient. PROC REG computes simple linear regression.

However, the world is complex and dynamic (Larsen-Freeman, 1997). The complexity of the pandemic development will be simplified and neglected by only considering a single independent variable in the regression model. Hence, the next section considers the occasions where independent variables are multiple in order to respond to the complexity of the pandemic development.

3.4.4 Multiple Linear Regression

Multiple linear and simple linear regression models both have one dependent variable with the continuous data type, but the difference between them lies in the number of independent variables. Corresponding coefficients can reflect the strength of each independent variable's effect on the target variable.

Generally speaking, Equation [4] presents how multiple linear regression is expressed.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon \quad [4]$$

Like the simple linear regression, y is the predicted data, ε is the error term, and β_0 is the constant. Unlike simple linear regression, it has many different regression coefficients corresponding to many independent variables (i.e., $x_1,x_2...x_k$). PROC GLM computes multiple linear regression.

Since many independent variables are likely to bring about the world to be complex and dynamic (Larsen-Freeman, 1997), there are inevitably many occasions where independent variables are multiple. We should consider how to feed them into the model, i.e., whether selecting some or putting all, and how to put them into the model, i.e., forward, backward, or other operations. In addition to the linear expression, we should also consider polynomial expressions since nonlinearity may be more appropriate to describe the complex relationship between COVID-19 emergent neologisms and collective human behavior (e.g., Lei *et al.*, 2021; Paun, 2021). Such considerations aim to maximize the model predictability and interpretability for the mathematical relationship between multiple independent variables and the target variable.

The following will introduce various fine-tuned regression models in order to develop their respective advantages in dealing with the weights of multiple independent variables, i.e., COVID-19 emergent neologisms, vector names, and PPE names.

3.4.5 Fine-Tuned Regressions

We fine-tuned multiple regression models by four widely used methods, i.e., stepwise regressions, regularization, the least angle regression, and polynomial methods. They have respective advantages in dealing with the issue of multiple independent variables.

Stepwise regression

Stepwise methods feed the regression models by automatic selection processes of independent variables (Efroymson, 1960). A variable is added to or subtracted from a set of independent variables based on pre-specified criteria. Two typical stepwise regression methods are forward selection and backward elimination.

Forward selection begins with an empty model with only an intercept. It adds each independent variable into the model each time. In each step where a new variable is added to the regression model, the model performance would be tested based on criteria (i.e., R^2 and RMSE were introduced in Section 3.4.5). The process would stop once the model performance does not improve at all with a new variable included. By contrast, backward elimination has an opposite mechanism: starting with a model with all variables and eliminating each independent variable each time. In each step, an existing variable is removed from the regression. The model performance would also be tested based on R^2 and RMSE. The modeling process would end once the model performance does not improve with removing an existing variable. According to R^2 and RMSE, the variables beneficial for the model predictability could be automatically extracted. We used SELECTION=FORWARD and SELECTION=BACKWARD in the model statement of SAS Studio to compute forward selection and backward elimination, respectively.

The stepwise method in SAS Studio also provides the statement SELECTION=STEPWISE, which combines the above two methods. In this model, if any effect is shown statistically insignificant, the least significant independent variable would be removed from the model, and the algorithm continues to the next step. The above operation ensures that less important independent variables could not be added to a model. At the same time, it also shows that some independent variables with statistical significance temporarily in the model are not necessarily

significant to a model. After the removal approach, stepwise regression will automatically begin the addition operation. Similar to the first cycle (i.e., removal approach), the effect that the addition approach finds the most statistically significant value will be added to the model, and the algorithm continues to the next step. The stepwise process will end when no effect outside the model is statistically significant, and every effect in the model is statistically significant. Compared with the separate operation of the forward selection and the backward elimination, stepwise regression combines them by completing all removal operations and then automatically beginning the addition operation to gain the most significant independent variables and effect sizes to the model.

Generally, stepwise regression boasts three types of advantages. Firstly, it can improve the model generalizability by subtracting less important independent variables from the model based on this stepwise method. Secondly, it is easy for the result interpretation. A small and simple model would undoubtedly have better interpretability than a complicated model. By reducing the dimensions of independent variables, the stepwise methods can provide a 'white box' model where the weights of all combinations of independent variables can be tracked. Last but not least, it is reproducible. Stepwise regression is more reproducible and objective than selecting independent variables manually based on expert opinions.

Nevertheless, automatic selection of independent variables cannot equal to expert opinions. Further analysis based on researchers' background knowledge is still necessary to help to analyze the entry or removal of certain variables from the model. Besides, we should also note the limitations of stepwise selection. Firstly, it cannot ensure to that the best possible combinations of variables are selected despite the computation advantage. Secondly, the outputs might be biased regression coefficients, confidence intervals, *p*-values, and R^2. With the

increased number of variables entering from the model, regression coefficients and R^2 would become larger, while confidence intervals and *p*-values would become smaller, causing a false-positive result. That is the reason why Hoyt *et al.* (2016), the developers of SPSS, announced "the significance values [a.k.a. *p*-values] are generally invalid when a stepwise method (stepwise, forward, or backward) is used (IBM Knowledge Center)" (https://www.ibm.com/docs/en/spss-statistics/29.0.0?topic=regression-linear-variable-selection-methods). Thirdly, it might not specify the problem case by case. For example, suppose the variable is labeled as a confounding factor by the stepwise method. In that case, it has theoretical meaning, but without high R^2, removing it from the model would be problematic.

To compensate for the limitations of stepwise methods, we also used regularization, which has been widely applied in clinical prediction models (Steyerberg, 2009).

Regularization

The term 'regularization' means to make things regular and acceptable (Mishra, 2018). Based on the fundamental formula of multiple linear regressions, a regularization term is used to penalize the errors caused by overfitting[5]. There are four types of regularization methods provided by SAS Studio, i.e., L1 regularization, L2 regularization, elastic regularization, and adaptive L1. L1 regularization, also known as LASSO (Least Absolute Shrinkage and Selection Operator), is a regression technique introduced first in geophysics (Santosa & Symes, 1986). The term LASSO coined by Tibshirani (1996) is used to add an L1 penalty equal to the coefficient value to restrict the coefficients' size and remove some least essential coefficients. Similarly, L2 regularization, known as Ridge Regression, adds a penalty L2 (i.e., the square of coefficients) after the loss

[5] Overfitting means that the model fits the observed data very well but does not perform accurately in the predicted data (IBM Cloud Education, 2021).

function of simple and multiple linear regression (i.e., the left of the plus sign). Equations for L1 and L2 regularization are as Equations [5] and [6] demonstrated, respectively.

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda\sum_{j=0}^{M}\left|W_j\right| \qquad [5]$$

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda\sum_{j=0}^{M}W_j^{\,2} \qquad [6]$$

The big difference between Equations [5] and [6] lies in the [5] estimating the data's median, while [6] estimating the data's mean. According to Equation [5], LASSO adds the penalty term in the cost function by adding the absolute value of weight (W_j) parameters. Hence, in L1 regularization, less critical features are reduced from the model. By contrast, L2 regularization adds the squared value of weights (W_j) in the cost function, as [6] shown. By calculating the loss function in the gradient calculation step, the loss function tries to minimize the loss by subtracting it from the mean of the data. Hence, unlike L1 regularization, L2 regularization would give less coefficient to the less important feature. L1 regularization is computed by the model statement SELECTION = LASSO, whereas L2 regularization is computed by SELECTION = RIDGE on SAS Studio.

Another type of regularization regression combines L1 regularization and L2 regularization by adding hyperparameters, called elastic net regression. Elastic Net first emerged by a critique on the LASSO, whose variable selection can be too dependent on data distribution and thus unstable. One possible solution is to combine the penalties of Ridge and LASSO to get the best of both regressions (Zou & Hastie, 2005). Elastic Net aims at minimizing the following loss function (i.e., the left of the plus sign), as [7] shown:

$$L_{enet}(\beta) = \frac{\sum_{i=1}^{n}(y_i - x_i'\beta)^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m}\beta_j^{\,2} + \alpha\sum_{j=1}^{m}|\beta_j| \qquad [7]$$

where α is the mixing parameter between Ridge ($\alpha = 0$) and LASSO ($\alpha = 1$). λ is another parameter. It is computed by SELECTION = ELASTICNET on SAS Studio.

SAS Studio also provides adaptive LASSO regularization. It is a modification of the above LASSO. In adaptive LASSO selection, weights are applied to each parameter constituting the LASSO constraint (Zou, 2006). SELECTION = ADAPTIVELASSO computes it.

We have introduced two types of fine-tune methods, stepwise methods and regularization. However, Efron *et al.* (2004) proposed an algorithmic framework, least angle regression, containing the advantages of the forward selection and LASSO. So, let us understand how the least angle regression works in the next part.

Least angle regression

The least angle regression (LAR) model starts by centering the covariates and target variable and scaling the covariates to guarantee the same corrected sum of squares. All the coefficients are set to zero at the beginning, as $y = y$. Then, the algorithm determines the feature with the strongest correlation with the current residual and takes a step toward this predictor. The length of this step determines the coefficient of this feature. The decision of the step length makes some other features, and the currently predicted target variable has the exact correlation with the current residual. At this point, the predicted target variable moves in the equiangular direction between these two features. This direction ensures that these two features retain an expected correlation with the current residual. The predicted target variable also moves in this direction until a third feature correlates with the current residual as the two already in the model. At this time, the

algorithm determines a new direction that is equiangular among these three features. The predicted target variable moves in this new direction until a fourth feature appears with the exact correlation with the current residual. This process loops until all features are in the model. SELECTION=LAR computes LAR.

Like stepwise methods, it is crucial to decide when to stop the selection process. Unlike the stepwise regression and regularization terms selecting a parsimonious set from a large data set, the LAR employs a more practical and less greedy method to modify LASSO and forward selection by ordinary least squares with more efficient computation time.

Theoretically, the LAR should perform the best. Nevertheless, considering the above stepwise regression and regularization terms have their characteristics in dealing with different input selection, we tried all of the methods to fine-tune the hyperparameters of the linear regression models in data analyses of the quantitative part. R^2 and RMSE decided on the final model.

Polynomial regression

In addition to feature selection methods, we still need to consider the polynomiality of input features. Unsurprisingly, the input feature is not usually linearly correlated with target variables; instead, they might be more complex due to their nonlinearity. Polynomial regression is a technique to fit a nonlinear equation by taking polynomial functions of independent variables. Thus, a polynomial of degree k in one variable is written as Equation [8]:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon \qquad [8]$$

Here, x, x^2, x^k are different independent variables. Mapping the relationship between the target

variable and independent variables can be completed by simple linear regression and multiple linear regression, as mentioned above.

We tried three power types: linear, binomial, and trinomial expressions. The linear relationship is the most widely investigated mathematical relationship. The binomial expression was the best model to describe the mathematical relationship between COVID-19 emergent neologisms and pandemic cases in the first six months after the pandemic outbreak, which is thus needed to be examined on its interpretability in the current thesis with more extended periods. The trinomial expressions are the tracking records. We only listed the setting of polynomiality in COVID-19 emergent neologisms (Table 3-a, 3-b, 3-c), and it is the same as vector names and PPE names.

Table 3 Polynomial operation of COVID-19 emergent neologisms

Table 3- a Linear inputs and outputs with an example of COVID-19 emergent neologisms

Independent Variables	Dependent Variables
Official names	Newly confirmed cases
Official names	Newly suspected cases
Official names	New deaths
Official names	Currently suspected cases
Pre-official names	Newly confirmed cases
Pre-official names	Newly suspected cases
Pre-official names	New deaths
Pre-official names	Currently suspected cases
Underspecifications	Newly confirmed cases
Underspecifications	Newly suspected cases
Underspecifications	New deaths
Underspecifications	Currently suspected cases
Stigmatizing names	Newly confirmed cases
Stigmatizing names	Newly suspected cases
Stigmatizing names	New deaths
Stigmatizing names	Currently suspected cases
English abbreviations	Newly confirmed cases
English abbreviations	Newly suspected cases
English abbreviations	New deaths
English abbreviations	Currently suspected cases

Table 3-a Continued

Independent Variables	Dependent Variables
Official names, Pre-official names, Underspecifications, Stigmatizing names, English abbreviations	Newly confirmed cases
Official names, Pre-official names, Underspecifications, Stigmatizing names, English abbreviations	Newly suspected cases
Official names, Pre-official names, Underspecifications, Stigmatizing names, English abbreviations	New deaths
Official names, Pre-official names, Underspecifications, Stigmatizing names, English abbreviations	Currently suspected cases

Table 3- b Binomial inputs and outputs with an example of COVID-19 emergent neologisms

Independent Variables	Dependent Variables
Official names, Official names^2	Newly confirmed cases
Official names, Official names^2	Newly suspected cases
Official names, Official names^2	New deaths
Official names, Official names^2	Currently suspected cases
Pre-official names, Pre-official names^2	Newly confirmed cases
Pre-official names, Pre-official names^2	Newly suspected cases
Pre-official names, Pre-official names^2	New deaths
Pre-official names, Pre-official names^2	Currently suspected cases
Underspecifications, Underspecifications^2	Newly confirmed cases
Underspecifications, Underspecifications^2	Newly suspected cases
Underspecifications, Underspecifications^2	New deaths
Underspecifications, Underspecifications^2	Currently suspected cases
Stigmatizing names, Stigmatizing names^2	Newly confirmed cases
Stigmatizing names, Stigmatizing names^2	Newly suspected cases
Stigmatizing names, Stigmatizing names^2	New deaths
Stigmatizing names, Stigmatizing names^2	Currently suspected cases
English abbreviations, English abbreviations^2	Newly confirmed cases
English abbreviations, English abbreviations^2	Newly suspected cases
English abbreviations, English abbreviations^2	New deaths
English abbreviations, English abbreviations^2	Currently suspected cases
Official names, Official names^2,Pre-official names, Pre-official names^2, Underspecifications, Underspecifications^2, Stigmatizing names, Stigmatizing names^2, English abbreviations, English abbreviations^2	Newly confirmed cases
Official names, Official names^2, Pre-official names, Pre-official names^2, Underspecifications, Underspecifications^2, Stigmatizing names, Stigmatizing names^2, English abbreviations, English abbreviations^2	Newly suspected cases
Official names, Official names^2, Pre-official names, Pre-official names^2, Underspecifications, Underspecifications^2, Stigmatizing names, Stigmatizing names^2, English abbreviations, English abbreviations^2	New deaths
Official names, Official names^2, Pre-official names, Pre-official names^2, Underspecifications, Underspecifications^2, Stigmatizing names, Stigmatizing names^2, English abbreviations, English abbreviations^2	Currently suspected cases

Table 3- c Trinomial inputs and outputs with an example of COVID-19 emergent neologisms

Independent Variables	Dependent Variables
Official names, Official names^2, Official names^3	Newly confirmed cases
Official names, Official names^2, Official names^3	Newly suspected cases
Official names, Official names^2, Official names^3	New deaths
Official names, Official names^2, Official names^3	Currently suspected cases
Pre-official names, Pre-official names^2, Pre-official names^3	Newly confirmed cases
Pre-official names, Pre-official names^2, Pre-official names^3	Newly suspected cases
Pre-official names, Pre-official names^2, Pre-official names^3	New deaths
Pre-official names, Pre-official names^2, Pre-official names^3	Currently suspected cases
Underspecifications, Underspecifications^2, Underspecifications^3	Newly confirmed cases
Underspecifications, Underspecifications^2, Underspecifications^3	Newly suspected cases
Underspecifications, Underspecifications^2, Underspecifications^3	New deaths
Underspecifications, Underspecifications^2, Underspecifications^3	Currently suspected cases
Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3	Newly confirmed cases
Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3	Newly suspected cases
Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3	New deaths
Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3	Currently suspected cases
English abbreviations, English abbreviations^2, English abbreviations^3	Newly confirmed cases
English abbreviations, English abbreviations^2, English abbreviations^3	Newly suspected cases
English abbreviations, English abbreviations^2, English abbreviations^3	New deaths
English abbreviations, English abbreviations^2, English abbreviations^3	Currently suspected cases
Official names, Official names^2, Official names^3, Pre-official names, Pre-official names^2, Pre-official names^3, Underspecifications, Underspecifications^2, Underspecifications^3, Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3, English abbreviations, English abbreviations^2, English abbreviations^3	Newly confirmed cases
Official names, Official names^2, Official names^3, Pre-official names, Pre-official names^2, Pre-official names^3, Underspecifications, Underspecifications^2, Underspecifications^3, Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3, English abbreviations, English abbreviations^2, English abbreviations^3	Newly suspected cases
Official names, Official names^2, Official names^3, Pre-official names, Pre-official names^2, Pre-official names^3, Underspecifications, Underspecifications^2, Underspecifications^3, Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3, English abbreviations, English abbreviations^2, English abbreviations^3	New deaths
Official names, Official names^2, Official names^3, Pre-official names, Pre-official names^2, Pre-official names^3, Underspecifications, Underspecifications^2, Underspecifications^3, Stigmatizing names, Stigmatizing names^2, Stigmatizing names^3, English abbreviations, English abbreviations^2, English abbreviations^3	Currently suspected cases

In Table 3-b and 3-c, many inputs are binomial or trinomial expressions for the same feature. The correlation analysis between features in the result chapter of the regression part showed statistical significance ($r > 0.7$) when including the same feature with different input powers. Hence, a multicollinearity issue is demonstrated by significant correlation coefficients between various features. The same feature with different powers may cause overfitting problems and further weaken the model's interpretability (Linardatos, Papastefanopoulos, & Kotsiantis, 2021). Our solution to address the multicollinearity issue within features employed stepwise regression, regularization terms, and least angle regression. Given that overfitting problems might cause the false positive modeling result, we used results before the fine-tuning to be the baseline to compare with the model performance after the fined-tuning to decide the final model.

3.4.6 Model Evaluation by RMSE and R^2 and Feature Selection by *p*-value

The two measurements for determining which regression model is better are based on RMSE and R^2. Root Mean Square Error (RMSE) is the standard deviation of the residuals[6]. In other words, it reports how robust the data is around the line of the best fitting. RMSE is commonly used in climatology to forecast and verify regression's experimental results (Laurini, 2018). The equation of RMSE is as [9] shown. RMSE is a standard way to measure the error of a model in prediction.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i\text{^} - y_i)^2}{n}} \qquad [9]$$

where $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values, and $y_1, y_2, \ldots, y_n$ are observed values. n is the number of observations. In general, the lower RMSE is, the better the model is.

[6] Residuals are a measure of how far from the regression line data points are.

The other measurement used in this thesis is R^2, a statistical measure of how close the data is to the fitted regression line. It is also known as the coefficient of determination or the coefficient of multiple determination for multiple regression. The definition of R^2 is reasonably straightforward: it is the percentage of the variation of the target variable that a linear model explains. It can be expressed in Equation [10] as is shown below:

$$R^2 = \text{Explained variation} / \text{Total variation} \quad [10]$$

The values of R^2 are always between 0 and 1. 0 indicates that the model explains none of the variability of the target data around its mean. In contrast, 1 suggests that the model explains all the variability of the target data around its mean. Generally, the higher the R^2 it is, the better the model fits the data.

In addition to R^2 and RMSE, the selection of the features is dependent on the *p*-value in the regression analysis. The feature with *p*-value smaller than 0.05 shows the good interpretability in the model and will be selected as the final feature.

The software *Origin* makes all plots, and SAS Studio operates the Pearson correlation and the above regression models. Appendix A presents the scripts for all the correlation and regression models.

3.5 Data Analyses in the Quantitative Part of Sina Microblog

The part of the textual analysis based on Sina Microblog data involving RQs 5-6 aims to answer whether emergent neologisms could monitor the public emotional change. We crawled the social posts on *dài kǒuzhào* 'wear masks' from one of the biggest social media platforms, Sina Microblog. The following part introduced how the data was scraped from Sina Microblog and how they were analyzed to show the psychological change over time.

3.5.1 Web Scraper

Web scraper-Free Web Scraping (https://chrome.google.com/webstore/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn), a Google extension, is used to collect data from Sina Microblog at different periods. It is a web data extraction tool with an easy point-and-click interface for the modern web. With such a hands-on interface, it can extract thousands of records from a website only taking a few minutes of scraper setup. Web Scraper utilizes a modular structure made of selectors, instructing the scraper on traversing the target site and guiding what data to extract without much effort.

3.5.2 Data Crawling

After installing Web Scraper on Google, we summarized the rule of crawling Sina Microblog. By typing the target phrase *dài kǒuzhào* 'wear masks' in the search box, we could receive the posts about *dài kǒuzhào* as Figure 2 shown.

Figure 2 Screenshot for typing *dài kǒuzhào* 'wear masks'

Then, we used the advanced filter to shift the periods and the posts. As Figure 3 shows, each time point should be set from 0:00 to 23:59 to obtain the post data at that time. We did not put an extended period above 24 hours because Sina Microblog does not provide all the posting information during an extended period. Another consideration is selecting eight-time points rather than continuous time points.

Figure 3 Screenshot for shifting the searches of *dài kǒuzhào* 'wear masks'

In this thesis, Web Scraper is used to collect the social posts at the eight-time points. They are December 21st, 2019, January 25th, 2020, May 4th, 2020, August 7th, 2020, December 31st, 2020, January 4th, 2021, May 29th, 2021, and August 2nd, 2021. Though it does not need codes, it should create the site maps. Using *Ctrl* + *Shift* + *I*, Web Scraper is turned on, as Figure 4 shown. Next, we made a sitemap by inputting the target website and naming a sitemap (See Figure 5). For example, we can create *weibo20191221* as the new sitemap to show all the operations and crawled data on December 21st, 2019. Before setting the menu of the sitemap, we should know the function of every selection. Figure 6 presents the definitions for each column. We added a new selector in the second step and input the corresponding operations, as Figure 7 indicates. One column not included in Figure 5 is *Delay (ms)*, which is how long it takes to delay the time. This selector is created for crawling content and time on every page. After establishing this selector, we created a new one for crawling the post time. By clicking *Element*, we gave the setting in this column in crawling the post time. The same operation returned to the Element and set for a new selector, i.e., contents. The last step is to add a page selector. At the bottom of each page is a 'next page'. As the Element parallel, a new selector was created named page, as Figure 8 demonstrates.

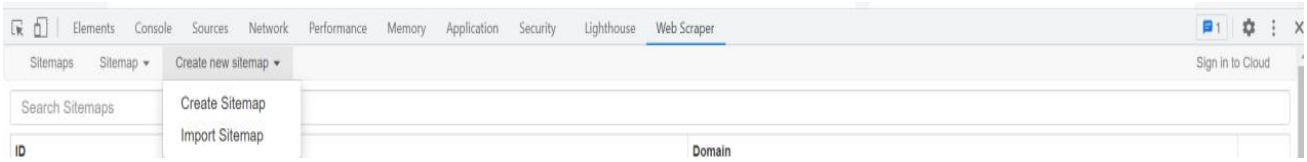

Figure 4 Web Scraper opening page

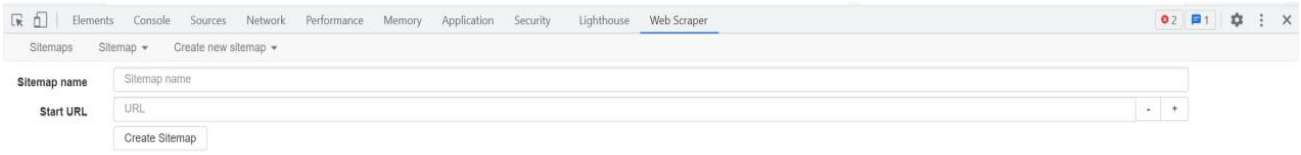

Figure 5 Creating sitemap in Web Scraper

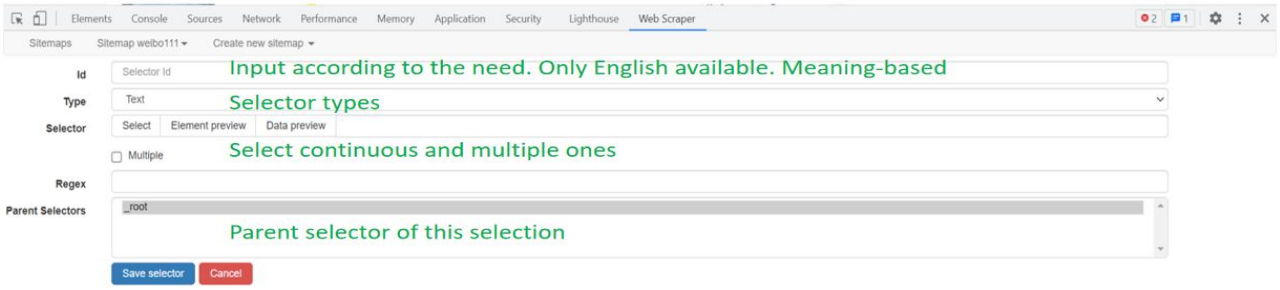

Figure 6 Definition on each column on Web Scraper interface

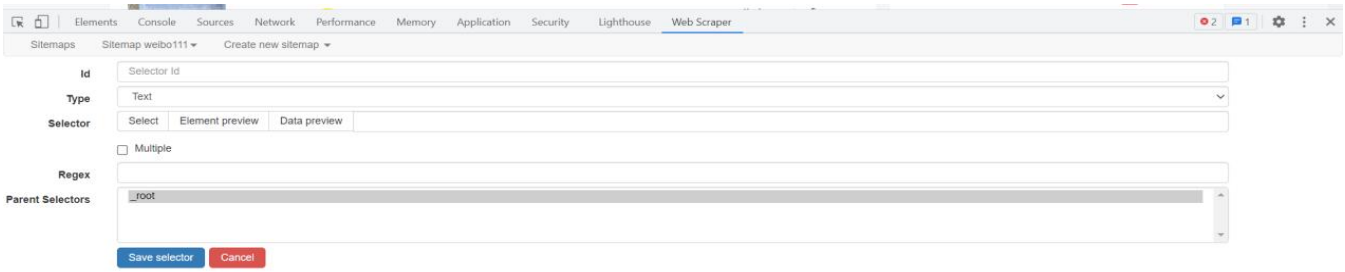

Figure 7 Selector for crawling contents and time

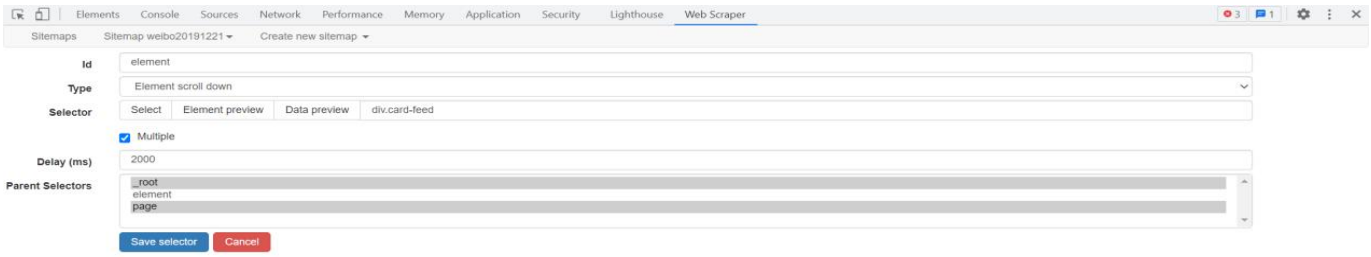

Figure 8 Selector for crawling pages

The above shows the operations for data crawling at a one-time point. They were repeated for the posts at the other seven time points. The basic information for how many posts were used for the following co-occurrence analysis is listed in Table 4.

Table 4 Summary of Sina Microblog Corpus on *dài kǒuzhào*

Time points	Number of posts	Number of Chinese characters
December 21st, 2019	790	26,593
January 25th, 2020	814	52,594
May 4th, 2020	661	46,986
August 7th, 2020	559	37,629
December 31st, 2020	864	23,542
January 4th, 2021	850	59,808
May 29th, 2021	990	73,265
August 2nd, 2021	988	140,386
Total	6,516	460,803

3.5.3 Data Pre-processing

Based on the above procedures, we obtained the Sina Microblog data at eight different time points with *.csv* format. Next, we pre-processed the crawled data. Based on the function of removing duplicates in Excel, all the repeated posts were removed from the data. The removed posts also include advertisements, Very Important Person posts, emojis, and logos to guarantee

that only the individuals' posts and the text were retained for further analysis. Excel and manual work did this part.

To make the meaning conveyance more efficient, we used *SnowNLP*, a package in Python, to extract the keywords and summary of the posts by *s.keywords()* and *s.summary()*, respectively. We also consolidated near-synonyms manually to have a more precise N-gram co-occurrence presentation. For example, we grouped the use of *dōngtiān pèidài kǒuzhào* 'wear masks in winter' to *dōngtiān dài kǒuzhào* 'wear masks in winter', given that they convey the same meaning.

3.5.4 Co-occurrence Calculated by VOSviewer

For the quantitative part of showing public attention change, this section uses VOSViewer to visualize the co-occurrence of the Chinese words in the Sina Microblog. This software runs on a Java environment. It can be used for map creation on network data and map visualization and exploration (van Eck & Waltman, 2021). In addition to the journal bibliometric data, it can also work on the self-built corpora, that is, our crawled Sina Microblog posts on *dài kǒuzhào* 'wear masks'. The function presentation for this software is shown in Figure 9.

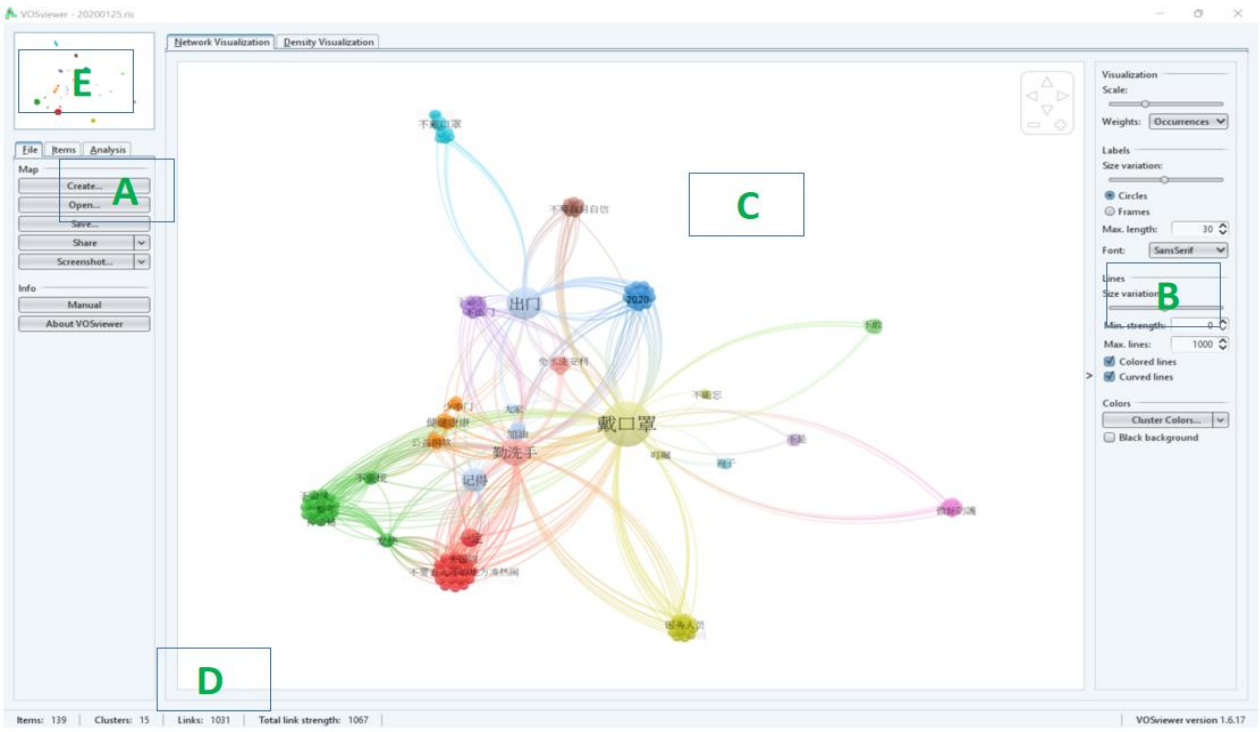

Figure 9 Main window of VOSviewer on the Sina Microblog data on January 25th, 2020. The letters designate (A) the main panel, (B) the options panel, (C) the action panel, (D) the information panel, and (E) the overview panel

Since this software can only process English texts, we gave some special settings in this software to make it available for processing Chinese texts. The procedures for visualizing the co-occurrence thus created a map based on bibliographic data because we used keyword co-occurrence for our data visualization. This type of data should be read by supporting files such as *.ris* format, which any reference manager file can generate. We used *Mendeley* to export that format for each time point. After inputting *.iris* data into the software, we should choose the type of analysis and counting method. The type of analysis is co-occurrence, unit analysis is keywords, and the counting method is full counting.

To illustrate how full counting works, we indicated its mechanism in Figure 10a and Figure 10b. Suppose there are four chunks labeled as C1, C2, C3, and C4, and three posts labeled as P1,

P2, and P3, as Figure 10a shown. P1 contains the chunks of C1, C2, and C3, P2 contains the chunks of C1 and C3, and P3 includes the chunks of C2 and C4. The full counting based on the example is indicated in Figure 10b. The link between C1 and C3 has the strength of 2 because these two chunks co-occurred in two posts, P1 and P2. The link between C1 and C2 has the strength of 1 because these two chunks co-occurred in one post, P1. The link between C2 and C3 has the strength of 1 because they co-occurred in one post, P1. The link between C2 and C4 has the strength of 1 because they co-occurred in one post, P3.

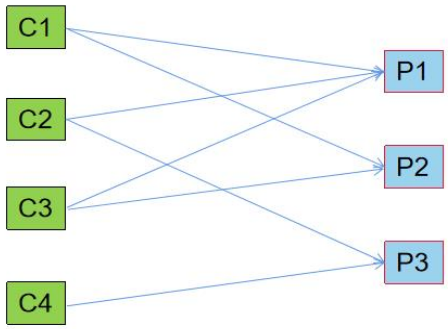

Figure 10- a Links between four chunks and three posts

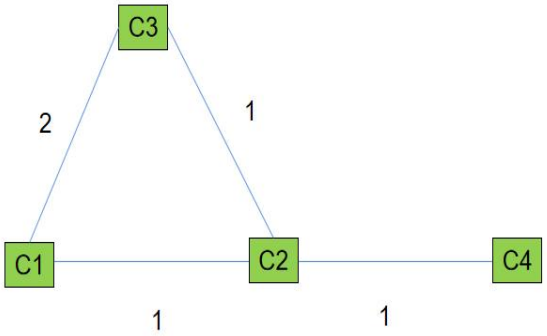

Figure 10- b Co-occurrence of chunks network constructed using full counting

Figure 10 Mechanisms of full counting

Based on the full counting, VOSviewer uses the association strength to normalize the co-occurrence frequencies (e.g., van Eck *et al.*, 2006; van Eck & Waltman, 2007). Association strength of items *i* and *j* is given by Equation [11]:

$$AS_{ij} = \frac{c_{ij}}{c_i c_j} \quad [11]$$

The association strength between *i* and *j* is the ratio between the observed number of occurrences of *i* and *j* and the number of co-occurrences of *i* and *j,* which are statistically independent.

The software VOSviewer furthermore used modularity to measure the structure of networks. Modularity is often used in optimization methods for detecting community structure in networks. Based on the association strength, VOSviewer divided the networks into different clusters shown by different colors. Networks with high modularity have dense connections between the nodes (i.e., *i* and *j*) within clusters; on the other hand, sparse connections between nodes will be clustered into different modules. Different colors represent different modules. The stronger is association strength between *i* and *j*, the thicker the link between them is. The more frequent the chunk is in the Sina Microblog Corpus, the bigger the chunk circle.

3.6 Chapter Summary

This chapter introduces how data is collected and analyzed in terms of qualitative and quantitative parts. Specifically, the motivation for focusing on the context of China and the Chinese language was reviewed because its data will be more available than that in other countries. Meanwhile, the uniqueness of Mandarin Chinese morphemes also provides further motivation. In the data collection, we elaborated on the four data sources, namely, Baidu Website and Baidu Index, Baidu News, the Official COVID-19 pandemic Website, and Sina Microblog,

concerning their functions and features and reported how they were collected according to the purpose of this thesis. In the data analyses, we divided them into three parts, i.e., qualitative and quantitative analyses. Important dates for significant social events are listed to explain the development and competition of the COVID-19 emergent neologisms over the fifteen months after the outbreak.

On the other hand, quantitative analyses consist of two parts. The regression part reviewed the algorithm and (dis)advantages under simple linear regression, multiple linear regression, and fine-tuned regression used to model the mapping relationship between the search frequencies of COVID-19 emergent neologisms and the pandemic cases, and buzzwords and the pandemic cases. The social media crawling part reported how to collect and analyze the data from Sina Microblog by Web Scraper. Working mechanisms and algorithms of VOSviewer for visualizing the association link between chunks are also reported in detail in this chapter.

CHAPTER FOUR EMERGENT NEOLOGISMS' REFLECTION ON POLICY AND THEIR PREDICTION IN PANDEMIC CASES

This chapter reports the findings from RQ 1 to RQ 4. RQs 1-2 are qualitative-oriented, examining how Chinese netizens used COVID-19 emergent neologisms over the fifteen months after the pandemic. In Section 4.1, we presented the proportion of each competing variant within the COVID-19 emergent neologisms in order to show their respective development and intra-competition during the 15 months after the outbreak of the COVID-19 pandemic. We then explained that the developmental pattern and competition might be attributed to the corresponding policy announcement and other popular social events.

RQs 3-4 are quantitative-oriented, which is going to answer how the search frequencies of COVID-19 emergent neologisms correlate and map the pandemic cases and whether COVID-19 emergent neologisms have bigger predictability than buzzwords in predicting the pandemic cases. The correlation and mapping relation are examined by Pearson correlation and simple linear, multiple linear, and fine-tuned regression models (i.e., stepwise methods, regularization, least angle regression, and polynomial expressions) with the sigmoidal function of log-linear regression as a baseline in Section 4.2. Lastly, Section 4.3 summarizes this chapter.

4.1 Evolution and Competition within COVID-19 Emergent Neologisms

As Figure 11 indicates, the evolution of each variant within the emergent neologisms has been shown over the fifteen months from December 21st, 2019 to March 30th, 2021. The most striking color is blue, which represents the under-specifications. Since the outbreak of the pandemic, under-specifications have been widely used among the public till now. However, they also

witnessed noticeable fluctuations, especially at the early time of the pandemic. For example, on December 21st and 25th, 2019, more than 90% of under-specifications were used to refer to the disease. However, the proportion of under-specifications decreased to 40% on the last day of December 2019 and stayed around 50% for the next month. From February 9th, 2020, the end of March, and early April, the proportion of under-specifications jumped from approximately 50% to around 70%. Since early April, the proportion of under-specifications grew to and remained at 75% over the remaining nine months. The general wide use of under-specifications on the internet responds to the seriousness of the pandemic since December 21st, 2019.

According to the emerging sequence of the variants of emergent neologisms, we then observed the development of pre-official names. Though it only took no more than 10% on the first day of the official announcement of the pandemic, the proportion of pre-official names grew by 15% quickly from January 15th to 30th, 2020. Interestingly, it witnessed a quick decrease in the following five months and the lowest proportion reached 0.2%. Pre-official names have maintained around 0.2% from February 2020 to March 2021. The reason might be attributed to the fact that WHO (2020) coined official names on February 11th, 2020, which may probably replace pre-official names. The other names that were used very frequently among people at the beginning of the pandemic are stigmatizing names. They were used to name the disease since the first day of the pandemic announcement and achieved the highest proportion with more than 50%, on January 5th, 2020, which exceeded under-specifications by 10%. The phenomenon stigmatizing names overtaking under-specifications reflects the public's panic to the emerging disease. Nevertheless, the stigmatizing names decreased to almost zero as the official names were created. The decrease and disappearance of stigmatizing names also reflects the

announcement by WHO (2020) who suggested not to include discrimination (i.e., disease nomenclatures involved in regions, countries, *etc.*) in the terminology to refer to any disease.

In addition to the blue color, green is the other color saturated in Figure 11. Unlike the other variants in the COVID-19 emergent neologisms, official names (e.g., *xīn guān* novel corona 'coronavirus') demonstrated a more stable developmental pattern since their birth on February 8th, 2020, retaining around 20%-30% during the whole 15 months in the range collected in our thesis. The appearance and use of Chinese official names echo the emergence of English abbreviations, but English abbreviations such as the term 'COVID-19 pandemic' were not widely used in mainland China. Hence, they only had a proportion of around 0.3% on January 25th, 2020. Even though they reached a peak of over 1% between March and April, 2020, they went back to 0.3% in the following eleven months.

Regarding the competition, under-specifications and pre-official names were the only linguistic variants used to refer to the disease at the beginning of the pandemic from December 21st to 26th, 2019. However, under-specifications dominated pre-official names with 95% of searches on the Baidu Index. However, stigmatizing names replaced the role of under-specifications on December 31st, 2019 and January 5th, 2020, when the epidemic developed to a national level. The most significant difference between stigmatizing names and under-specifications can reach more than 10%, with the former exceeding the latter. As the speed of the epidemic grew to an unprecedented and uncontrollable degree (between January 15th to 20th, 2020), these two linguistic variants with similar proportions (i.e., stigmatizing names and under-specifications 37.1% and 37.3%) occupied more than half of the use in the disease nomenclature. Stigmatizing names were substituted by official names since February 20th, 2020. In the rest of the months in our dataset (since January 25th, 2020), stigmatizing names seemed not to be

selected for the disease reference with almost disappearance in Figure 11. In contrast, official names since their substitution for stigmatizing names indicated a stable development around 20%-30% and kept the second biggest in the disease nomenclature.

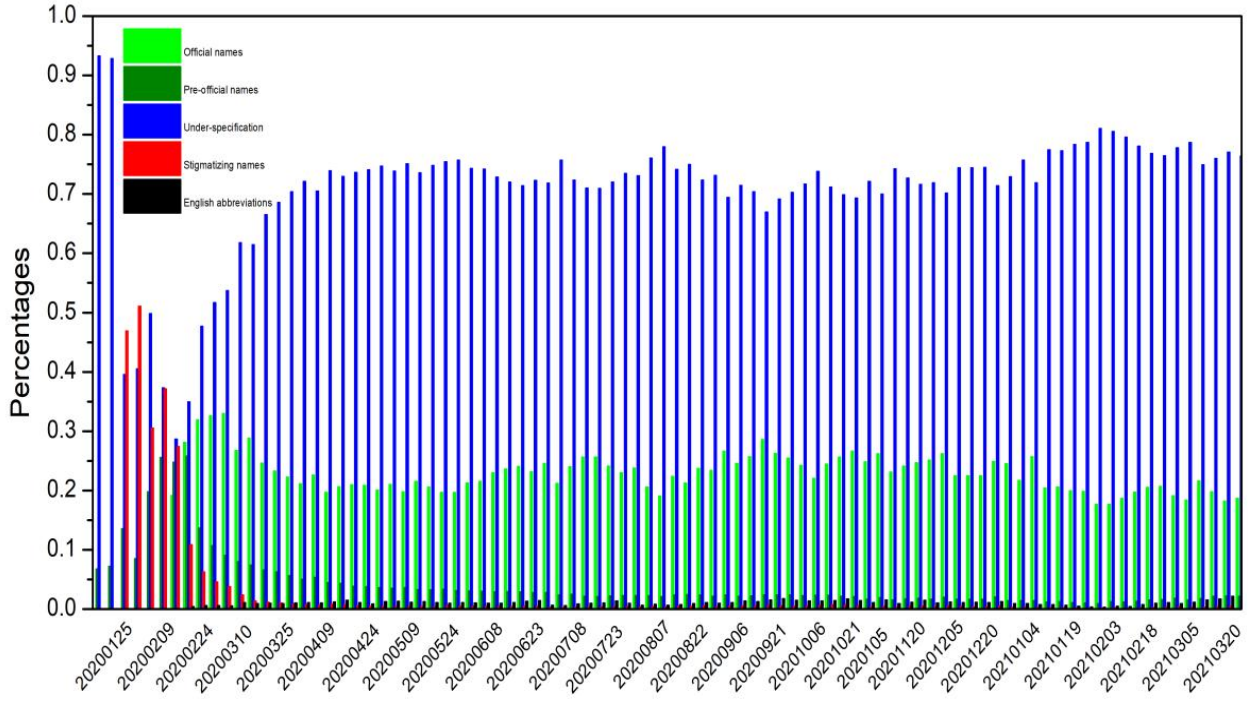

Figure 11 Percentages of variants in the emergent neologisms on the Baidu Index

4.2 Regression Modeling by COVID-19 Emergent Neologisms

This section examines the statistical correlation between emergent neologisms and pandemic cases in Section 4.2.1.

In the regression modeling, Section 4.2.2 reports the model performance by the sigmoidal function, which is the formula of the S-curve and also the baseline of the other model performance. Before using the linear and polynomial regression modeling, we should examine whether the data distribution of COVID-19 emergent neologisms meets the requirement of regression analysis. Hence, the report for this part will be presented before each following

regression report. The prediction result based on the simple linear regression model is shown in Section 4.2.3. Furthermore, we tried multiple linear regression and reported their performance in Section 4.2.4. Section 4.2.5 continued to try the polynomial expressions by single versus multiple variables. To avoid the multicollinearity issue, we employed the fine-tuned regressions with stepwise methods, regularization, and the least angle regression on polynomial expressions of all variables in Section 4.2.6. Since one of the important purposes is to compare the model performance by using COVID-19 emergent neologisms as predictors with the model performance by using buzzwords as predictors, correlation and mapping examination both include buzzwords.

4.2.1 Pearson Correlation

The Pearson correlation analysis is motivated by the initial observation of the development of COVID-19 emergent neologisms and pandemic cases during the 15 months after the pandemic outbreak. Figure 12-a indicates the number of newly confirmed cases and newly suspected cases in Greater China. At the same time, Figure 12-b demonstrates the frequency of the variants in the COVID-19 emergent neologisms, Figure 12-c the frequency of the variants in the vector words, and Figure 12-d the frequency of the variants in the PPE words. All figures cover the same duration from January 20th to March 25th, 2021 in order to compare their development with the pandemic cases according to observable patterns. In the visualization, we retained the periods with five-day intervals as the *x*-axis, but for a more straightforward observation of trends, we logarithm the *y*-axis.

COVID-19 cases showed a more complex picture than the development of COVID-19 emergent neologisms and buzzwords: more rises and falls over time. Specifically, such

complexity has been demonstrated in newly confirmed and currently suspected cases. Both of them declined quickly during the following four months after reaching the peak and experienced many "boomlets" such as the late June, August 2020 and the late January 2021, as Figure 12-a shows (around 10, 000 cases in the late January 2020). However, the overall trend is still upward at first and downward later.

Similarly, under-specifications, official names, pre-official names, and stigmatizing names as a whole showed an upward trend at the early stage of the pandemic and a downward trend after January 25th, 2020. Despite subtle fluctuations, English abbreviations also showed growth, with their highest point on February 19th and a gradual decline in the following months. Words on vectors and PPE also showcased initial growth within a short time (between December 21st, 2019 and February 9th, 2020) and gentle declines in the following thirteen months. Hence, pandemic data and COVID-19 emergent neologisms and buzzwords increased early and generally decreased later.

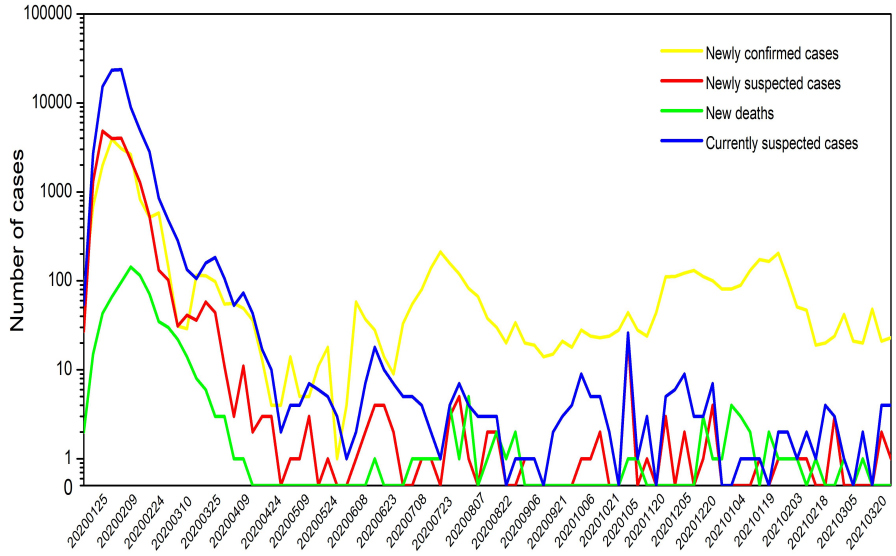

Figure 12- a Development of national newly confirmed and suspected cases

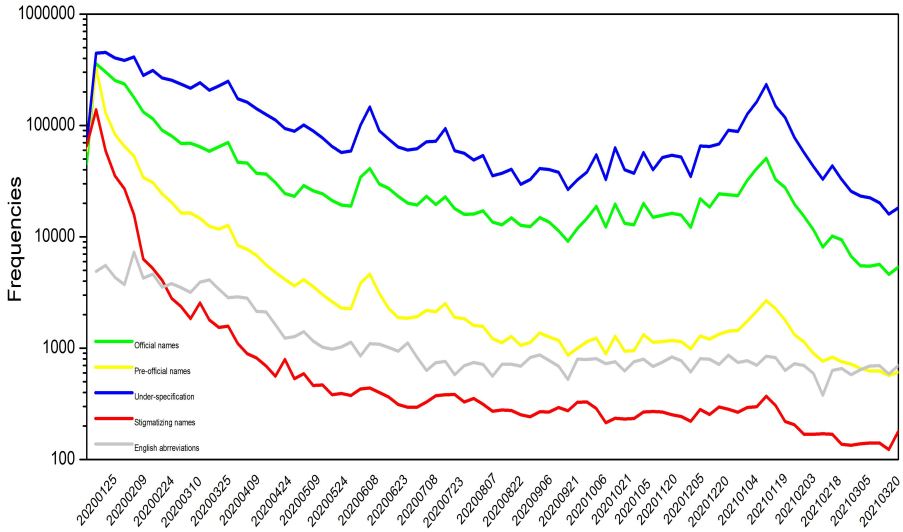

Figure 12- b Development in frequencies of COVID-19 emergent neologisms

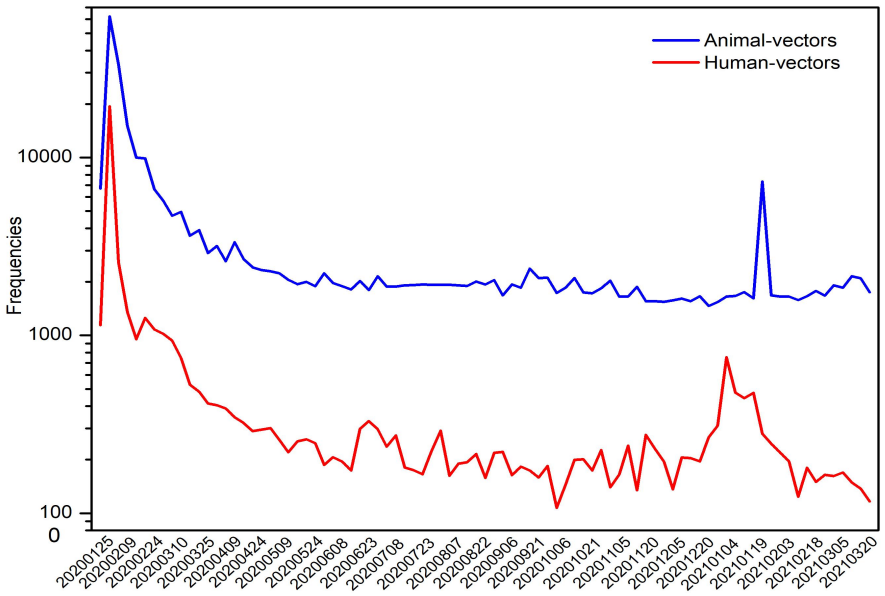

Figure 12- c Development in frequencies of vector terms

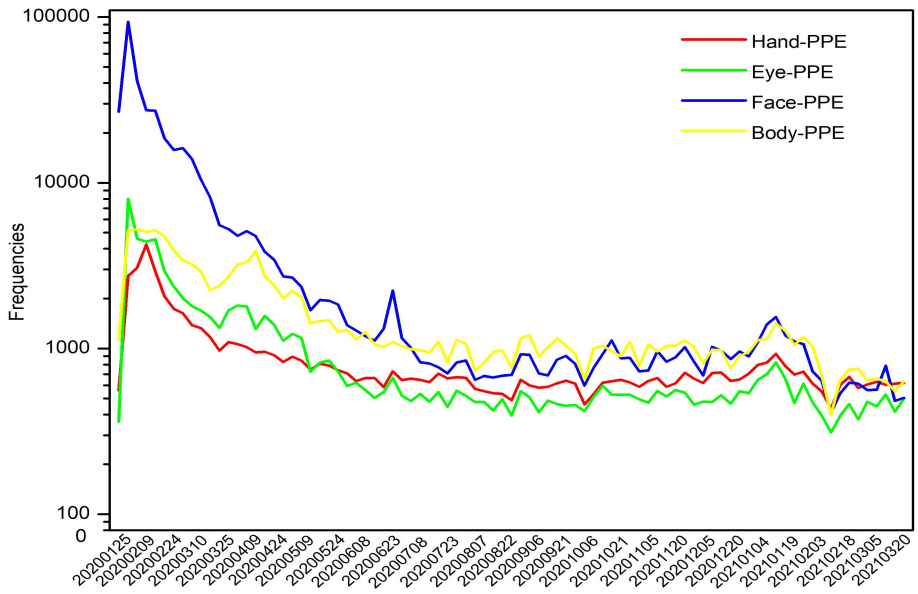

Figure 12- d Development in frequencies of PPE terms

Figure 12 Comparison of pandemic cases and emergent neologisms and buzzwords in Greater China

Then, it is natural to correlate the searching trends of the COVID-19 emergent neologisms and buzzwords with the patterns in the newly confirmed and newly suspected cases to verify our initial observations. The Correlation Matrix is shown in Figure 13. Comparatively, terms on hand PPE have the highest correlation values with epidemiological data such as their correlation with newly suspected cases ($r = 0.906$, $p < 0.01$), newly confirmed cases ($r = 0.899$, $p < 0.01$), and new deaths ($r = 0.889$, $p < 0.01$). Terms on eye and body PPE also show high correlation values with newly suspected cases ($r = 0.756$, $p < 0.01$ and $r = 0.752$, $p < 0.01$). However, the correlation values become smaller between eye PPE and the other pandemic cases (lower than 0.7). In contrast, the correlation values show a more stable relationship between body PPE and the other pandemic cases (still larger than 0.7). By contrast, terms on vectors have relatively low correlation values with pandemic data. The largest correlation value is $r = 0.579$ between animal vectors and newly suspected cases. Regarding COVID-19 emergent neologisms, official names

have the highest correlation with newly suspected cases (r = 0.848, p < 0.01). They also significantly correlate with newly confirmed cases (r = 0.780, p < 0.01) and currently suspected cases (r = 0.766, p < 0.01). However, the correlation between official names and new deaths becomes weaker (r = 0.668, p < 0.01). Similarly, the under-specification terms have a more significant correlation with the newly suspected cases (r = 0.726, p < 0.01) and confirmed cases (r = 0.714, p < 0.01). Differently, the correlation value between under-specifications and new deaths is bigger (r = 0.719, p < 0.01). In contrast, the value between under-specifications and currently suspected cases is smaller (r = 0.668, p < 0.01) than the correlation values between official names and new deaths versus official names and currently suspected cases, respectively. The English abbreviations also significantly correlate with newly suspected cases (r = 0.639, p < 0.01), newly confirmed cases, (r = 0.636, p < 0.01), and currently suspected cases (r= 0.566, p < 0.01). In the correlation between English abbreviations and pandemic cases, the most significant correlation lies in their relation to new deaths (r = 0.761, p < 0.01). Comparatively, pre-official and stigmatizing names have a weaker correlation with most r values less than 0.5.

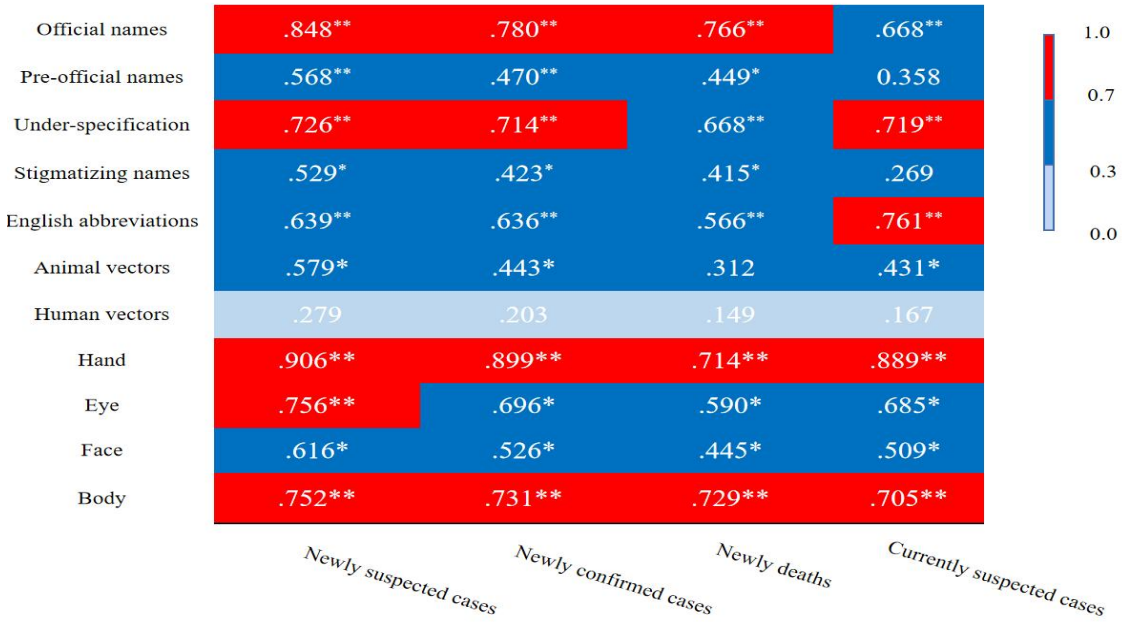

Figure 13 Correlation matrix of emergent neologisms and buzzwords and pandemic cases ($^{*}p < .05$; $^{**}p < .01$)

Overall, many variants of emergent neologisms and the variants of buzzwords are highly correlated with the pandemic cases. We then applied regression analyses to interpret the observed strong correlation values between COVID-19 emergent neologisms/buzzwords and the epidemiological data (Brumercikova & Bukova, 2020; Cao *et al.*, 2020).

Before the trials of different regressions, one noteworthy thing needs attention. Our research did not consider the interaction among predictors because it might explain modeling parameters more like a 'black box'; instead, we only focused on the main effects of independent variables (i.e., emergent neologisms and buzzwords) on the DVs. The regression analysis part is not only to find out the mathematical relationship between emergent neologisms and pandemic cases but also to explore whether emergent neologisms have predictability on pandemic development compared with buzzwords. The datasets were thus divided into three parts: 80% as training data, 10% as validation data, and 10% as test data in the following regression modeling.

4.2.2 Log-Linear Regression Performance

As has already been reviewed the wide interpretability of the S-curve (i.e., the sigmoidal function of log-linear regression) on the development of neologisms, we examined whether it can be applied to the mapping relationship between COVID-19 emergent neologisms and pandemic development, and buzzwords and pandemic development.

We only fed the log-linear regression with every single variant as an independent variable in this part because the logic is developed from a simple arrangement of the independent variables to a deep arrangement. The simple arrangement refers to considering every single variant in the COVID-19 emergent neologisms and buzzwords as independent variables for the log-linear regression model. In contrast, the deep arrangement means that the different combinations of the variants in the COVID-19 emergent neologisms and buzzwords are embedded into the log-linear regression model. The other regression models used by the present thesis will deal with the mapping relationship between COVID-19 emergent neologisms and pandemic cases and buzzwords and pandemic cases based on these two types of independent variables'arrangement. In this way, we can find a better regression model by controlling the same independent variables'arrangement and extract better predictors by controlling the same regression model.

Table 5 presents the performance by the sigmoidal function based on single variables of emergent neologisms.

Table 5 Log-linear regression performance on single variables of COVID-19 emergent neologisms

Independent Variables	Dependent Variables	R^2	RMSE
Official names	Newly confirmed cases	0.469	24.816
Pre-official names	Newly confirmed cases	0.226	35.186
Under-specifications	Newly confirmed cases	0.492	15.025
Stigmatizing names	Newly confirmed cases	0.181	25.682
English abbreviations	Newly confirmed cases	0.338	16.908
official names	Newly suspected cases	0.768	108.547
Official names	Newly suspected cases	0.418	441.433
Pre-official names	Newly suspected cases	0.785	15.339
Under-specifications	Newly suspected cases	0.333	84.849
English abbreviations	Newly suspected cases	0.782	30.814
Official names	New deaths	0.573	1.133
Pre-official names	New deaths	0.241	1.270
Under-specifications	New deaths	0.695	0.692
Stigmatizing names	New deaths	0.162	0.982
English abbreviations	New deaths	0.680	0.631
Official names	Currently suspected cases	0.718	693.541
Pre-official names	Currently suspected cases	0.360	1980.554
Under-specifications	Currently suspected cases	0.785	73.547
Stigmatizing names	Currently suspected cases	0.273	305.932
English abbreviations	Currently suspected cases	0.820	212.558

According to Table 5, there is no R^2 larger than 0.9 (an acceptable predictability result) in the sigmoidal function of log-linear regression by COVID-19 emergent neologisms, with the enormous value of 0.820 occurring in the pair of English abbreviations as an independent variable and currently suspected cases as a dependent variable. The other good predictability scores with R^2 above 0.7 lie in the relationship between official names and newly suspected cases (R^2 = 0.768), under-specifications and newly suspected cases (R^2 = 0.785), English abbreviations and newly suspected cases (R^2 = 0.782), official names and currently suspected cases (R^2 = 0.718), as well as under-specifications and currently suspected cases (R^2 = 0.785). On the other hand, the R^2 values for most pairs of COVID-19 emergent neologisms and pandemic cases are very low by the sigmoidal function of log-linear regression models.

Table 6 Log-linear regression performance on single variables of vector names

Independent Variables	Dependent Variables	R^2	RMSE
Names on animal vectors	Newly confirmed cases	0.209	30.221
	Newly suspected cases	0.368	189.482
	New deaths	0.202	1.085
	Currently suspected cases	0.309	728.657
Names on human vectors	Newly confirmed cases	0.092	22.616
	Newly suspected cases	0.172	39.806
	New deaths	0.083	0.938
	Currently suspected cases	0.136	150.016

Table 6 indicates the performance of the vector names modeling the pandemic cases by log-linear regression. Worse than the model performance on emergent neologisms, vector names can only contribute to the R^2 values below 0.4, which is considered to be weakly predictable to pandemic cases.

Table 7 Log-linear regression performance on single variables of PPE names

Independent Variables	Dependent Variables	R^2	RMSE
Hand PPE names	Newly confirmed cases	0.481	24.810
	Newly suspected cases	0.755	39.806
	New deaths	0.591	1.358
	Currently suspected cases	0.724	1742.505
Eye PPE names	Newly confirmed cases	0.373	31.241
	Newly suspected cases	0.696	392.638
	New deaths	0.483	1.383
	Currently suspected cases	0.662	3117.314
Face PPE names	Newly confirmed cases	0.281	38.591
	Newly suspected cases	0.521	662.805
	New deaths	0.330	1.500
	Currently suspected cases	0.460	3540.174
Body PPE names	Newly confirmed cases	0.404	16.104
	Newly suspected cases	0.826	13.234
	New deaths	0.654	0.676
	Currently suspected cases	0.853	56.471

Compared to vector names, the model performance based on PPE names shown in Table 7 seems to have stronger predictability with the R^2 on body PP names in predicting newly suspected cases

(R^2 = 0.826) and currently suspected cases (R^2 = 0.853), respectively. The R^2 values more than 0.7 only have two pairs (i.e., hand names predicting newly suspected and currently suspected cases).

The modeling results shown from Table 5 to Table 7 can be summarized that COVID-19 emergent neologisms and PPE names might have better predictability than vector names. However, both did not show R^2 values higher than 0.9 in the sigmoidal function of log-linear regression.

4.2.3 Simple Linear Regression on COVID-19 Emergent Neologisms

Since the sigmoidal function of log-linear regression has been found ineffective to develop the importance of the COVID-19 emergent neologisms and buzzwords in the model interpretability (i.e., no R^2 larger than 0.9), we then examined the model performance by simple linear regressions. Again, to compare the performance of simple linear regressions with log-linear regression, the linear regression model considers every single variant as an independent variable in this part. Table 8 indicates the model performance of simple linear regression by comparison with that of log-linear regression based on COVID-19 emergent neologisms.

Table 8 Performance between simple linear and log-linear regression on single variables of COVID-19 emergent neologisms

IV	DV	Linear-R^2	Linear-RMSE	Log-linear-R^2	Log-linear-RMSE
Official names	Newly confirmed cases	0.608	394.128	0.469	24.816
Official names	Newly suspected cases	0.719	444.490	0.226	35.186
Official names	New deaths	0.446	18.339	0.492	15.025
Official names	Currently suspected cases	0.587	2578.842	0.181	25.682
Pre-official names	Newly confirmed cases	0.221	555.965	0.338	16.908
Pre-official names	Newly suspected cases	0.322	690.473	0.768	108.547
Pre-official names	New deaths	0.128	23.013	0.418	441.433
Pre-official names	Currently suspected cases	0.201	3586.386	0.785	15.339
Under-specifications	Newly confirmed cases	0.509	441.143	0.333	84.849
Under-specifications	Newly suspected cases	0.527	576.639	0.782	30.814
Under-specifications	New deaths	0.517	17.124	0.573	1.1329
Under-specifications	Currently suspected cases	0.446	2988.088	0.241	1.270
Stigmatizing names	Newly confirmed cases	0.179	570.575	0.695	0.692
Stigmatizing names	Newly suspected cases	0.280	711.530	0.162	0.982
Stigmatizing names	New deaths	0.073	23.737	0.68	0.631
Stigmatizing names	Currently suspected cases	0.172	3651.395	0.718	693.541
English abbreviations	Newly confirmed cases	0.405	485.855	0.36	1980.554
English abbreviations	Newly suspected cases	0.408	645.078	0.785	73.547
English abbreviations	New deaths	0.579	15.994	0.273	305.932
English abbreviations	Currently suspected cases	0.320	3308.647	0.820	212.558

According to the comparison result shown in Table 8, we might not decide which model performs better since there is still no modeling result with R^2 larger than 0.9. The biggest R^2 shown in Table 8 is just above 0.7, occurring between official names and newly suspected cases.

Table 9 Performance between simple linear and log-linear regression on single variables of vector names

IV	DV	Linear-R^2	Linear-RMSE	Log-linear-R^2	Log-linear-RMSE
Animal vector names	Newly confirmed cases	0.197	564.514	0.209	30.221
Animal vector names	Newly suspected cases	0.335	683.919	0.368	189.482
Animal vector names	New deaths	0.097	23.416	0.202	1.085
Animal vector names	Currently suspected cases	0.186	3621.849	0.309	728.657
Human vector names	Newly confirmed cases	0.041	616.626	0.092	22.616
Human vector names	Newly suspected cases	0.078	805.396	0.172	39.806
Human vector names	New deaths	0.022	24.373	0.083	0.938
Human vector names	Currently suspected cases	0.028	3956.951	0.136	150.016

According to Table 9, the simple linear regression does not perform better than the performance by log-linear regression by using the vector names as independent variables. A similar phenomenon that neither simple linear nor log-linear regression perform effectively goes to PPE names as independent variables with no R^2 larger than 0.9 shown in Table 10.

Table 10 Performance between simple linear and log-linear regression on single variables of PPE names

IV	DV	Linear-R^2	Linear-RMSE	Log-linear-R^2	Log-linear-RMSE
Hand PPE names	Newly confirmed cases	0.808	275.867	0.481	24.810
Hand PPE names	Newly suspected cases	0.821	354.511	0.755	39.806
Hand PPE names	New deaths	0.510	17.255	0.591	1.358
Hand PPE names	Currently suspected cases	0.791	1834.067	0.724	1742.505
Eye PPE names	Newly confirmed cases	0.484	452.424	0.373	31.241
Eye PPE names	Newly suspected cases	0.571	548.935	0.696	392.638
Eye PPE names	New deaths	0.348	19.904	0.483	1.383
Eye PPE names	Currently suspected cases	0.469	2923.546	0.662	3117.314
Face PPE names	Newly confirmed cases	0.277	535.449	0.281	38.591
Face PPE names	Newly suspected cases	0.380	660.517	0.521	662.805
Face PPE names	New deaths	0.198	22.076	0.330	1.500
Face PPE names	Currently suspected cases	0.259	3455.278	0.460	3540.174
Body PPE names	Newly confirmed cases	0.535	429.486	0.404	16.104
Body PPE names	Newly suspected	0.565	553.055	0.826	13.234
Body PPE names	New deaths	0.532	16.863	0.654	0.676
Body PPE names	Currently suspected cases	0.497	2847.049	0.853	56.471

Two issues draw our attention from the low performance ($R^2 < 0.9$) of the sigmoidal function of log-linear regression and simple linear regression models. On the one hand, both models might be too simple to map the complex relationship between COVID-19 emergent neologisms and pandemic cases, leading to ineffective performance in the model prediction. A more complex model should be tried to improve the modeling result. On the other hand, a single variable (i.e., linear expressions) might also be too simple to fit the development of a pandemic. Even if the variants are involved in the COVID-19 emergent neologisms and buzzwords, it needs to realize that different variants play distinctive functions. Hence, each variant should be endowed with

varying weights in the model fitting. A multiple linear regression model should be, therefore, tried to verify whether a more complex regression model by endowing different variants with different weights in the formula can perform better in the prediction. In addition, we also found that emergent neologisms and PPE names might have better predictability in the model performance in predicting the pandemic development by the above trials on log-linear regression, which will be, therefore, further focused on in the following modeling.

4.2.4 Multiple Linear Regression on Emergent Neologisms

Before showing the modeling result based on multiple linear regression, we examined whether the linear expressions of the COVID-19 emergent neologisms meet the requirement of conducting a regression analysis. The premise check selects newly confirmed cases as the dependent variable because they are the most focused pandemic cases by the public compared to the other three types of pandemic cases.

Figure 14 indicates whether the linear relationship between all the variants in the COVID-19 emergent neologisms and newly confirmed cases.

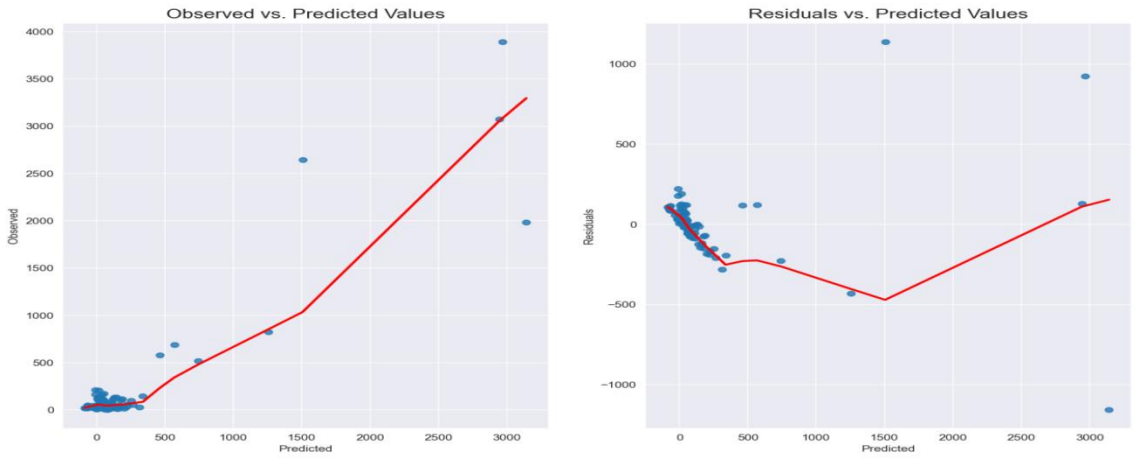

Figure 14 Linearity check of all the variants in emergent neologisms and pandemic cases

All points shown in observed data versus predicted values plot of Figure 14 roughly evenly distributed around the diagonal. All points are roughly evenly distributed on both sides of the horizontal line in the residual-predicted value plot, so the linearity is met. It satisfies the first hypothesis.

Second, the average of residuals is almost 0 (Residual_Mean =-8.633e-13), which meets the second hypothesis that no autocorrelation exists in the error terms in all the variants of emergent neologisms. Third, Variance Inflation Factor (VIF) shows the multicollinearity issue in the linear regression of all the variant in the emergent neologisms because each variant has VIF higher than 10 (the threshold of multicollinearity issue). Table 11 shows the VIF result.

Table 11 VIF result of the linear regression of all the variant in the emergent neologisms

	Features	VIF
0	Official names	26.074
1	Pre-official names	26.425
2	Under-specifications	28.522
3	Stigmatizing names	18.379
4	English abbreviations	14.989

Fourth, the examination of homoscedasticity of errors is checked by Breusch-Pagan test. According to Table 12, the homoscedasticity of errors (p-value < 0.05) is rejected.

Table 12 Breusch-Pagan test

	Value
Lagrange multiplier statistic	5.493e+01
p-value	1.352e-10
f-value	2.7741e+01
f p-value	2.923e-16

In sum, linear expression of considering all the variants in the emergent neologisms to be independent variables can be put into linear regression model but the collinearity issue should be cautious. The modeling result based on the linear expression of all the variants in the emergent neologisms could be used as a baseline for whether optimization performs better.

Table 13 demonstrates the multiple linear regression model performance by putting all the variants in each type of emergent neologism into the model. Again, to compare the model performance among log-linear regression, simple linear regression, and multiple linear regression, all the independent variables put into the model are linear expressions. Table 13-a demonstrates the performance of three regression models based on COVID-19 emergent neologisms, Table 13-b on vector names, and Table 13-c on PPE names.

Table 13 Model performance by multiple linear regression performance

Table 13- a Performance of all the variants of COVID-19 emergent neologisms

IV	DV	R^2	RMSE
All emergent neologisms	Newly confirmed cases	0.866	236.491
All emergent neologisms	Newly suspected cases	0.979	124.319
All emergent neologisms	New deaths cases	0.707	13.658
All emergent neologisms	Currently suspected cases	0.942	991.692

Table 13- b Performance of all the variants of vector names

IV	DV	R^2	RMSE
All vector names	Newly confirmed cases	0.470	461.148
All vector names	Newly suspected cases	0.756	416.825
All vector names	New deaths	0.223	21.856
All vector names	Currently suspected cases	0.530	2768.181

Table 13- c Performance of all the variants of PPE names

IV	DV	R^2	RMSE
All PPE names	Newly confirmed cases	0.872	229.493
All PPE names	Newly suspected cases	0.841	340.454
All PPE names	New deaths	0.615	15.575
All PPE names	Currently suspected cases	0.866	1494.447

According to Table 13-a, 13-b, and 13-c, it is noticeable that considering all variants of each type of emergent neologisms outperforms the only consideration of a single variant in the model. On the one hand, the R^2 value can reach 0.719 by using official names to model newly suspected cases, while the R^2 value can grow to 0.979 by using all the COVID-19 emergent neologisms to model newly suspected cases. The same is true for vector names and PPE names. When using the

single animal name to model newly suspected cases, R^2 is only 0.335, but the R^2 value skyrockets to 0.756 based on all vector names. The consideration of all the PPE variants also improve the model performance from 0.821 between hand PPE and newly suspected cases to 0.872 after considering all variants of PPE names to predict the newly confirmed cases.

Interestingly, when we fed the log-linear regression model with all variants of COVID-19 emergent neologisms and buzzwords, the results did not show a noticeable improvement, as indicated in Table 14.

Table 14 Log-linear regression model performance on all variants versus single variants

Table 14- a Performance on emergent neologisms

IV_1	DV_1	Log-linear-all-R^2	Log-linear-all- RMSE	IV_2	DV_2	Log-linear-single-R^2	Log-linear-single-RMSE
All emergent neologisms	Newly confirmed cases	0.349	508.164	Official names	Newly confirmed cases	0.469	24.816
				Pre-official names	Newly confirmed cases	0.226	35.186
	Newly suspected cases	0.360	670.803	Under-specifications	Newly confirmed cases	0.492	15.025
	New deaths	0.404	19.033	Stigmatizing names	Newly confirmed cases	0.181	25.682
	Currently suspected cases	0.309	3336.283	English abbreviations	Newly confirmed cases	0.338	16.908
				Official names	Newly suspected cases	0.768	108.547
				Pre-official names	Newly suspected cases	0.418	441.433
				Under-specifications	Newly suspected cases	0.785	15.339
				Stigmatizing names	Newly suspected cases	0.333	84.849
				English abbreviations	Newly suspected cases	0.782	30.814
				Official names	New deaths	0.573	1.133
				Pre-official names	New deaths	0.241	1.270
				Under-specifications	New deaths	0.695	0.692
				Stigmatizing names	New deaths	0.162	0.982

Table 14-a Continued

IV_1	DV_1	Log-linear-all-R^2	Log-linear-all- RMSE	IV_2	DV_2	Log-linear-single-R^2	Log-linear-single-RMSE
All emergent neologisms	Newly confirmed cases	0.349	508.164	English abbreviations	New deaths	0.680	0.631
				Official names	Currently suspected cases	0.718	693.541
	Newly suspected cases	0.360	670.803	Pre-official names	Currently suspected cases	0.360	1980.554
	New deaths	0.404	19.033	Under-specifications	Currently suspected cases	0.785	73.547
	Currently suspected cases	0.309	3336.283	Stigmatizing names	Currently suspected cases	0.273	305.932

Table 14- b Performance on vector names

IV_1	DV_1	Log-linear-all-R^2	Log-linear-all- RMSE	IV_2	DV_2	Log-linear-single-R^2	Log-linear-single-RMSE
All vector names	Newly confirmed cases	0.419	480.155	Animal	Newly confirmed cases	0.209	30.221
	Newly suspected cases	0.468	612.040		New suspected cases	0.368	189.482
	New deaths	0.432	18.577		New deaths	0.202	1.085
	Currently suspected cases	0.366	3194.437		Currently suspected cases	0.309	728.657
				Human	Newly confirmed cases	0.092	22.616
					New suspected cases	0.172	39.806
					New deaths	0.083	0.938
					Currently suspected cases	0.136	150.016

Table 14- c Performance on PPE names

IV_1	DV_1	Log-linear-all-R^2	Log-linear-all-RMSE	IV_2	DV_2	Log-linear-single-R^2	Log-linear-single-RMSE
All PPE names	Newly confirmed cases	0.391	491.414	Hand	Newly confirmed cases	0.481	24.810
	Newly suspected cases	0.410	644.340		Newly suspected cases	0.755	39.806
	New deaths	0.442	18.417		New deaths	0.591	1.358
	Currently suspected cases	0.358	3216.534		Currently suspected cases	0.724	1742.505
				Eye	Newly confirmed cases	0.373	31.241
					Newly suspected cases	0.696	392.637
					New deaths	0.483	1.383
					Currently suspected cases	0.662	3117.314
				Face	Newly confirmed cases	0.281	38.591
					Newly suspected cases	0.521	662.805
					New deaths	0.330	1.500
					Currently suspected cases	0.460	3540.174
				Body	Newly confirmed cases	0.404	16.104
					Newly suspected cases	0.826	13.234
					New deaths	0.654	0.676
					Currently suspected cases	0.853	56.471

According to Table 14-a, Table 14-b, and Table 14-c, the performance of log-linear regression on three types of emergent neologisms with the R^2 around 0.4 seemed no improved compared with that on single variant as the difference between multiple linear and simple linear regression models demonstrates. In Chapter Six, there is one part for speculating why log-linear regression (i.e., an S-curve) does not perform well on all variants of COVID-19 emergent neologisms.

To some extent, we responded to one of the above issues, i.e., the arrangement of independent variables. Since log-linear regressions did not perform well, regardless of single and

multiple variants, we only compared the performance of simple linear and multiple linear regressions on single and all variants of emergent neologisms and buzzwords. We found that the models based on all variants generally outperform those on the single variant, which confirms our hypothesis of the importance of different variants in the emergent neologisms. However, it is not the end to this issue. We only tried multiple linear regression, but the pandemic trends are nonlinear, as shown in Figure 11. Hence, our next trials will go deeper by trying polynomial expressions to describe the mathematical relationship between emergent neologisms and pandemic development.

4.2.5 Polynomial regression

Through observing the trend of pandemic cases and the development of emergent neologisms from Figure 11, we found that binomial and trinomial expressions can be tried because they showed a parabola curve, especially at the early stage of the pandemic. Higher polynomial expressions such as the quartic polynomial expression are not necessary to be tried because trinomial expressions have already satisfied the description of the pandemic cases and emergent neologisms. Following the arrangement of independent variables, i.e., single variant versus all the variants, when running the single variant with the binomial expression as independent variables, we compared its performance to the single variant with linear regression. When running all the variants with the binomial expression as independent variables, we compared its performance to the single variant with linear regression. The same method goes to compare the arrangement of trinomial expressions with linear and binomial expressions on single and all variants.

Before conducting binomial regression modeling, we examined whether the binomial expression of the COVID-19 emergent neologisms satisfied the requirement of conducting a regression analysis. Figure 15 indicates whether the binomial regression of all the variants in the COVID-19 emergent neologisms satisfies the linear relationship with newly confirmed cases.

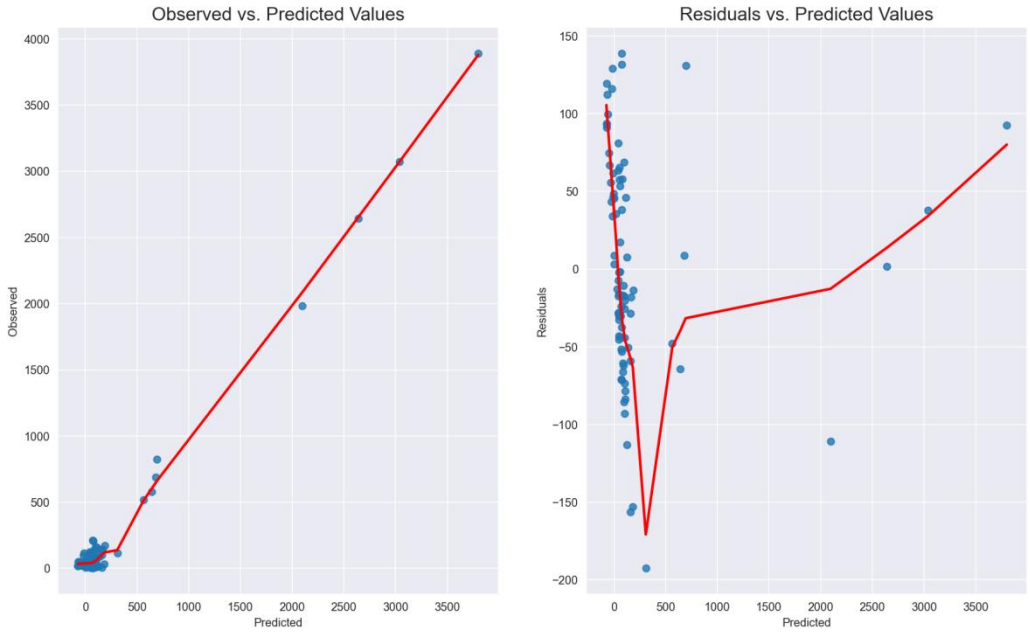

Figure 15 Linearity check of all variants in emergent neologisms with binomial expressions and pandemic cases

All points shown in observed data versus predicted values plot of Figure 15 roughly evenly distributed around the diagonal. All points are roughly evenly distributed on both sides of the horizontal line in the residual-predicted value plot, so the linearity is roughly met. It satisfies the first hypothesis.

Second, the average of residuals is almost 0 (Residual_Mean =-8.575e-08), which meets the second hypothesis. Third, VIF shows the multicollinearity issue in the binomial expression of all the variant in the emergent neologisms because most variant has VIF higher than 100, which shows the certainty of multicollinearity issue. Table 15 shows the VIF result.

Table 15 VIF result of the binomial expression of all the variant in the emergent neologisms

	Features	VIF
0	Official_names	549.471
1	Preofficial_names	1289.465
2	Underspecifications	275.489
3	Stigmitizing_names	2881.038
4	English_abbreviations	89.730
5	Official_names^2	413.900
6	Preofficial_names^2	5234.338
7	Underspecifications^2	229.759
8	Stigmitizing_names^2	10959.477
9	English_abbreviations^2	47.898

Fourth, the examination of homoscedasticity of errors is checked by Breusch-Pagan. According to Table 16, homoscedasticity of errors is rejected.

Table 16 Breusch-Pagan test

	Value
Lagrange multiplier statistic	22.019
p-value	0.015
f-value	2.575
f p-value	0.010

In sum, the binomial expression of all the variants in the emergent neologisms can be put into binomiela regression model but still the collinearity issue should be cautious. The modeling result based on the binomial expression of all the variants in the emergent neologisms could be used as a baseline for whether optimization performs better.

Table 17 demonstrates the performance of the binomial and simple linear regression models on single variables of emergent neologisms.

Table 17 Binomial expressions and simple linear regression on single variants of emergent neologisms

Binomial IVs	DVs	Single binomial-R^2	Single binomial-RMSE	Single IVs	Single linear-R^2	Single linear-RMSE
Official names, Official names^2	Newly confirmed	0.643	378.778	Official names	0.608	394.128
Official names, Official names^2	Newly suspected	0.720	446.248	Official names	0.719	444.490
Official names, Official names^2	New deaths	0.669	14.269	Official names	0.446	18.338
Official names, Official names^2	Currently suspected	0.604	2541.493	Official names	0.587	2578.842
Pre-official names, Pre-official names^2	Newly confirmed	0.680	358.158	Pre-official names	0.221	555.965
Pre-official names, Pre-official names^2	Newly suspected	0.784	391.818	Pre-official names	0.322	690.473
Pre-official names, Pre-official names^2	New deaths	0.540	16.813	Pre-official names	0.128	23.013
Pre-official names, Pre-official names^2	Currently suspected	0.683	2272.951	Pre-official names	0.201	3586.386
Underspecifications, Underspecifications^2	Newly confirmed	0.673	362.434	Underspecifications	0.509	441.1425
Underspecifications, Underspecifications^2	Newly suspected	0.775	400.128	Underspecifications	0.527	576.639
Underspecifications, Underspecifications^2	New deaths	0.554	16.556	Underspecifications	0.517	17.124
Underspecifications, Underspecifications^2	Currently suspected	0.614	2508.353	Underspecifications	0.446	2988.088

Table 17 Continued

Binomial IVs	DVs	Single binomial-R^2	Single binomial-RMSE	Single IVs	Single linear-R^2	Single linear-RMSE
Stigmatizing names, Stigmatizing names^2	Newly confirmed	0.537	431.055	Stigmatizing names	0.179	570.575
Stigmatizing names, Stigmatizing names^2	Newly suspected	0.643	503.889	Stigmatizing names	0.280	711.530
Stigmatizing names, Stigmatizing names^2	New deaths	0.296	20.765	Stigmatizing names	0.073	23.737
Stigmatizing names, Stigmatizing names^2	Currently suspected	0.574	2636.119	Stigmatizing names	0.172	3651.395
English abbreviations, English abbreviations^2	Newly confirmed	0.470	461.222	English abbreviations	0.405	485.855
English abbreviations, English abbreviations^2	Newly suspected	0.449	626.048	English abbreviations	0.408	645.078
English abbreviations, English abbreviations^2	New deaths	0.675	14.144	English abbreviations	0.579	15.994
English abbreviations, English abbreviations^2	Currently suspected	0.331	3302.132	English abbreviations	0.320	3308.647

As Table 17 shows, the result of binomial and linear expressions complies with our expectation that binomial expressions help improve the model noticeably because every R^2 in binomial expressions is more significant than that in simple linear regressions, and almost every RMSE in binomial expressions is smaller than that in linear regression. The same result also goes for binomial expressions on vector and PPE names (See Table 18 and Table 19). The modeling result by using, for example, animal vector-related names to predict new deaths was only 0.097 as the R^2 value but jumped to 0.482 when binomially. Though the R^2 differences between single

binomial and single linear regression models are not as evident as vector names, the peak value of R^2 increased from 0.821 (i.e., between hand PPE and newly suspected cases) in the single linear independent variables to 0.836 (i.e., between body PPE and newly suspected cases) in the single binomial independent variables. All the RMSE values for single vector names and PPE names in the binomial expressions are lower than simple linear expressions.

Table 18 Binomial expressions and linear regression on single variants of vector names

Binomial-IVs	DVs	Binomial-R^2	Binomial-RMSE	Single linear-IV	Single linear-R^2	Single linear-RMSE
Animal, Animal^2	Newly confirmed	0.669	364.347	Animal	0.197	564.514
Animal, Animal^2	Newly suspected	0.822	355.759	Animal	0.335	683.919
Animal, Animal^2	New deaths	0.482	17.844	Animal	0.097	23.416
Animal, Animal^2	Currently suspected	0.675	2300.077	Animal	0.186	3621.849
Human, Human^2	Newly confirmed	0.519	439.471	Human	0.041	616.626
Human, Human^2	Newly suspected	0.654	496.324	Human	0.078	805.396
Human, Human^2	New deaths	0.499	17.555	Human	0.022	24.373
Human, Human^2	Currently suspected	0.493	2873.816	Human	0.028	3956.951

Table 19 Binomial expressions and linear regression on single variants of PPE names

Binomial-IVs	DVs	Binomial-R^2	Binomial-RMSE	Single linear-IVs	Single linear-R^2	Single linear-RMSE
Hand, Hand^2	Newly confirmed	0.835	257.148	Hand	0.808	275.867
Hand, Hand^2	Newly suspected	0.832	346.043	Hand	0.821	354.511
Hand, Hand^2	New deaths	0.609	15.497	Hand	0.510	17.255
Hand, Hand^2	Currently suspected	0.834	1645.641	Hand	0.791	1834.067
Eye, Eye^2	Newly confirmed	0.601	400.146	Eye	0.484	452.424
Eye, Eye^2	Newly suspected	0.649	499.619	Eye	0.572	548.935
Eye, Eye^2	New deaths	0.588	15.905	Eye	0.348	19.904
Eye, Eye^2	Currently suspected	0.581	2612.866	Eye	0.469	2923.546

Table 19 Continued

Binomial-IVs	DVs	Binomial-R^2	Binomial-RMSE	Single linear-IVs	Single linear-R^2	Single linear-RMSE
Face, $Face^2$	Newly confirmed	0.609	396.010	Face	0.277	535.449
Face, $Face^2$	Newly suspected	0.678	478.832	Face	0.380	660.517
Face, $Face^2$	New deaths	0.586	15.958	Face	0.198	22.076
Face, $Face^2$	Currently suspected	0.605	2538.473	Face	0.259	3455.278
Body, $Body^2$	Newly confirmed	0.746	319.377	Body	0.535	429.486
Body, $Body^2$	Newly suspected	0.836	341.972	Body	0.565	553.055
Body, $Body^2$	New deaths	0.563	16.388	Body	0.532	16.863
Body, $Body^2$	Currently suspected	0.730	2099.319	Body	0.497	2847.049

Even if the binomial expressions outperformed the linear expressions, there is no R^2 larger than 0.9 in the binomial expressions. However, there are more pairs larger than 0.9 in terms of R^2 in the trinomial expressions: official names with newly confirmed cases (R^2 = 0.936), newly suspected cases (R^2 = 0.964), and currently suspected cases (R^2 = 0.939), stigmatizing names with newly confirmed cases (R^2 = 0.909), hand PPE with newly suspected cases (R^2 = 0.908), as well as eye PPE with newly suspected cases (R^2 = 0.982). More details are given in Appendix B. Appendix B-1 shows the model performance on emergent neologisms on trinomial regressions. Appendix B-2 shows the model performance on vector names on trinomial regressions. Appendix B-3 shows the model performances on PPE names on trinomial regressions.

According to Section 4.2.4, all variants outperformed single variants since each variant has its distinctive role in the model interpretability, so we examined whether the binomial regression model on all variants would continue to perform better than single all variants.

Table 20 Binomial and multiple linear regression on all variants of emergent neologisms

IVs	DVs	Binomial all-R^2	Binomial all-RMSE	Multiple linear all-R^2	Multiple linear all-RMSE
All emergent neologisms	Newly confirmed	0.987	74.622	0.866	236.491
All emergent neologisms	Newly suspected	0.996	58.573	0.979	124.319
All emergent neologisms	New deaths	0.944	6.152	0.707	13.658
All emergent neologisms	Currently suspected	0.985	520.371	0.942	991.692

Table 20 indicates better performance in the binomial expressions on all variants in the emergent neologisms than multiple linear regression. The most significant difference between these two models lies between all emergent neologisms and new deaths (Binomial all-R^2 = 0.944, Binomial all-RMSE = 6.152 versus Multiple linear all-R^2 = 0.707, Multiple linear all-RMSE = 13.658). Table 21 shows the model performance of binomial expressions on all vector names by comparison with multiple linear expressions on all vector names. The R^2 values between all vector names and new deaths improved from 0.223 in the multiple linear regression to 0.705 in the binomial regression. In contrast, the RMSE values decreased from 21.856 in the multiple linear regression to 13.638 in the binomial expressions. Table 22 shows the improvement based on the binomial expressions on all PPE names. The modeling result between all PPE names and currently suspected cases by binomial expressions enhances the R^2 value to 0.993 from the original value of 0.866 by the multiple linear regressions. The RMSE value decreases noticeably from 1494.447 in the multiple linear regressions to 362.944 in the binomial expressions, verifying the importance of arranging independent variables into the model.

Table 21 Binomial and multiple linear regression on all variants of vector names

IVs	DVs	Binomial all-R^2	Binomial all-RMSE	Multiple all-R^2	Multiple all-RMSE
All vector names	Newly confirmed	0.740	327.284	0.470	461.148
All vector names	Newly suspected	0.826	356.583	0.756	416.825
All vector names	New deaths	0.705	13.638	0.223	21.856
All vector names	Currently suspected	0.709	2203.119	0.530	2768.181

Table 22 Binomial and multiple linear regression on all variants of PPE names

IVs	DVs	Binomial all-R^2	Binomial all-RMSE	Single all-R^2	Single all-RMSE
All PPE names	Newly confirmed	0.941	159.224	0.872	229.493
All PPE names	Newly suspected	0.957	182.42	0.841	340.454
All PPE names	New deaths	0.786	11.903	0.615	15.575
All PPE names	Currently suspected	0.993	362.944	0.866	1494.447

Table 23 shows the modeling performance by using all the word subcategories to explain the pandemic case by trinomial expressions and lists the performance of binomial expressions based on all variants in the COVID-19 emergent neologisms and buzzwords since binomial expressions have been shown to improve the model performance significantly.

Table 23 Model performance of trinomial and binomial expressions on all variants of emergent neologisms

IVs	DVs	Trinomial all-R^2	Trinomial all-RMSE	Binomial all-R^2	Binomial all-RMSE
All emergent neologisms	Newly confirmed	0.996	45.021	0.987	74.621
All emergent neologisms	Newly suspected	0.999	17.247	0.996	58.573
All emergent neologisms	New deaths	0.995	1.844	0.944	6.152
All emergent neologisms	Currently suspected	0.999	80.280	0.986	520.371
All vector names	Newly confirmed	0.911	193.676	0.740	327.284
All vector names	Newly suspected	0.901	272.066	0.826	356.583
All vector names	New deaths	0.738	13.013	0.705	13.638
All vector names	Currently suspected	0.860	1545.646	0.709	2203.119
All PPE names	Newly confirmed	0.983	88.587	0.941	159.224
All PPE names	Newly suspected	0.997	42.827	0.957	182.420
All PPE names	New deaths	0.966	4.859	0.786	11.903
All PPE names	Currently suspected	0.996	265.028	0.993	362.944

Compared with binomial expressions, the trinomial expression seems to perform better regardless of R^2 and RMSE, as shown in Table 23. However, we could not determine the trinomial expressions as our final model. After all, every model's performance will be improved when adding one/more predictor(s). Since the trinomial expressions have the most significant number of IVs (including linear, binomial, and trinomial) in the model prediction, their "better" performance can be expected. Nevertheless, such a "better" performance seems more superficial partly because the two hypothesis examination has showed the tendency of having the multicollinearity issue when considering all the variant of the emergent neologisms and each variant has a strong correlation ($r > 0.7$) with the other categories in emergent neologisms and PPE names, as shown in Figure 16a and b.

We also noticed that vector names did not have R^2 larger than 0.9, even in trinomial expressions. The reason might be a multicollinearity issue shown in very high correlation values between human vector names and animal vector names ($r = 0.920$). Such multicollinearity is also offered in different variants in the emergent neologisms (Figure 16a) and other variants in the PPE names (Figure 16b), with those correlation values larger than 0.7. The reason that causes the multicollinearity issue may be the arrangement of binomial or trinomial expressions. The consequence of only considering R^2 and RMSE would be a false-positive performance (Colquhoun, 2018) where a superficially better but not good modeling result occurs.

Noticing the potential multicollinearity problem, we tried to reduce such problem in the following predictive models by hyperparameter adjustment via stepwise methods, regularization, and the least angle regression models. Considering the overall shape of a parabola and constantly small rises and falls in pandemic cases, binomial expressions would be at first explored based on these fine-tuned methods. The potentially 'falsely positive' modeling performance in the

binomial expressions without fine-tuned methods as the first four lines shown in Table 23 was used as the baselines for checking the model result.

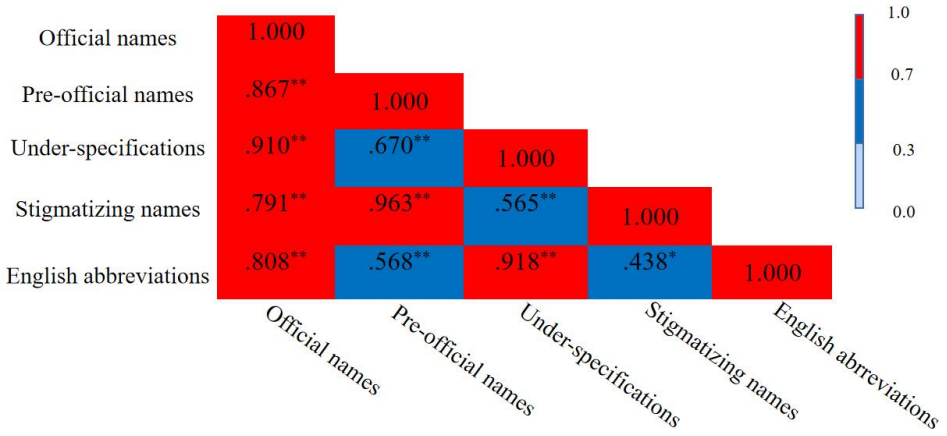

Figure 16- a Correlation matrix within emergent neologisms

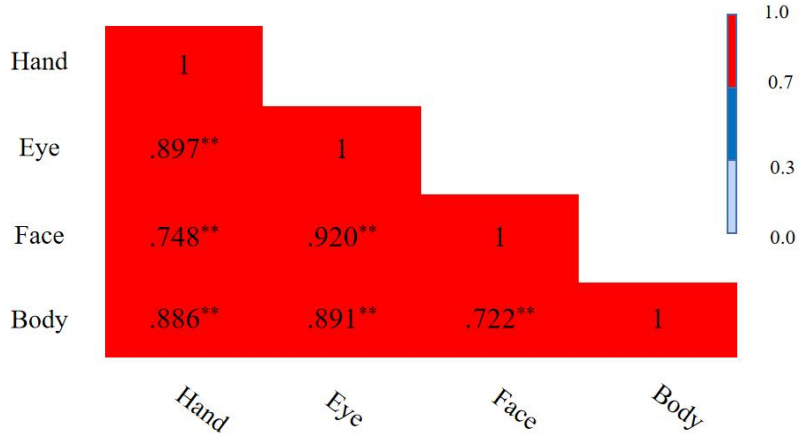

Figure 16- b Correlation matrix within PPE names

Figure 16 Correlation matrix within emergent neologisms and PPE names

4.2.6 Fine-Tuned Model Optimization

To address the multicollinearity issue and maximize the model performance, we used the following eight models: 1) Forward Selection, 2) Backward Elimination, 3) Stepwise Regression,

4) Least Angle Regression, 5) LASSO Regularization, 6) Adaptive LASSO Regularization, 7) Ridge Regularization, 8) Elastic Net Regularization. These eight regression models work similarly to linear regression. In addition to their respective advantage introduced in the methodology chapter, they also have different focuses with 1) - 4) strengthening the more influencing factors but weakening the less influencing factors in the model, while 5) - 8) tuning the function by adding penalty term in the error function. These eight models have also been used more widely for dealing with the multicollinearity issue. If any of the performance in these eight models exceeded the corresponding false-positive model performance (i.e., binomial expressions and trinomial expressions on all variants) and at the same time outperformed the other model performances in the eight models, the model would be better to fit the interrelationship between specific word type(s) and COVID-19 pandemic development.

We only presented the performance of the eight fine-tuned regression models by using emergent neologisms to predict the pandemic data. Moreover, all these models are formulated by the main effects of binomial expressions. Additionally, the other reason that we did not use the trinomial expressions as the final model is that the performance in all eight fine-tuned models after dealing with multicollinearity does not outperform the false-positive modeling result, that is, the trinomial baseline. More details can be checked in Appendix C.

Table 24 demonstrates the model performance after being optimized in selecting independent variables by comparison with binomial expressions of all variants of emergent neologisms.

Table 24 Model optimization on all variants of emergent neologisms

Regression types	IVs	DVs	Binomial fine-tuned R^2	Binomial fine-tuned RMSE	Binomial without fine-tuned R^2-baseline	Binomial without fine-tuned RMSE-baseline
Forward Selection	All emergent neologisms binomial main effects	Newly confirmed	0.993	61.777	0.987	74.622
	All emergent neologisms binomial main effects	Newly suspected	0.998	39.296	0.996	58.573
	All emergent neologisms binomial main effects	New deaths	0.948	5.369	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	0.987	424.273	0.985	520.371
Backward Elimination	All emergent neologisms binomial main effects	Newly confirmed	0.990	76.620	0.987	74.622
	All emergent neologisms binomial main effects	Newly suspected	0.997	51.806	0.996	58.573
	All emergent neologisms binomial main effects	New deaths	0.987	2.397	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	0.991	191.841	0.985	520.371
Stepwise Regression	All emergent neologisms binomial main effects	Newly confirmed	0.995	41.926	0.987	74.6212
	All emergent neologisms binomial main effects	Newly suspected	0.999	32.628	0.996	58.573
	All emergent neologisms binomial main effects	New deaths	0.926	6.031	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	0.999	43.880	0.985	520.371
Best: Least Angle Regression	All emergent neologisms binomial main effects	Newly confirmed	**0.989**	**67.329**	0.987	74.622
	All emergent neologisms binomial main effects	Newly suspected	**0.997**	**53.157**	0.996	58.573
	All emergent neologisms binomial main effects	New deaths	**0.960**	**5.209**	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	**0.997**	**232.994**	0.985	520.371

Table 24 Continued

Regression types	IVs	DVs	Binomial fine-tuned R^2	Binomial fine-tuned RMSE	Binomial without fine-tuned R^2-baseline	Binomial without fine-tuned RMSE-baseline
LASSO Regularization	All emergent neologisms binomial main effects	Newly confirmed	0.987	81.052	0.987	74.622
	All emergent neologisms binomial main effects	Newly suspected	0.996	59.894	0.996	58.573
	All emergent neologisms binomial main effects	New deaths	0.975	4.579	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	0.846	1775.776	0.985	520.371
Adaptive LASSO Regularization	All emergent neologisms binomial main effects	Newly confirmed	0.990	77.093	0.987	74.622
	All emergent neologisms binomial main effects	Newly suspected	0.993	65.153	0.9956	58.573
	All emergent neologisms binomial main effects	New deaths	0.982	3.638	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	0.987	521.379	0.985	520.371
Elastic Regularization	All emergent neologisms binomial main effects	Newly confirmed	0.913	215.536	0.987	74.622
	All emergent neologisms binomial main effects	Newly suspected	0.995	58.857	0.996	58.573
	All emergent neologisms binomial main effects	New deaths	0.958	6.169	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	0.956	820.0313	0.985	520.371
Ridged Regression (α= 0.005, the best parameter)	All emergent neologisms binomial main effects	Newly confirmed	0.871	222.753	0.987	74.622
	All emergent neologisms binomial main effects	Newly suspected	0.981	78.732	0.996	58.573
	All emergent neologisms binomial main effects	New deaths	0.823	9.338	0.944	6.152
	All emergent neologisms binomial main effects	Currently suspected	0.912	1128.440	0.985	520.371

Table 25-26 presents the eight fine-tuned model performances of using vector names and PPE names to predict the pandemic development by comparison with binomial expressions without fine-tuned models.

Table 25 Model optimization on all variants of vector names

Regression types	IVs	DVs	Binomial with fine-tuned R^2	Binomial with fine-tuned RMSE	Binomial without fine-tuned R^2-baseline	Binomial without fine-tuned RMSE-baseline
Forward Selection	All vector names binomial main effects	Newly confirmed	0.694	288.724	0.740	327.284
	All vector names binomial main effects	New suspected	0.826	396.657	0.826	356.583
	All vector names binomial main effects	New deaths	0.828	11.599	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.902	1034.805	0.709	2203.119
Backward Elimination	All vector names binomial main effects	Newly confirmed	0.740	351.184	0.740	327.284
	All vector names binomial main effects	New suspected	0.858	263.329	0.826	356.583
	All vector names binomial main effects	New deaths	0.733	14.022	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.732	2355.730	0.709	2203.120
Stepwise Regression	All vector names binomial main effects	Newly confirmed	0.796	1791.897	0.740	327.284
	All vector names binomial main effects	New suspected	0.823	390.969	0.826	356.583
	All vector names binomial main effects	New deaths	0.806	12.732	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.721	1810.696	0.709	2203.119
Least Angle Regression	All vector names binomial main effects	Newly confirmed	0.811	295.701	0.740	327.284
	All vector names binomial main effects	New suspected	0.829	375.165	0.826	356.583
	All vector names binomial main effects	New deaths	Stopped	22.771	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.749	2271.326	0.709	2203.119
LASSO	All vector names binomial main effects	Newly confirmed	0.724	335.850	0.740	327.284
	All vector names binomial main effects	New suspected	0.838	380.253	0.826	356.583
	All vector names binomial main effects	New deaths	0.699	15.075	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.732	1872.487	0.709	2203.119

Table 25 Continued

Regression types	IVs	DVs	Binomial with fine-tuned R^2	Binomial with fine-tuned RMSE	Binomial without fine-tuned R^2-baseline	Binomial without fine-tuned RMSE-baseline
Adaptive LASSO	All vector names binomial main effects	Newly confirmed	0.740	374.214	0.740	327.284
	All vector names binomial main effects	New suspected	0.850	365.671	0.826	356.583
	All vector names binomial main effects	New deaths	0.638	15.144	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.851	1465.684	0.709	2203.119
Elastic Net	All vector names binomial main effects	Newly confirmed	0.650	318.703	0.740	327.284
	All vector names binomial main effects	New suspected	0.817	306.355	0.826	356.583
	All vector names binomial main effects	New deaths	0.699	13.692	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.694	2527.160	0.709	2203.119
Ridge (α= 0.005, the best parameter)	All vector names binomial main effects	Newly confirmed	0.740	327.284	0.740	327.284
	All vector names binomial main effects	New suspected	0.826	356.583	0.826	356.583
	All vector names binomial main effects	New deaths	0.705	13.638	0.705	13.638
	All vector names binomial main effects	Currently suspected	0.709	2203.119	0.709	2203.119

Table 26 Model optimization of all variants of PPE names

Regression types	IVs	DVs	Binomial with fine-tuned R^2	Binomial with fine-tuned RMSE	Binomial without fine-tuned R^2-baseline	Binomial without fine-tuned RMSE-baseline
Forward Selection	All PPE binomial	Newly confirmed	0.958	149.830	0.941	159.224
	All PPE binomial	New suspected	0.989	102.581	0.957	182.420
	All PPE binomial	New deaths	0.843	9.816	0.786	11.903
	All PPE binomial	Currently suspected	0.993	391.880	0.993	362.944
Backward Elimination	All PPE binomial	Newly confirmed	0.988	74.309	0.941	159.224
	All PPE binomial	New suspected	0.979	110.261	0.957	182.420
	All PPE binomial	New deaths	0.791	11.021	0.786	11.903
	All PPE binomial	Currently suspected	0.992	366.927	0.993	362.944

Table 26 Continued

Regression types	IVs	DVs	Binomial with fine-tuned R^2	Binomial with fine-tuned RMSE	Binomial without fine-tuned R^2-baseline	Binomial without fine-tuned RMSE-baseline
Stepwise Regression	All PPE binomial	Newly confirmed	0.884	234.470	0.941	159.224
	All PPE binomial	New suspected	0.838	374.419	0.957	182.420
	All PPE binomial	New deaths	0.880	8.388	0.786	11.903
	All PPE binomial	Currently suspected	0.990	377.127	0.993	362.944
Least Angle Regression	All PPE binomial	Newly confirmed	0.925	147.027	0.941	159.224
	All PPE binomial	New suspected	0.944	223.020	0.957	182.420
	All PPE binomial	New deaths	0.684	13.820	0.786	11.903
	All PPE binomial	Currently suspected	0.994	370.794	0.993	362.944
LASSO	All PPE binomial	Newly confirmed	0.969	122.601	0.941	159.224
	All PPE binomial	New suspected	0.983	109.373	0.957	182.420
	All PPE binomial	New deaths	0.733	10.905	0.786	11.903
	All PPE binomial	Currently suspected	0.815	1865.005	0.993	362.944
Adaptive LASSO	All PPE binomial	Newly confirmed	0.654	241.471	0.941	159.224
	All PPE binomial	New suspected	0.889	252.332	0.957	182.420
	All PPE binomial	New deaths	0.561	17.206	0.786	11.903
	All PPE binomial	Currently suspected	0.994	370.640	0.993	362.944
Elastic Net	All PPE binomial	Newly confirmed	0.955	143.030	0.941	159.220
	All PPE binomial	New suspected	0.989	87.059	0.957	182.420
	All PPE binomial	New deaths	0.716	14.386	0.786	11.903
	All PPE binomial	Currently suspected	0.978	629.185	0.993	362.944
Ridge (α= 0.005, the best parameter)	All PPE binomial	Newly confirmed	0.941	159.224	0.942	159.223
	All PPE binomial	New suspected	0.956	182.419	0.957	182.420
	All PPE binomial	New deaths	0.786	11.903	0.786	11.904
	All PPE binomial	Currently suspected	0.993	362.944	0.992	362.943

Generally speaking, the COVID-19 emergent neologisms have good predictability because the model performance before and after the optimization is higher than 0.9. By comparing the Binomial R^2-baseline and Binomial RMSE-baseline and the other optimized regression models, a better model is achieved by Least Angle Regression. It can solve the multicollinearity issue of feeding the model with all word subcategories of emergent neologisms as independent variables and adjust the coefficients of baseline models for better prediction. Our final model adjusted by Least Angle Regression based on the COVID-19 emergent neologisms is thus as Equation [12][7] shown:

Newly confirmed = -136.8357 + 0.0123*Official_names - 0.0839*Pre-official_names - 0.0003*Under-specifications + 0.6127*Stigmatizing_names - 1.8E-7*Official_names^2 + 1.279E-6*Pre-official_names^2 - 9.2E-6*Stigmatizing_names^2

Newly suspected = -48.4422 + 0.0148*Official_names - 0.0547*Pre-official_names - 0.0031*Under-specifications + 0.01170*Stigmatizing_names + 3.044E-8*Official_names^2 + 1.79E-7*Pre-official_names^2 - 1.28E-6*Stigmatizing_names^2 + 1.03E-5*English_abbreviations^2

New deaths = 5.4026 + 0.0018*Official_names - 0.0003*Pre-official_names - 0.0005*Under-specifications - 0.0147*English_abbreviation - 5.62E-9*Official_names^2 + 2.22E-6*English_abbreviations^2

Currently suspected = 143.0311 + 0.0067*Official_names - 0.0745*Pre-official_names - 0.0045*Under-specifications - 0.5525*Stigmatizing_names + 0.0702*English_abbreviation + 8.03E-7*Official_names^2 - 1.84E-6*Pre-official_names^2 - 2.1716E-8*Under-specification^2 + 1.05E-5*Stigmatizing_names^2 + 4.52E-5*English_abbreviations^2 [12]

[7] The coefficient of each variant of the COVID-19 emergent neologisms is selected based on *p*-value. Appendix D reporting the coefficients is listed after the body text.

The proposed Equation [12] is acceptable. On the one hand, we thought the equation was proper because the parabola curve has already shown the necessity of the binomial expression in the modeling relationship; on the other hand, our equation conforms to the actual development of this sudden major pandemic. At the beginning stage, it stays at a low occurrence, and the growth is not exponential, whereas skyrocketing increases suddenly happen when the pandemic becomes serious. The equations can reflect this prominent feature. Since stigmatizing names and under-specifications took up the most significant proportion at the early stage and official names and English abbreviations did not appear, we supposed stigmatizing names and under-specifications as 10,000, official names and English abbreviations as both 0, and pre-official names as 100. Based on the extrapolated data at the early stage, there would be 5059.325 newly confirmed cases. As the pandemic went deeply, the stigmatizing names were avoided by WHO, and official names and English abbreviations came out. Later, stigmatizing names should be 0, while official names are set to 10,000. Since Chinese people do not widely use English abbreviations and pre-official names, they are both set 10. Under-specifications are also the most significant terms to refer to the pandemic, so they are still 10,000. If so, newly confirmed cases became –35.119. Though negative values are not acceptable in the newly confirmed cases, they can reflect the controlled situation at the later stage of the pandemic—the same as the other three pandemic data. The bigger intercept for stigmatizing names also reflects the big difference between the early and the later stage in the equation compared with the other variants of the COVID-19 emergent neologisms.

However, from the perspective of VIF, it seems that the multicollinearity issue has not been fully solved because VIF values more significant than 10 will undoubtedly lead to this issue.

$$\mathrm{VIF} = \frac{1}{1 - R^2} \qquad [13]$$

According to Equation [13], R^2 larger than 0.9 would have a more substantial possibility of having a multicollinearity issue. The proposed regression model, Least Angle Regression, has all pairs of emergent neologisms predicting pandemic cases with R^2 larger than 0.9, which should lead to a multicollinearity issue as well.

Our explanation for a this 'conflict' is that VIF values can only show the tendency of having multicollinearity issues or not, but cannot confirm that this issue must exist. If the qualitative interpretation of the Equation [12] makes sense from the theoretical perspective (i.e., every variant of the COVID-19 emergent neologisms should have its distinctive characteristic in the regression model), the result still has significance. In addition, the fine-tuned regression models also deal with the multicollinearity issue by the adjusted hyperparameters, which gives a double guarantee of the collinearity avoidance.

According to Table 26, we did not find any pairs with R^2 larger than 0.9, even if these optimized regression models have already processed them. There are two "stops" mentioned in Table 26. When the software finds any selection at a local minimum of the SBC[8] criterion, it will stop the operation. This result echoes the previous finding that vector names may not have big prediction as the COVID-19 emergent neologisms to the pandemic cases. Even if PPE names have more pairs with R^2 larger than 0.9 compared with vector names, they still do not show all R^2 larger than 0.9 in all the pairs, which is why we did not conclude PPE names to be better independent variables to predict pandemic cases.

Unlike the predictability of the COVID-19 emergent neologisms, PPE names do not show all the R^2 larger than 0.9, and the optimized models perform better than binomial regressions on

[8] SBC is the abbreviation of Schwarz Bayesian Criterion, which refers to a criterion for selecting a model. The model with a lower SBC is generally selected.

all variants of PPE names. Chapter Six will discuss why the COVID-19 emergent neologisms have bigger predictability than buzzwords in modeling the pandemic cases.

4.3 Chapter Summary

In this chapter, we answered RQs 1-4 related to the distribution of the emergent neologism searches, the association between important policy announcement and the internet searches to refer to the disease, and the regression modeling between pandemic cases and internet searches to refer to the disease. The pandemic development can be matched with the important policy announcement over the fifteen months after the outbreak. The correlation analysis confirms the close relationship between the emergent neologisms and pandemic cases in the regression modeling section. Based on the strong correlation, we then examined the sigmoidal performance function of log-linear regression and simple linear regression on a single variant and all variants of emergent neologisms compared with buzzwords. Both regression models did not perform better than 0.9 by their R^2. However, we found better interpretability of all variants than the single variant in the regression model. On this basis, we used multiple linear regression to verify our hypothesis and proved the importance of all variants.

Furthermore, based on the initial observations of the developmental patterns of emergent neologisms and pandemic cases, a parabola curve is shown noticeably, which motivates the thesis to try polynomial regression models, i.e., binomial and trinomial. Though trinomial regression performs the best, regardless of the R^2 value and RMSE, it is not as appropriate as binomial regression from theoretical analysis and practical results. However, the binomial regression is still not the final model selection because of multicollinearity issues. Based on eight fine-tuned models, Least Angle Regression performs better, and emergent neologisms are better

to model the pandemic development in the fifteen months after the epidemic outbreak than buzzwords.

CHAPTER FIVE EMERGENT NEOLOGISMS' MONITOR ON PUBLIC ATTENTION

Chapter Four has revealed how COVID-19 emergent neologisms are associated with the important policy announcement and the pandemic cases. However, there is a remaining question about the relationship between the use of emergent neologisms and the change in public attention. Based on the Framing Effect under Prospect Theory, we hypothesized that the use of COVID-19 emergent neologisms could be an important indicator of monitoring the change in public attention. Another consideration supporting us to propose this hypothesis lies at previous relevant studies (e.g., Chew & Eysenbach, 2010; Gesser-Edelsburg *et al.*, 2016): people tend to choose vague names (e.g., *yìqíng* 'pandemic') to refer to the disease in order to weaken the negative impact of the disease on their psychology. On the other hand, the official names including the nature of the virus (e.g., *xīn guān fèiyán* novel corona pneumonia 'coronavirus') are still widely used by the public, which to some extent reflects the public's self-protection awareness from the disease. Different linguistic variants (e.g., vague names versus official names) are selected at different times of the epidemic, reflecting different emotion among the public behind the different variants. Hence, we will explore the monitoring effect on the change of public attention by tracking the use of different variants of the COVID-19 emergent neologisms.

In Section 5.1, we thus proposed an assumption on the public emotional change during different periods of the pandemic based on the general development of the COVID-19 emergent neologisms. Such assumption needs evidence from empirical data, so we crawled Sina Microblog posts to verify such a hypothesis through N-gram co-occurrence calculation in Section 5.2. Section 5.3 summarizes this chapter.

5.1 Hypothesis on the Change of Public Attention by Emergent Neologisms

In this section, we first summarized the general pattern based on the development of the COVID-19 emergent neologisms in Section 5.1.1 and then tried to deduce the change of public emotion at different stages in Section 5.1.2.

5.1.1 Overall Pattern in the COVID-19 Emergent Neologisms

According to Figure 11, the COVID-19 emergent neologisms presents three main stages: the competition and fluctuations at the early stages, while more stable use at the later stages. The early stages refer to the time range from December 21st, 2019 to March 30th 2020, whereas the later stages refer to the time after March 2020. In the first three months after the outbreak, the most noticeable "chaos" happened. It witnessed the fierce rivalry among a variety of COVID-19 emergent neologisms with the dominant use of stigmatizing names with a dominating value of more than 40% and under-specifications around 40% by Chinese netizens. After this period, fluctuations in the first three months have been constantly replaced by the stable development of under-specifications and official names since the coinage of official names to the COVID-19 pandemic. The overall steady development of using COVID-19 emergent neologisms lies in the second and third stages. The second stage experienced a general stable period, while the third stage still witnessed small fluctuations, sometimes with the invasion of a new wave under the generally stable development. For example, stigmatizing names increased their proportions from September to November, 2020 and from February to March, 2021, compared with the other months.

Extending these three stages of the development of COVID-19 emergent neologisms summarized above, we wondered whether they might also correspond to the three main stages of

the change of public emotion at the corresponding time range. Therefore, we tried to deduce the public emotional changes in the next part.

5.1.2 Hypothesis on Public Emotional Change from Emergent Neologisms

According to the three stages of the COVID-19 emergent neologisms deduced from Figure 11, we predicted that there might also be three general stages for the public emotional changes during the fifteen months after the pandemic outbreak.

At the early stage of the emerging pandemic, people may feel frightened, while relaxed emotions gradually replace their fearful emotion when the pandemic received effective control. Indeed, when new waves came back, the public awareness of self-protection would be raised again. The following stages state these three developmental patterns for describing the change of public emotion in more detail.

Stage I. State of being frightened. Figure 17 shows how fear spreads among the public at the first stage. When a new pandemic emerges and there are many emerging infected cases accordingly overnight, the first idea coming to the public's mind may probably be what the disease is. Since people had no idea about the nature of the disease, the names labeled by the affected region (e.g., *wǔhàn bìngdú* 'Wuhan virus') may be temporarily used as a first choice to refer to the disease among the community at this stage in addition to under-specifications (e.g., *yìqíng* 'pandemic'). Buzzwords such as how it happened (e.g., biānfú 'bat') and how to protect themselves (e.g., kǒuzhào 'mask') were also the public's focus. At this period, people were so frightened because there was no official announcement about the nature of the disease, making the disease an "intangible demon" and increasing the horrible atmosphere.

Meanwhile, many unconscientious social media that were involved in the misinformation reportage aggravated the terror among the public (e.g., Kim & Tandoc, 2022). More importantly, to prevent the disease from further spreading among the community, the Chinese mainland government issued stay-home announcement. Though the announcement was confirmed to be very effective in controlling the large-scale spread of the pandemic, the public who were required to stay at home still had probability to be infected were under panic in the environment where the pandemic is not stable. In addition to concerning their health, terror also comes from the concern of whether the disease would heavily influence their workplace and working status. Therefore, Stage I is full of fear and chaos.

Figure 17 Frightened stage

Stage II. State of being relaxed. Three months later, the disease received effective control thanks to international scientists' collaboration and the world's assistance. Hence, the fear moves

to the indifferent stage, as Figure 18 shows. Such control was not only reflected by the decreases of newly suspected and newly confirmed cases per day, but also the public's deeper understanding of the symptoms and the shape of the virus (through *xīn xíng guān zhuàng bìngdú fèiyán* novel type corona shaped viral pneumonia 'COVID-19') based on scientists' findings. It helped unveil the myth of the "intangible demon". People found that avoiding the animal vectors and wearing masks when outside is workable, therefore relieving their fear. What makes people more relaxed is the announcement that they could return to the workplace and talk with others face to face after the boredom for nearly four months only if mask wearing and social distancing were guaranteed. Based on the overall controllable environment at that time, people believed that the COVID-19 pandemic would end in a short time. Hence, Stage II presents the relaxation among the public.

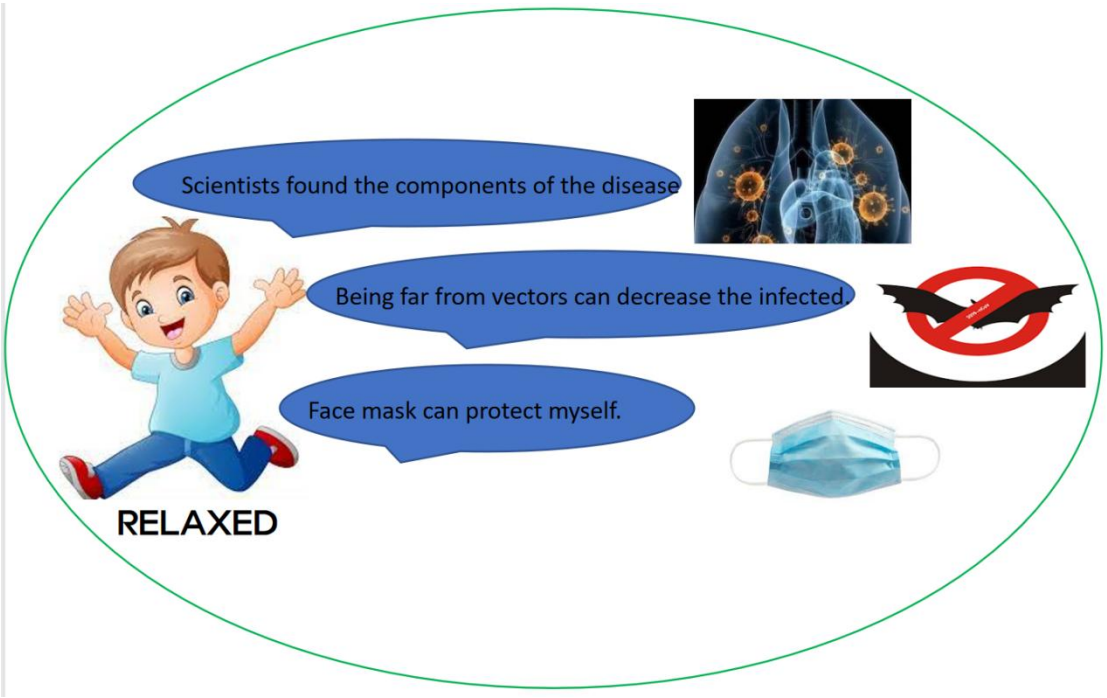

Figure 18 Relaxed stage

In addition to the above two major stages, i.e., the state of being frightened and the state of being relaxed, some minor fluctuations in the state of being relaxed also attracted our attention. According to the pandemic development from December 21st, 2019 to the end of March 2021, even if it has been effectively controlled, new waves have still been experiencing come-and-go. Based on such an observation, we continued a prediction for the public emotion under the overall controlled condition. The general public might also be cautious about the comeback status of the pandemic (as Figure 19 shows).

Stage III. State of being cautious. When the second wave came (nearly October 2020 - February 2021 in mainland China), it can still arouse the awareness of the relaxed people that the pandemic has not ended, though not as tremendously influential as Stage I. After all, new waves reappeared with the varied strains of the COVID-19 pandemic. On the other hand, the international community has been under an abyss of misery because of being heavily impacted by the pandemic (https://www.worldometers.info/coronavirus/). Under such circumstance, people began to stock up on face masks and regularly clean hands again.

Nevertheless, this stage did not make people in too much panic as Stage I did. Instead, they became calmer because they knew what the COVID-19 was, how it happened, and how it should be avoided. They also knew that the prevention from the COVID-19 needs more effort from vaccination, medicine, and individual's collaboration with the government's announcement. As public's cautiousness returned, the discussion on varieties of PPE came back (e.g., *xǐshǒuyè* wash hands liquid 'hand sanitizer').

Figure 19 Cautious stage

In sum, the state of being frightened (Stage I) may only take a short time because the accumulated human knowledge and pandemic tracking systems with big data help us understand an emerging pandemic soon. Such a stage might not return. However, the state of being relaxed and cautious may cycle before achieving 'herd immunity' as shown in Figure 20. The state of being relaxed develops to the state of being cautious as a new wave comes, but the state of being cautious will shift back to the state of being relaxed when the old wave goes. The cycle presented in Figure 19 is a real event happening in Greater China. The public are more likely to let their guard down when the danger goes and pick their self-protection awareness up when the virus come back. Then, we realized a close relationship between the public's psychological status and pandemic development: the more people show relaxation during the pandemic period, the more likely a new wave of pandemic strains come back.

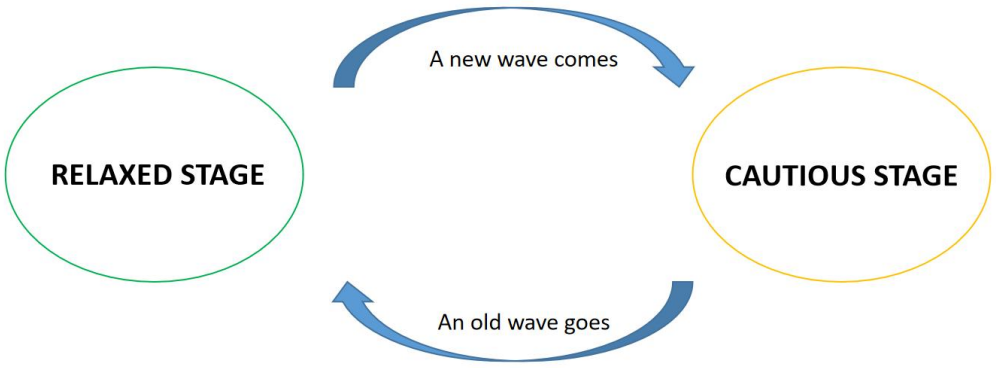

Figure 20 A cycle of relaxed and cautious stages in the emergent pandemic (before 'herd immunity')

However, such cycle cannot happen forever in a pandemic because the public's frustration and indifference would increase with the number of cycle. Former President Cyril Ramaphosa has pointed out this problem in South Africa 'We have let our guard down and are paying the price' and even has been called 'pandemic fatigue' (South African Government, 2021). So, the Health Center and related departments can be aware of and take effective precaution on such a "cycling" issue because the more "cycling" times an emerging pandemic experiences before 'herd immunity', the more time and prices the human has to pay.

5.2 Verification of the Hypothesis based on Sina Microblog Posts

We predicted the change of public emotion in three stages based on the competition and development of the COVID-19 emergent neologisms. This section reports the empirical evidence to prove whether our prediction makes sense. The following operations are based on the 2-6 Chinese words co-occurring with *dài kǒuzhào* 'wear face masks'. As the most important protective equipment to isolate from the virus, *kǒuzhào* 'masks' has undoubtedly become daily necessities. The decision of wearing masks versus not wearing masks, or the willingness to wear masks versus the unwillingness to wear masks, is undoubtedly impacted by individual's

judgement on the seriousness of the pandemic and their attitude to the pandemic. Hence, by tracking the public emotion about wearing masks at different times, the psychological states among the public to the pandemic and self-protection awareness could be reflected.

Section 5.2.1 reports the calculation of 2-6 Chinese words co-occurring with the target chunk, i.e., *dài kǒuzhào* 'wearing masks' at the following five points. They are December 21st, 2019, January 25th, 2020, May 4th, 2020, August 7th, 2020, and December 31st, 2020. Section 5.2.2 reports the calculation of what 2-6 Chinese words co-occurring with the target chunk, i.e., *dài kǒuzhào* 'wearing masks' on January 4th, 2021 and March 30th, 2021. Section 5.2.3 reports that the calculation of 2-6 Chinese words co-occurring with the target chunk, i.e., *dài kǒuzhào* 'wear masks' on May 29th, 2021, and August 4th, 2021. Separating the time points in the second year (January 4th, 2021 and March 30th, 2021 versus May 29th, 2021 and August 4th, 2021) is out of three considerations. Firstly, given that the end date for internet searches of the COVID-19 emergent neologisms and the Sina Microblog data both has March 30th, 2021 as one ending time, the separation at this time point can serve for easy comparison between the public emotion predicted by the internet searches of COVID-19 emergent neologisms and the real public emotion in the Sina Microblog data. Secondly, the exploration of public emotion during the popular social events would inevitably be compared with the Issue-Attention Cycle. It motivates us to extend the time range in the Sina Microblog data to present a more comprehensive picture of the public emotional change during the different periods. We answered whether there is any difference between the public emotion in 2020 and that in 2021 and whether the Issue-Attention Cycle can be applied to interpret the public emotion well during the COVID-19 pandemic. Thirdly, the announcement on getting vaccinated around March 30th, 2021 leads to the separation at this date. However, some newly confirmed cases were still reported around May 2021, even if

the infected people completed the injection. With the marketing of vaccination in mainland China and the changed relationship between vaccination and infection rates, the general public may change their emotion to *dài kǒuzhào* 'wear masks' at different periods.

5.2.1 Change of Public Emotion to *dài kǒuzhào* in the First Year (2019.12.21 - 2020.12.31)

The first period reported in this section includes five time points. The first time point is the first announcement of a confirmed case in mainland China, i.e., December 21st, 2019. According to Figure 20 shown. There is no doubt that *dài kǒuzhào* 'wearing masks' is the most focused chunk in the figure. We found that the chunks associated with this target chunk are involved in various events. The most frequent 2-6 Chinese words connected with *dài kǒuzhào* 'wearing masks' (i.e., relatively bigger circles) are *wùmái tài dà* 'Fog is too heavy', *míngxīng* 'superstars', *jīchǎng* 'airport', and *dàibǔ* 'arrest'. Those highly connected chunks are closely associated with the occasions and locations where it is reasonable to decide to wear masks in December 2019. Heavy fog has been a severe problem in the northern regions of China. To avoid the harm of fog to the respiratory tract, wearing masks becomes one of the most efficient ways of preventing the harmful effects. In addition to common people, superstars are also a particular group who need to wear masks, especially at the airport. Be pursuing fashion or avoiding paparazzi, masks seem to is necessary for superstars. The connection of *dài kǒuzhào* 'wear masks' therefore conforms to reality. The last chunk of solid association with *dài kǒuzhào* 'wear masks' is *dàibǔ* 'arrest', which echoes the status of that period when the rebels were used to wearing face masks at Hong Kong. In addition to the most frequent connection to *dài kǒuzhào* 'wear masks' on December 21st, 2019, Figure 21 also presents the other chunks connected with the target chunk, such as *dōngtiān shìhé* 'It is proper for winter' and *yuǎnlí líugǎn* 'stay away from the flu'. *kōngqì bù*

xīnxiān ‘Air is not fresh’, which also reflects the necessity of wearing masks under the low temperature.

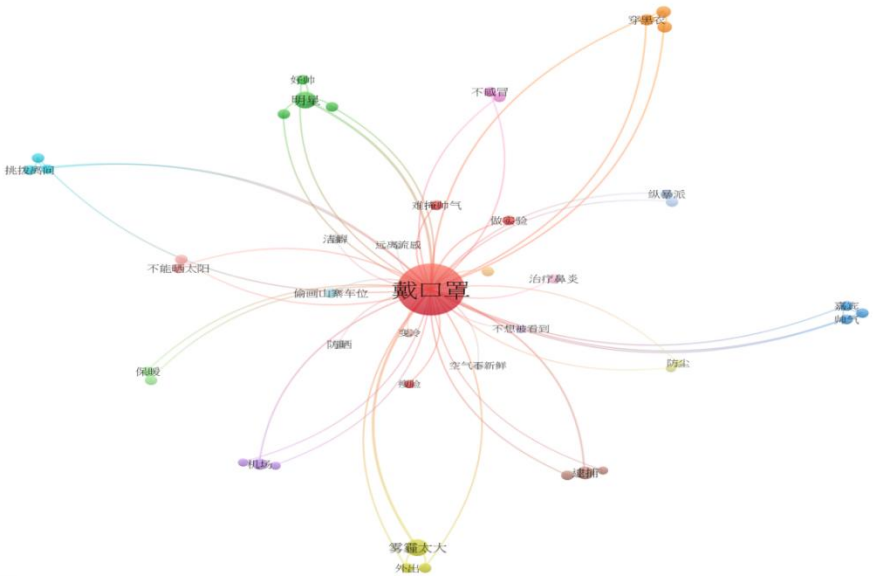

Figure 21 Chunk co-occurrence with the target chunk *dài kǒuzhào* ‘wear masks’ on December 21st, 2019

Though the chunks shown in Figure 21 respond to the social events that happened at the end of December 2019, we did not notice any mention of wearing masks to prevent the pandemic. Hence, on the first official day of announcing the confirmed case, the public has not realized the seriousness of the pandemic.

However, when it came to January 25th, 2020, a different picture was presented in Figure 22. The circle for the target chunk becomes surprisingly large, meaning there are more mentions and discussions of wearing masks compared to the end of December 2019. The most associated chunks are *chūmén* ‘go outside’, *qín xǐshǒu* ‘keep washing hands’, *jìdé* ‘remember’, *yīdìng* ‘must’, *bù yào huāng* ‘Don’t panic’, *píngpíngānān* ‘May everyone be sound and peaceful’, and *jiànjiànkāngkāng* ‘Be fit’. According to the chunks, the general public has realized the seriousness of the pandemic by highlighting the importance of hand-washing and mask-wearing.

In addition, it also pays attention to the occasion where it is required to wear masks (reflected by *jìdé* 'remember' and *yīdìng* 'must') when going outside (reflected by *chūmén* 'go outside'). Even if the public raises strong awareness of carrying all necessary equipment, the encouragement and comfort words like *bù yào huāng* 'Don't panic', *píngpíngānān* 'May everyone be sound and peaceful', and *jiànjiànkāngkāng* 'Be fit' are also frequently used in the posts in order to achieve the purpose of mutually cheering up. The other chunks are highly related to the announcement during the pandemic, such as *jìn jùlí jiēchù* short distance contact 'close contact', *fāshāo* 'have a fever', *késou* 'cough', *bù yào qù rén duō de dìfang còurènao* 'Don't have fun to crowded places'.

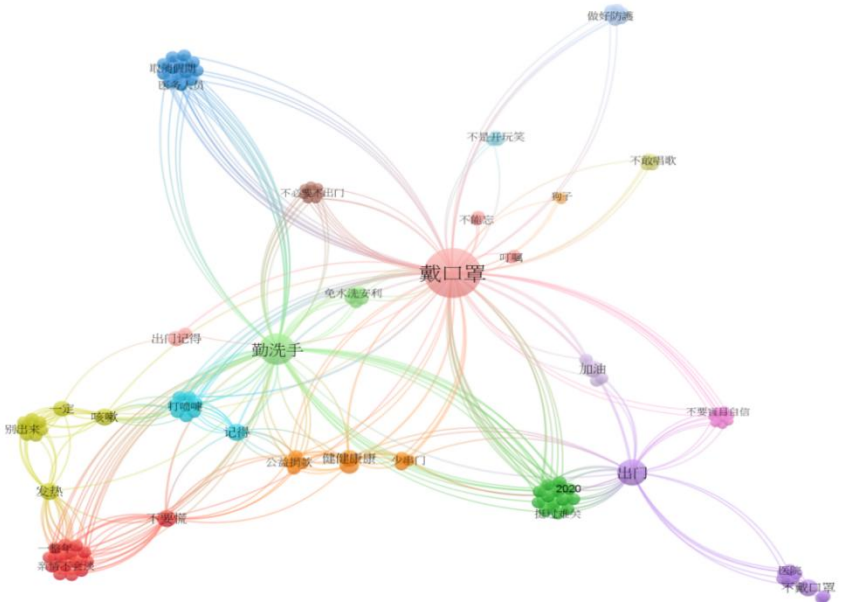

Figure 22 Chunk co-occurrence with the target chunk *dài kǒuzhào* 'wear face masks' on January 25th, 2020

With the complete lockdown in mainland China around Chinese New Year in 2020 as well as the sky-rocketing growth of the newly confirmed cases day by day, the general public realized how terrible the disease is by including all disease-related words in their posts.

Thanks to the concerted efforts of the Chinese government and the whole world, the pandemic cases witnessed effective controls, with almost no newly confirmed and newly

suspected cases appearing after May 2020, as Figure 23 shown. While praying for the quick end of the pandemic (i.e., *yìqíng* 'pandemic' and *kuàikuài jiéshù* quick end 'Hope it will end quickly') and reminding the others to look after themselves (e.g., *zhào gù zìjǐ* mask look.after yourself 'look after yourselves by face masks'), the general public shows an obvious dispreference for wearing masks. The most associated chunks are *rè* 'hot', *38 dù* '38 degree', *xiàtiān* 'summer', *tàiyáng dǐxià* 'under the sun', *quán shēn shàng xià yòu rè yòu nián* whole body up down again hot again sticky 'hot and sticky of the whole body'. They clearly show how unwilling the public was to wear face masks under the hot temperature and when the pandemic has received effective controls. In addition to complaints about wearing masks in summer, some posters directly used an evaluation *lèi* 'tired' to summarize the current status of wearing masks psychologically and physically. Worse, the disadvantage of wearing face masks always also has disfiguration, which might probably influence finding boyfriends and may trigger *cùsǐ* 'sudden death'.

On the other hand, on the rightmost side of Figure 23, the red color presents the strong connection between the runner and *bù dài kǒuzhào* no wear mask 'Don't wear masks'. Such a connection represents a stark contrast with January 25th, 2020 when it requires everyone to wear masks outside. However, after five months of the pandemic outbreak, some people do not wear face masks especially outside and this behavior seemed like tacit consent by the relevant departments.

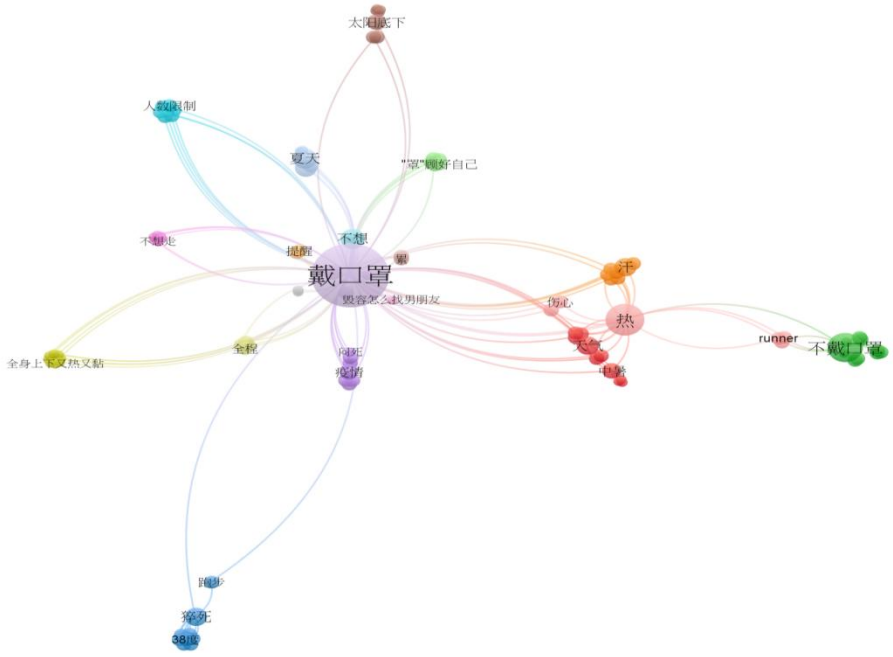

Figure 23 Chunk co-occurrence with the target chunk *dài kǒuzhào* 'wear masks' on May 4th, 2020

Logically, the negative attitude generated on May should continue to August when the temperature increases up to 40 degrees. Meanwhile, August is the first summer term after the pandemic outbreak for school students and staff. This is an excellent time to go outside after the stay-at-home enforcement policy for consecutive three months. Hence, August should be theoretically a crowded time. Plus, with the high temperature, it is easy to assume that the negative emotion among the public should become very obvious.

Nevertheless, Figure 24 demonstrates a different scene. Even if the general attitude remains negative, which is reflected by the chunks of *jùyǒu tiǎozhànxìng* have challenges 'very challenging', *shénme shíhou jiéshù* what time-end 'When will end', *bù yuànyì* no willing 'unwilling', as well as *réngrán* 'still', there are some self-consolation emotion that appeared among the public in order to relieve the negative emotion. For example, some people think that wearing masks is a good way of making the face look small (e.g., *liǎn xiǎo* face small 'display

face small'), covering up the acne (e.g., *zhē zhù dòu* cover up acne 'hide acne' and *tóufa yóu* hair oily 'Hair is so greasy'), and helping to pretend a hard-working state (e.g., *jiǎzhuāng nǔlì* pretend persistence 'pretend efforts'). Meanwhile, some bloggers posted extra information about *tú hùfūpǐn* apply cream 'apply skin products', *jiǎn ge dòng* cut a hole 'cut a hole on the face mask', as well as the information about *huíyìng jiěsuǒ respond unlock* 'respond to unlock the iPhone by face ID' under the period when mask wearing seems to be 'forever' behavior. Having more advantages than the polarity of sentiment analysis by NLP approaches, the chunking co-occurrence can provide more detailed information about the public's emotional change, including self-consolation and extra information focus under the general negative scores of the attitude. Certainly, some other posts care about epidemic prevention by pointing out the unqualified face masks (*bù hégé de qīngliáng kǒuzhào* not qualified DE cool mask 'unqualified cool and refreshing masks'). These posts refresh the purpose of the necessity of wearing masks in the external environment (*wàichū* go.outside 'go outside'), especially in confined spaces (e.g., *zuò dìtiě* take subway 'take the subway').

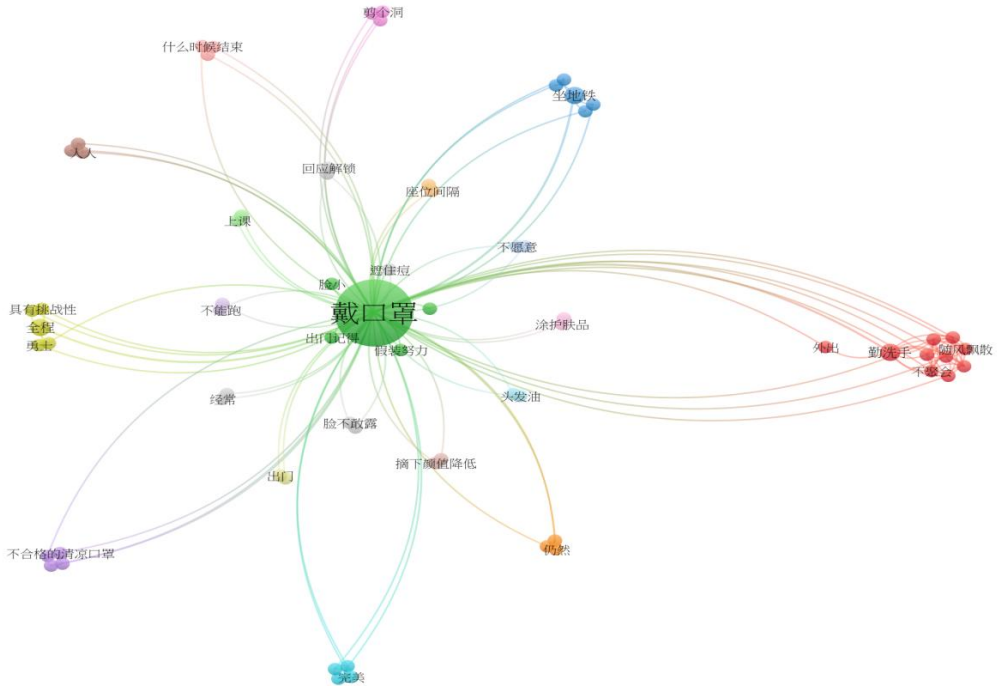

Figure 24 Chunk co-occurrence with the target chunk *dài kǒuzhào* 'wear masks' on August 7th, 2020

The last time point in this section is the last day of 2020. As typically, people made a summary of the whole year on December 31st, 2020. As Figure 25 indicates, the most connected chunks are *2020 nián* 2020 year '2020', *2020 nián biànde jí wéi bù píngfán* 2020 year become very exceptionally no ordinary '2020 becomes extremely unusual', *03 nián fēidiǎn* 2003 year SARS 'SARS in 2003', and 512 *dìzhèn* 512 earthquake 'Sichuan earthquake on May 12th, 2008. These chunks expressed the public's sighing emotion to the whole year when experiencing the cruelty of the pandemic, frustration of lockdown, *etc.*, making them unwilling to summarize (e.g., *bù tài xiǎng zài zhèlǐ zǒngjié* 'Don't want to make a summary', *yīqiè fǎngfú dōu zài sān sì yuè* all seem all at March April 'Seems the time is still at March and April'). Some netizens also compared the COVID-19 pandemic with the SARS in 2003 and the Sichuan earthquake in 2008 to express the sad emotion. In addition, there are a small number of people who expressed extremely negative

feelings by using *jìng gàn zhè tuō le kùzi fàngpì de shì* only do this take.off LE trouser fart DE thing ‘Do things with formalism’ to convey the dissatisfaction to some measures taken by the government on mask wearing.

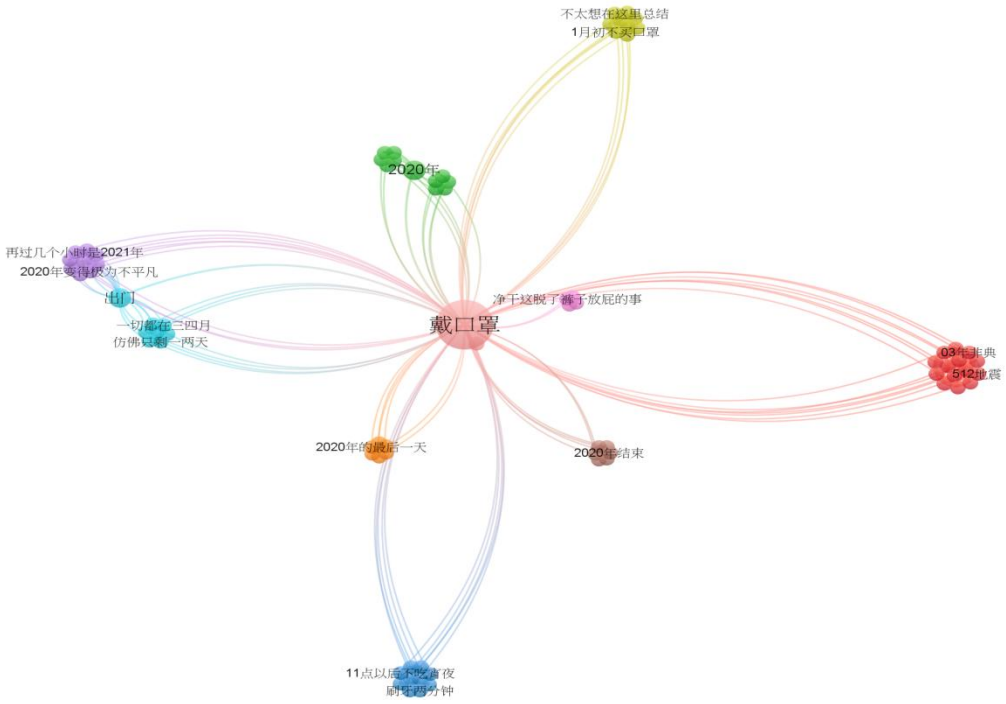

Figure 25 Chunk co-occurrence with the target chunk *dài kǒuzhào* ‘wear masks’ on December 31st, 2020

Overall, the general emotion is undoubtedly negative in the whole year of 2020. However, different from the result of polarity analysis for the sentiment by NLP approaches, our analysis based on the N-gram co-occurrence provided more specific emotional information under the general negative attitude. That is, at the most severe time in which the general negative attitude was shown to the seriousness of the pandemic, people also had the intensive awareness to protect themselves even if in the deep summer. In addition, Chinese netizens shifted their attention to some other exciting activities, such as applying the skin product to preserve moisture, adjusting the iPhone face ID algorithm to recognize the face with mask wearing, *etc.*, to neutralize the general negative attitude.

To a more significant extent, the N-gram co-occurrence visualization of the whole year in 2020 supports our hypothesis of the three-state psychological change from the frightening emotion to the cautious emotion included with the relaxing emotion. The development of the public attention in the first-round Sina Microblog data can be partly covered by the Issue-Attention Cycle because it begins with no attention among the public and experiences the chaotic emotional change. However, the less attention to mask wearing and the replacement by the other important issue described in the Issue-Attention Cycle are not reflected in the Sina Microblog data in 2020. Instead, the attention to mask wearing has remained among the public. The remaining attention may be probably influenced by the constant new strains of the virus. On the other hand, the public also tries to "compromise" to and "co-exist" with the environment where mask wearing has become a necessary self-protection method by finding some interesting activities.

So, in the next section, we further checked how public emotion changed in the second year after the pandemic outbreak and examined how it interacts with the Issue-Attention Cycle. One thing that should be noted is that the last day of 2020 commonly shows the public's summary of the past year. We assumed that it would be similar to the last day in 2021, so we did not include the posts of this date in the next section.

5.2.2 Change of public emotion to *dài kǒuzhào* in the Next Year (2021.1.4 -2021.8.2)

On the first day after the New Year public holiday of 2021 (January 4th), people need to go back to work, which may cause the phenomenon of crowdedness in the confined environment, thereby providing the spread of the virus with sufficient environment again. Unsurprisingly, the most frequent chunk connected with the target phrase *dài kǒuzhào* 'wear masks' is still *qín xǐshǒu*

often wash.hand ‘Wash hands regularly’ and *jiānchí* ‘insist’ under the low temperature, as Figure 26. The other relevant chunks such as *bǎohù zìjǐ* ‘self-protection’, *bǎochí shèjiāo jùlí* ‘keep social distancing’, *bù zhāduī* no gathering ‘Don’t gather together’ are also given much attention again, as last January. However, different from last January 25th, 2020, this January provides netizens with a great deal of consideration about the vaccination and concern the virus mutation. For example, *biànyì xīn guān bìngdú* varied novel corona virus ‘mutated COVID-19 virus’ and *hésuān jiǎncè jiéguǒ chéng yángxìng* nucleic.acid test result present positive ‘Nucleic acid test is positive’ is closely connected with wearing face masks. In addition, the consideration on whether the protection rate can reach 100% after getting injected is also the public’s focus when discussing the disease.

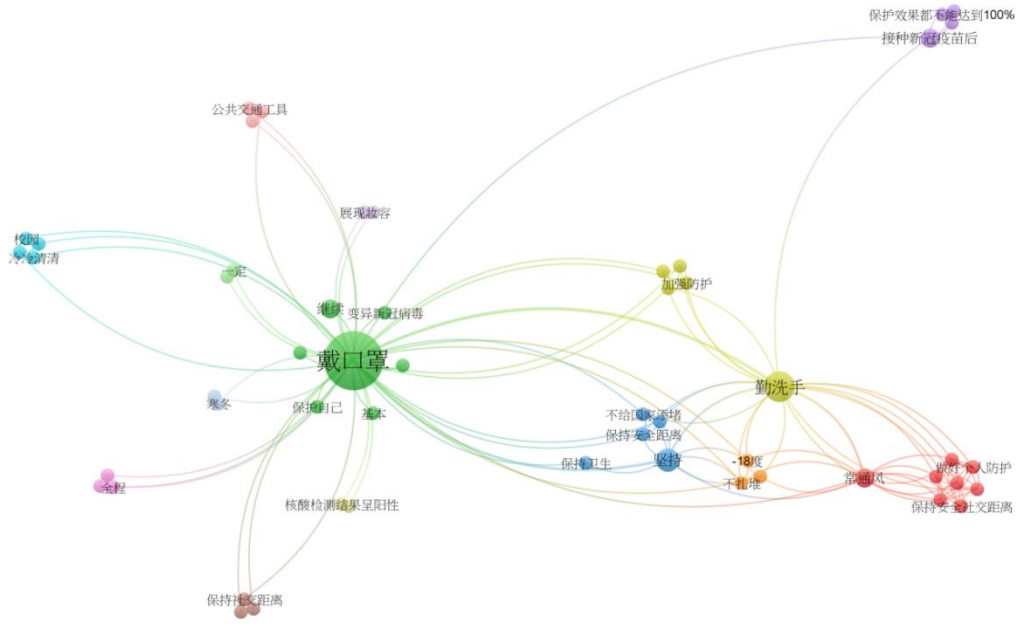

Figure 26 Chunk co-occurrence with the target chunk *dài kǒuzhào* ‘wear masks’ on January 4th, 2021

The association between getting injected and wearing masks is not that strong in January 2021 because the vaccination went to the Chinese market and was not widely injected by the public.

By contrast, as Figure 27 shows, *jiēzhòng xīn guān yìmiáo* inoculate novel corona vaccination 'get Sinovac' becomes the most substantial connection with the target chunk *dài kǒuzhào* 'wear masks' on May 29th, 2021. Another difference is evident because the public did not show much negative attitude compared with last May when the public considered that the pandemic will end soon and mask wearing seemed unnecessary and uncomfortable under the hot weather. By contrast, people realized that the pandemic remains severe and would not end in a short time. Even if getting injected, there is also the possibility of getting infected.

Meanwhile, mask wearing has become unavoidable after all the pandemic seemingly has 'endless' variants. Based on the above perception, the public kept cautious by still frequently using *bǎochí ānquán shèjiāo jùlí* 'keep social distance safe ', *qín xǐshǒu* often wash.hand 'Wash hands regularly', and *jiāqiáng fánghù* 'strengthen protection', which are highly connected with the target phrase. In addition, the epidemic in Guangdong reoccurred around January 2021, and netizens attributed the epidemic's come-back to the status at that period where the public might be unwilling to wear masks.

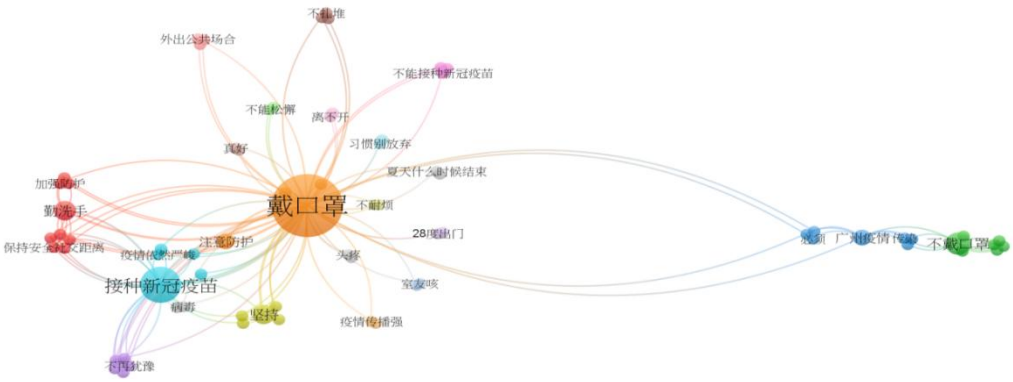

Figure 27 Chunk co-occurrence with the target chunk *dài kǒuzhào* 'wear masks' on May 29th, 2021

Even if the deep summer (August 7th, 2021) came again, the public's major attitude to mask wearing tends to be positive. Unlike last August, when people complained about mask wearing under the hot weather, this August highlights the circulation of the pandemic (e.g., *yìqíng fǎnfù* pandemic repeat 'the pandemic circulates' and *nánjīng yìqíng* 'Nanjing pandemic'). Even if more than 70% of people completed the vaccination injection, there are still a small group of people who were infected. Such cases increased the public's dependence on masks. The chunks connected with *dài kǒuzhào* 'wear masks' like *bù fàngsōng* no relax 'Don't relax', *14 tiān méi líkāi dāngdì* 14 days no leave local 'Don't travel outside the *cè tǐwēn* 'measure temperature' are still frequently used among the public. Figure 28 shows more details.

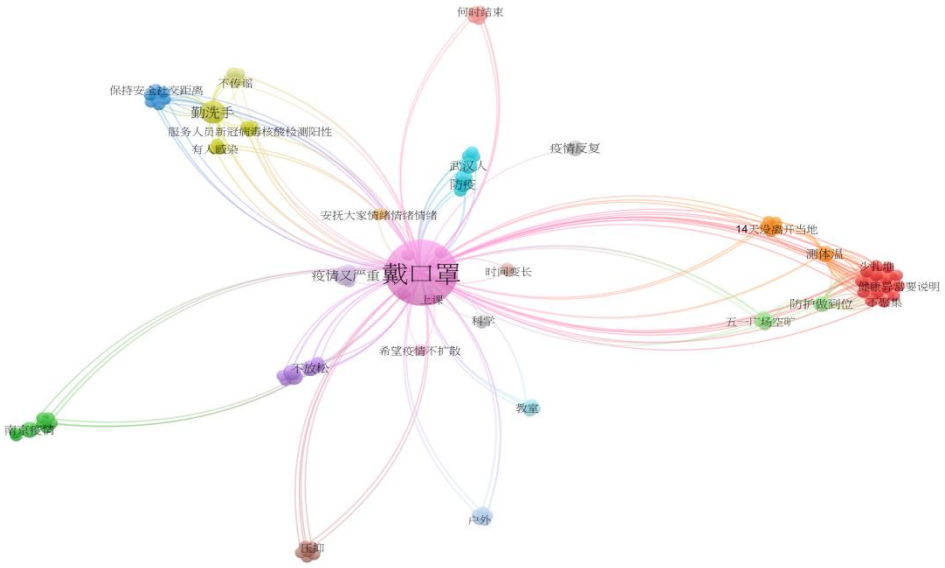

Figure 28 Chunk co-occurrence with the target chunk *dài kǒuzhào* 'wear masks' on August 7th, 2021

5.3 Chapter Summary

This chapter summarized our assumption of how public emotion might change during the different periods of the pandemic. Based on the competition among the COVID-19 emergent

neologisms over the fifteen months, we found that the early stages demonstrated more "confusing" development, while the later stages showed a more stable pattern. However, there are some minor fluctuations during the later stages. Based on such observations, we hypothesized that there might be three stages to describe the change of public emotion. That is, the State of being frightened, the State of being relaxed, and the State of being cautious. At the early stage of the pandemic, people may have little idea about the virus' shape and nature by predicting from the observation that varieties of emergent neologisms were used to refer to the disease. Such confusing status in the emergent neologisms is considered to be impacted by frightening emotion among the public. In the later development of the pandemic, emergent neologisms presented a more stable picture by only under-specifications and official names. Under such a stable circumstance, the relaxing emotion would replace the frightening emotion at the early stage. However, some new waves and strains can still be witnessed under a generally stable environment after April 2020, which undoubtedly brings about the emergence of third-stage emotion among the public, i.e., the State of being cautious. These careful feelings made underused emergent neologisms (e.g., stigmatizing names and pre-official names) reoccur on the Baidu Index. However, the increased ratios of the underused emergent neologisms did not influence the whole development.

The three-stage public emotion (i.e., frightened, relaxed, and cautious) over the fifteen months after the pandemic outbreak were then verified by the Sina Microblog data by the search phrase *dài kǒuzhào* 'wear masks'. We only selected some critical time points for the verification. In the first part, December 21st, 2019, January 25th, May 4th, August 7th, and December 31st, 2020 were selected. We found that the emotion are generally negative and neutral in the first part, where people generally got into a panic on January 25th and all discussions were about how to

protect themselves from the disease. The posts are full of reminders about pandemic control and prevention, and mutual encouragement about insistence by 2-6 Chinese words co-occurrence calculation.

Shortly, the relaxed emotions witnessed growth on May 4th, 2020. For example, the chunks about having opportunities to go outside and gather with friends grew noticeably. However, the transient relaxation was replaced by an unsatisfactory attitude on August 2020 of still being required to wear masks in the open place during the deep summer. Though the public's generally relaxing attitude remained and dominated on August 7th, the awareness of the self-protection still came back. When it came to the last day of 2020, the public expressed their wishes that the pandemic would end soon. Hence, Sina Microblog data throughout 2020 efficiently confirmed our hypothesis of public emotional change based on the search frequencies of emergent neologisms. Compared with the sentiment polarity analysis by the NLP approaches, the N-gram co-occurrence algorithm elaborates the public emotions by more fine-grained context.

Besides, we also found some interesting results: the general public did not show as unwillingness to mask wearing on May and August 2021 as in the same months last year. Rather, they highlighted the importance of mask wearing in epidemic prevention from Sina Microblog data. Unlike what the Issue-Attention Cycle (i.e., little attention -- more attention -- chaos -- steady drop -- replacement) predicts, the Sina Microblog data in May and August 2021 presents a constant self-protection awareness. In the deep summer, people still think they must wear masks as one of the self-protection ways. Hence, the public attention does not show many negative attitudes toward mask wearing. We assumed that the contradiction to the Issue-Attention Cycle might be the intervention of vaccination, which would be discussed in detail in Chapter Six.

CHAPTER SIX DISCUSSION

Based on Chapter Four and Chapter Five, we have reported how the COVID-19 emergent neologisms reflect important policy announcement, predict the pandemic cases, and monitor the change of public emotion. In this section, we will discuss these interesting findings. First, the role of COVID-19 emergent neologisms in reflecting the important policy announcement, predicting the pandemic cases, and monitoring the change of public attention is better than the buzzwords. Hence, Section 6.1 will discuss the advantage of using the COVID-19 emergent neologisms as predictors to respond to the collective human behavior compared with buzzwords.

Section 6.2 will discuss the findings based on the regression modeling. The regression result demonstrates that the binomial expressions based on all variants of the COVID-19 emergent neologisms are a better formula to predict the pandemic cases than other regression models. Hence, we will discuss why the combination of all the variants of the COVID-19 emergent neologisms performed better than single variants and why the binomial expressions of the COVID-19 emergent neologisms performed better than other expressions in Section 6.2.1. Section 6.2.2 will discuss why the Least Angle Regression is a better model. Before the regression modeling, we also examined the correlation relationship between the COVID-19 emergent neologisms and the pandemic cases. However, the high correlation does not necessarily represent the high regression result, which motivates us to discuss such a 'contradictory' finding in Section 6.2.3. In retrospect, one of the theoretical issues examined in this thesis is how much the classical S-curve can be applied to explain the development of the COVID-19 emergent neologisms, which will be discussed in Section 6.2.4.

Section 6.3 will discuss the findings based on the N-gram co-occurrence calculation. The monitoring effect of the COVID-19 emergent neologisms to the change of public attention has been verified by the N-gram co-occurrence visualization. Compared with sentiment polarity analysis by the NLP approach, the co-occurrence algorithms provide more fine-grained information about the public emotion. The advantage of using the N-gram co-occurrence of social media data to monitor the change of public attention than the sentiment polarity analysis will be discussed in Section 6.3.1. Another theoretical issue of whether the traditional Issue-Attention Cycle can be applied to the change of public attention during the COVID-19 pandemic will be discussed in Section 6.3.2.

Section 6.4 summarizes this chapter.

6.1 A Better Indicator to Collective Human Behavior: Emergent neologisms

The COVID-19 emergent neologisms have been empirically evidenced by their more sensitive function to respond to the collective human behavior than buzzwords. The more sensitiveness by the COVID-19 emergent neologisms may come from four reasons. First, they are the most core words to describe the pandemic. By contrast, although the buzzwords are popular among the Chinese netizens, they are not the core words to refer to the disease. The very close association between neologisms and the disease can be dated back to the 17th century in Europe. At that time, there were more extraordinary recurrent plagues in Europe, causing the death of nearly 100,000 Londoners (over one-fifth of the total population of this city) and almost 1 million French. The under-specified references like ‘epidemic’ and ‘pandemic’ were coined to describe the great plagues at that time. More specifically, the symptom of black pustules on victims’ skin was officially named the Black Plague at first. Still, after a while, people were more likely to use its

synonym Black Death, which might raise the public's awareness of self-protection by directly using 'death' in the terminology. Hence, the terminologies for the disease include varieties of names for different purposes.

As mentioned above, official names and also the names achieving the function of attracting the public's awareness would be used. Comparatively, the names to raise the public's awareness of self-protection seem to have a more extensive communication power by more prominent word usage like 'death'. Furthermore, abbreviations are also widely used to name the disease. For example, the term 'Spanish influenza' in 1890 and the term 'Poliomyelitis' in 1878 were reduced to the term 'Spanish flu' and the term 'polio', respectively. Recent decades have also witnessed the coinage of new words to refer to epidemics such as AIDS and SARS. The former was shortened from 'acquired immune deficiency syndrome' in 1982, while the latter was shortened from 'severe acute respiratory syndrome' in 2003.

Our categorization of different variants of emergent neologisms involving under-specifications, official names, and English abbreviations conforms to the conventional way of naming the disease, which should well respond to the disease. Stigmatizing names, though having not been mentioned explicitly by the WHO or other authorities before the COVID-19 pandemic, were also used in the previous disease, including 'Spanish' in 'Spanish influenza', which makes the inclusion of stigmatizing names in the COVID-19 emergent neologisms acceptable. In this case, pre-official names which refer to the variants mentioning the similar symptoms but not making them explicit also make sense.

Second, the sensitiveness of the COVID-19 emergent neologisms is also reflected in the wide use of shortened linguistic form to refer to the disease, thereby promoting the interaction between the emergent neologisms and collective human behavior. According to the Zipfian law

(1949), the shorter the word length is, the more frequently the word occurs in the natural language. Our observation shows under-specified references, usually with only two Chinese characters (e.g., *fèiyán* 'pneumonia'). Due to their short word length, they can effectively reduce the users' memory load and increase communication efficiency related to the pandemic topic compared with other variants. This can explain why Baidu netizens more widely used them. On the other hand, under-specified references are generic words to represent disease, regardless of time and place. The studies in L2 acquisition reported that generic words are more likely to be frequently used by people (Ellis, 2006; Ellis & Larsen-Freeman, 2009; Ellis & Collins, 2009). The effect of word length can be also reflected in official names.

Our data did show that the use proportion of *xīnguān* novel corona 'COVID-19' exceeds the usage of its full name *xīnxíng guānzhuàng bìngdú fèiyán* novel type corona shape viral pneumonia 'COVID-19' in the percentage by Baidu netizens, thereby confirming the important role of frequency in impacting the communication power of the variants of emergent neologisms. Though English abbreviations did not show noticeable usage compared with the other variants over the fifteen months after the pandemic outbreak, their occurrence in the Baidu netizens' searches echoes the general public's preference for the variant with shorter lengths by a higher frequency used in "COVID-19" than "Corona Virus Disease".

The third reason for explaining the advantage of emergent neologisms to respond to the collective human behavior than buzzwords is under the Framing Effect of Prospect Theory. The development of stigmatizing names, official names, and under-specifications in the emergent neologisms are better to reflect the impact of public emotion under the gain-loss competition. In the first month of the pandemic (i.e., December 21st-31st, 2019), under-specifications accounted for the most significant proportion. The most significant use of under-specifications show that

the public had little idea of the nature of the disease; on the other hand, they conveyed the purpose of the public's making the deadly disease vague. The vagueness and abstractness can to a larger extent weaken the panic among the public. When entering the mid January 2020, stigmatizing names witnessed exponential growth, exceeding under-specifications for one whole January, 2020. This result can be explained by emotional change among the public. The city of Wuhan is the province that experienced the most prominently confirmed cases and deaths in the whole January 2020 in China. According to the gain-loss framework, the wide usage of stigmatizing names during that period seems to show the attitude of the people from other provinces that they wanted to separate themselves from the Wuhan people. The wide use of stigmatizing names at the most serious time of the pandemic was to intensify the other people's loss. During January and early February, the term 'Wuhan' seems to be a sensitive word to the people who live outside Wuhan. In addition to the region where the first pandemic case was officially reported, Wuhan also experienced being locked down. This region might aggravate its sensitivity to the public. The wide use of stigmatizing names is more likely to be used by the people outside Wuhan who thought 'objectively' stating the fact of a pandemic. Even some websites (e.g., https://news.sina.cn/zt_d/yiqing0121) which records the pandemic cases divided the pandemic data by the city of Wuhan or Hubei province versus outside Wuhan or Hubei, which can furthermore reflect how sensitive the general public was to the term of 'Wuhan'. There is no doubt that such variants containing the regional information deliver the discrimination, which should be abandoned (Ghebreyesus, 2020).

By contrast, the end of February witnessed a nationwide spreading of the COVID-19 pandemic, not merely centering on the city of Wuhan. Since this time, the proportion of stigmatizing names declined exponentially and was gradually replaced by official names and

under-specifications on the Baidu Index. They even disappeared from the public's conversation after April 2020. Since then, the COVID-19 pandemic was not an issue excluded for one city but a nationwide problem. The gain-loss framework interprets that this dramatic change of using official names and under-specifications to replace stigmatizing names is confronting the self's loss rather than substituting the self's loss with the other's loss. Under such a circumstance, the public outside Wuhan could not use stigmatizing names again to so-called 'objectively' state this disease. Rather, the general public tended to employ more neutral and unbiased names such as under-specifications or official names to refer to the disease because they realized that the disease was also posing a significant threat to their health and life, i.e., self's loss.

Meanwhile, we also noticed that the gain and loss framing is not a trade-off effect. Rather, they showed interchangeable use by people. Generally, the public uses under-specifications to refer to the disease, which is consistent with the gain framing strategy. The purpose of using them is to make the disease opaque. When the pandemic becomes extremely serious, and at the same time, the public does not master the knowledge about the disease, pre-official names and stigmatizing names were frequently used, even dominating the other variants of emergent neologisms. The wide use of these two types of emergent neologisms reflects the loss framing strategy preferred by the public. The uncertainty achieved by pre-official names (e.g., *bùmíng yuányīn fèiyán* unknown reason pneumonia 'pneumonia with unknown reasons') and the downplayed characteristics achieved by stigmatizing names (e.g., *wǔhàn fèiyán* 'Wuhan pneumonia') highlight the seriousness and negative effect of the pandemic. When the information about what causes the disease, how it transmits, and how it can be avoided has been gradually clarified, the official names reflecting the disease by *xīn* 'novel' and *guān* 'corona' have been widely used among people. At this period, the gain framing strategy comes back by

such type of emergent neologisms in order to objectively describe the disease. Different from prior studies on replacement-based disease neologisms at different periods (e.g., Chew & Eysenbach, 2010; Gesser-Edelsburg *et al.,* 2016) where only the first-used name and the finalized name were examined, this thesis uses non-replacement tracking methods and captures the nature of non-linearity and interchangeability of emergent neologisms over the long-term development.

Apart from the impact of frequency and the public emotion under the gain-loss framework, the competition and development of emergent neologisms are more apparent to be influenced and discussed by the public than buzzwords, which also reflect that they are good indicators. Stigmatizing names showed dramatic declines and almost disappeared in a very short time, though they exceeded under-specifications for a month. Their 'disappearance' reflects the effect of WHO's announcement that any disease should not be named by linking the names of malignant diseases with regions, animals, or individuals in order to avoid any discrimination (Ghebreyesus, 2020) on February 11st, 2020. After WHO announced the official name, a friendly reminder that jumps out each time when using Baidu once the users were found to search for stigmatizing names. In addition, the pandemic has spread to almost every corner around the world, which may lead to fewer use of stigmatizing names to make "Wuhan" and "China" stand alone. The use of English abbreviations is the smallest proportion probably because the netizens targeted are Chinese, and they are more accustomed to select Chinese terminologies to describe the pandemic. However, the English abbreviations, also created by the WHO officially and used internationally, were also used by social media to echo the international usage.

From the above, emergent neologisms should be a better indicator to respond to the collective human behavior than buzzwords. Concerning the competition of different word

categories to refer to disease at different time periods, we summarized an interaction network combining the variant selection of emergent neologisms, collective human behavior, and pandemic stages in Figure 29.

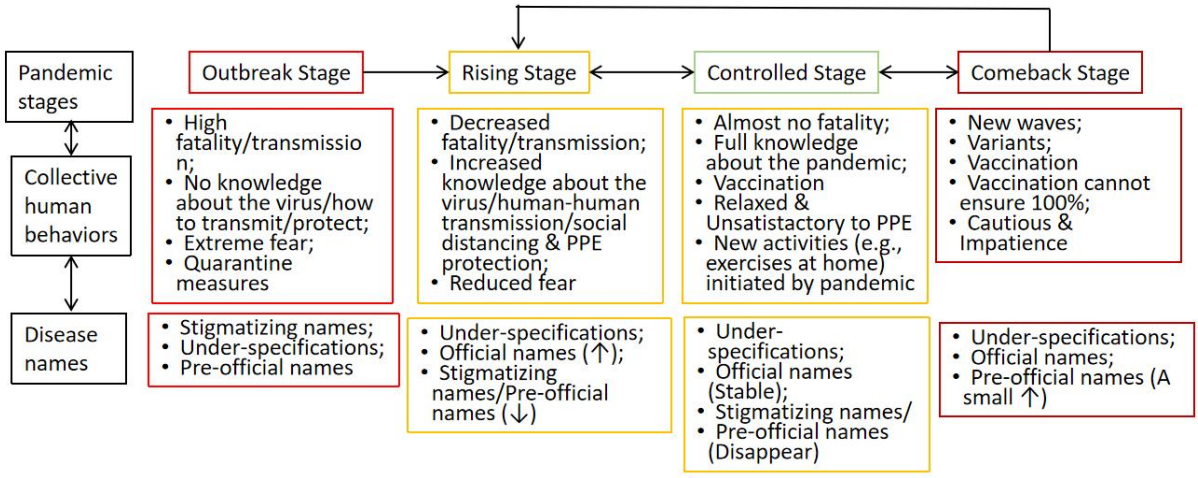

Figure 29 Interaction of pandemic stages, human behavior, and emergent neologisms

At the beginning of the disease, people were unclear about what the epidemic is and what it could do to them. They think they would not get sick if avoiding vectors such as snakes and bats. At that time, people assumed no human-to-human transmission, so this is a semi-denial stage where they only focused on what the disease might bring. Hence, at the beginning of the disease, people would like to use under-specifications (e.g., *yìqíng* 'pandemic') because they think under-specifications can to some extent weaken the fear to the ongoing pandemic because of the abstractness and they are easy to communicate because of the short length. Pre-official names as *bùmíng yuányīn fèiyán* 'pneumonia with unknown reasons' are also used to show their limited knowledge about the disease.

As the disease becomes more serious, people realized that avoiding the vectors is still difficult to avoid the virus. It is a disease that has human-to-human transmission. To effectively

control it, people want to know what causes the disease. But there might be no general panic till that time because pathogen is still quite abstract knowledge for the laypeople. The controlled stage quickly arrived when many scientists discovered rich information about the disease. People also realized that the epidemic is extremely severe and can transmit among human. But, at the same time, they were unable to acquire enough PPE for themselves or their family, which caused a general panic at this time So, stigmatizing names are primarily used to indicate the emotion of being fear of the place where the first case of the pandemic occurred. Pre-official names also convey people's fear and helplessness by highlighting the pneumonia for unknown reasons. Under-specifications are still widely used because such hedging words can help reduce the negative effect by the panic, which can reflect the small proportion of official names at that time even if authorities announced them. When it came to the come-back stage, a range of social and medical efforts have helped effectively control the spread of the epidemic, so more and more people used the official names when describing the disease apart from under-specifications.

Though the interaction framework proposed by the present thesis still needs to be confirmed by longer-period and more dynamic data of internet searches of emergent neologisms and collective human behavior, this framework that captures the important role of internet searches on the change of collective human behavior can shed light on the studies on emergency related communication. Prior studies have found that the important role of textual features (e.g., syntactic and discoursal levels) is functioned on health communication. For example, based on the quantitative and qualitative analysis of media discourse in Australian, Chinese, and Japanese documentary about environmental protection, Hook *et al.* (2017) revealed in rich and compelling detail the complex relationship between risk and responsibility in the climate change discourse. Ji *et al.* (2021) developed machine learning classifiers for health professionals with or without

Chinese proficiency to assess public-oriented health information in Chinese based on the definition of effective health communication by the WHO (relevance, credibility, understandability, actionability, accessibility). They found optimized SVM classifier achieved statistically significant higher ROC area. In their study, textual features in the effective health communication are considered to be independent variables, while the quality of health information are considered to be dependent variables. Ji *et al.* (2022) investigated the correlation between syntactic and structural features of written posts on health forums and the people with psychiatric disorders at risk of medication nonadherence. Using Bayesian machine learning techniques and publicly accessible online health forum data, their study illustrates the viability of developing cost-effective, informative decision aids to support the monitoring and prediction of patients at risk of medication nonadherence. Under the big data era, it will be interesting to investigate how internet searches play a role in health communication.

In addition, the framework can also provide reference to neologism related studies in general, beyond health emergency-related communication. The regression techniques based on training and testing data and the N-gram co-ocurrence analysis based on social media posts can be used to extend the studies on evolutional neologisms. Future work can also examine the comparison of the skewed S-curve and traditional S-curve on the development of neologisms.

6.2 Importance of Emergent Neologisms to Predict Pandemic

6.2.1 Better Predictability of Emergent Neologisms to Pandemic

As discussed above, the COVID-19 emergent neologisms has been considered to be a better indicator of reflecting, predicting, and monitoring the collective human behavior compared with buzzwords due to the high frequency, the short word length, the generic semantic meaning, the

closer association with the public emotion and social environment in the COVID-19 emergent neologisms. Relative to what the disease is caused (vector names) and how it can be avoided (PPE names), what the disease is (the COVID-19 emergent neologisms) is more urgent to be addressed and settled down. Hence, they should have bigger predictability theoretically than buzzwords to the pandemic development.

For the arrangement of the variants in the COVID-19 emergent neologisms in the regression modeling, we tried to feed the regression model with the single variant versus all variants, and linear, binomial, versus trinomial expressions. The modeling performance shows that binomial expressions based on all variants in the COVID-19 emergent neologisms are better arrangement of the independent variables by comparing R^2 and RMSE.

Each variant in the emergent neologisms has its respective focus. According to the percentage figure (Figure 11), under-specifications and pre-official names became the frequently used references at the end of December 2019 and early January 2020. However, when entering the middle of January in 2020, stigmatizing names exceeding under-specifications took the most significant proportion in the emergent neologisms till the middle of February. Within a short period, stigmatizing names were replaced by official names and under-specifications for almost one year. English abbreviations, which can be considered to be English version of Chinese official names, also substituted the proportion of stigmatizing names. The preference of different variants to refer to the disease at different times represents their different interpretability so that the different variants should be all reflected in the regression model.

Relative to linearity, non-linearity is more likely to conform to the commonality of the world. The special issue of *BioMed Research International* highlighted the complex systems of theoretical modeling, technical analysis, and numerical simulations in physics and mathematics

(Ji *et al.*, 2017). A complex system exhibit emerging properties due to the interaction within their subsystems when certain unspecific environmental conditions are satisfied. The system itself shows temporal and/or spatial patterns on a scale which is more prominent than the scale on which the subsystems interact (Fuchs, 2013, p. 3). Such complexity with self-adapted and self-interacted features is expressed non-linearly in the model. The pandemic event can be understood to be a complex system where the subsystems interact with each other. The subsystems of the pandemic can be all-inclusive such as the governmental policy enactment, the public emotion, or the use of emergent neologisms. From the perspective of the competition of gain-loss framing strategies, the use of different variants to refer to the disease at different periods contains information about who is the loss receiver. For example, the terms such as *wǔhàn bìngdú* 'Wuhan virus' show the other's loss, while the terms such as *xīn xíng guān zhuàng bìngdú fèiyán* 'COVID-19' show the self's loss. This indicates the complexity within different variants in the emergent neologisms. The arrangement of embedding all the variants in the regression models improves the predictability, conforming to the argument of the self-interacted complex system.

In addition, the proportional development of the emergent neologisms presents an immediate increase at the early stage and a gentle decline at the later stage, confirming the accuracy of binomial expressions. This finding verifies the reliability of the finding based on the first six-month pandemic data and internet searches in Lei *et al.* (2021).

6.2.2 The Good Fitting by Least Angle Regression

The Least Angle Regression is found to be better in modeling the mapping relationship between emergent neologisms and pandemic development based on the largest R^2 and the smallest RMSE.

Theoretically, compared with stepwise regression and regularization, the Least Angle Regression that brings the stepwise and regularization methods into full play should be unsurprisingly a better model in the mathematical mechanism. Practically, few prior studies investigated the mathematical relationship between neologisms and pandemic development. One recent study examining the mapping relationship is Lei *et al.* (2021) who found that the binomial expression is a better formula to model the relationship between emergent neologisms and pandemic cases over the six-month data since the pandemic outbreak. Our finding is also consistent with their finding. However, our result adjusted the hyperparameters by multiple methods (e.g., polynomial expressions, different arrangement of independent variables, and fine-tuned regression), which leads to a higher and more accurate model performance.

6.2.3 Inapplicability of S-curve on the COVID-19 Emergent Neologisms

The S-curve is not applied to the COVID-19 emergent neologisms. Though the development of emergent neologisms in our dataset shows fewer use and mentions at the early stage, which is similar to the innovation stage of the classical S-curve (Kroch, 1989; Chambers & Trudgill, 1992; Denison, 2002), the development of emergent neologisms differs from the traditional S-curve in two aspects.

On the one hand, the development of emergent neologisms witnessed the second stage as an exponential increase, while the traditional one has relatively gentle growth. On the other hand, the development of emergent neologisms at the final stage violates the finalization of new coinage in the replacement change by being distracted by the public. Hence, the development of emergent neologisms based on the fifteen-month pandemic data and internet searches generally followed the three-stage development in the skewed S-curve proposed by Lei *et al.* (2021).

Figure 30 illustrates the searches of the COVID-19 emergent neologisms on the Baidu Index over the fifteen months after the pandemic outbreak. At the emergence stage, it seems like the innovation stage of the traditional S-curve because it had a small number of percentages among the public to refer to the disease. Few people were used to employing it at the early stage of the pandemic. However, the Emergence Stage took a short time to move on to the next stage, i.e., the Diversification Stage, where it showcases a steeper increase compared to the second stage of the traditional S-curve. At this stage, people encountered a variety of word variants to name the disease because there has been no officially coined term for the disease. The public would select one term among the diverse variants according to the development of the pandemic. Unlike the time range between the second and third stage in the traditional S-curve, the skewed S-curve shows a shorter time to move from the Diversification Stage to the Distraction Stage. The traditional S-curve takes the most of the time in the Selection and Propagation, while the skewed S-curve reflects the uniqueness of terminologies for referring to the emerging event. Unlike the other important social events, such as political and economic events, the pandemic, which caused hundreds of people dead and sick, should promote the government and the relevant department to create appropriate nomenclatures to refer to the disease quickly. Hence, the red color representing the Distraction Stage in Figure 30 accounted for the most significant regional shade. The Distraction Stage is also distinct from the Fixation Stage in the traditional S-curve. In the S-curve, the Finalization Stage only presents how a new coinage entered the language system. However, the Distraction Stage in the skewed stage highlights the usage of the new variant by Baidu netizens. The epidemic event, especially the COVID-19 pandemic, experienced many new waves and would not end in a short period. Therefore, the Distraction Stage in the skewed S-curve reflects the successful entering of the new coinage in the language system and the change

of the public attention according to the change of disease. The Distraction Stage indicates more fluctuations in the following days compared with the traditional S-curve by extending the six-month to fifteen-month period.

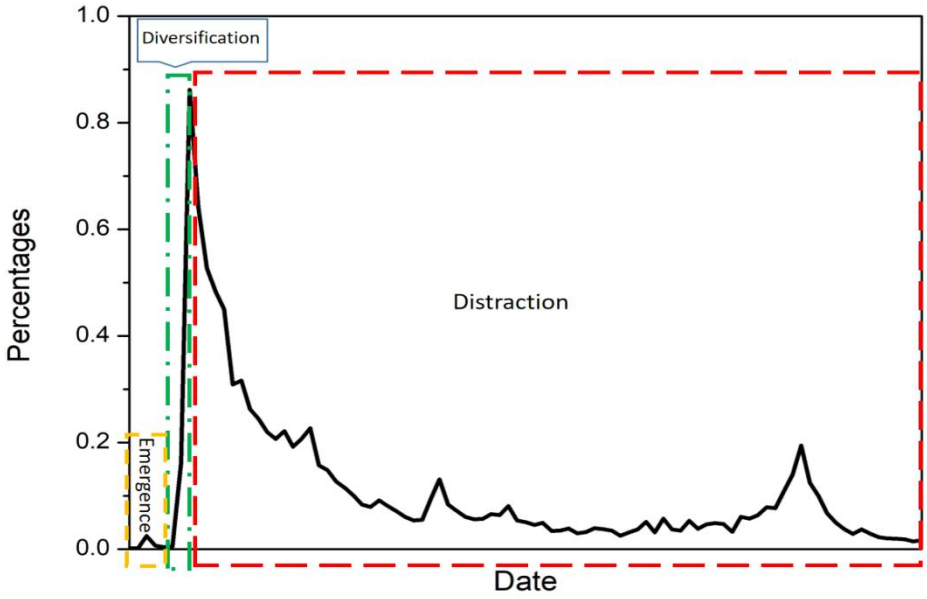

Figure 30 Skewed S-curve on emergent neologism of emergent neologisms

The skewed S-curve proposed by Lei *et al.* (2021) and validated by our thesis can be considered a visualization of the gain-loss framing strategies. As has been discussed above, the skewed S-curve reflects the change of the public attention during different periods of the pandemic. The most apparent stages showing the competition between gain- and loss-framing strategies lie in the Diversification Stage and Distraction Stage. At the Diversification Stage, the loss was considered the others, thereby leading to various variants used among the public. When the government had effectively controlled the pandemic, the phenomenon of using different variants disappeared and was replaced by more stable use and less attention among the public.

6.3 Monitoring Effect of Emergent Neologisms to Public Attention

6.3.1 Monitor of Psychological Change by N-gram Co-occurrence on Social Media

Based on the general use of emergent neologisms on the Baidu Index over the fifteen months after the pandemic outbreak, we hypothesized and verified the three-state psychological change: the state of being frightened, the state of being relaxed, and the state of being cautious.

These three states of the psychological change in the public can respond to our real experience from December 2019 to March 2021. The state of being frightened is set between the middle of January and early February 2020. This period is during the Chinese New Year. Conventionally, Chinese people would visit friends and relatives and gather together to have a Chinese New Year's Eve dinner. However, people found that they were disallowed to the gathering, visit, and travel at the most important traditional festival. They were restricted to staying at home and had to receive temperature tests at least twice per day by the community. The all-around suspension of entertainment activities has already caused a pandemic and negative attitudes among the public. On the other hand, the increase rate had more than 2 000 confirmed cases every day during the most painful period. Even many of our friends and relatives were found to be infected, exacerbating the panic. What's worse, while the panic spreads among people, the frequent change between misinformation (or fake news) and their refutation promotes the terror to an unexpected climax.

In February, the pandemic witnessed the adequate control by seeing the declining number of newly confirmed cases. At the same time, people were allowed to do entertainment activities while keeping appropriate social distances because the nature of the virus was discovered. People knew the whole process of how the pandemic emerges and how to protect themselves from being

infected, reducing the frightening emotion among the public. This is the second state, i.e., the state of being relaxed. Though the pandemic received effective control on a large scale, there were still new waves occurring from time to time under the overall relaxing state because of new variants of the virus, for example, local cases of single digits or foreign imported cases. A great deal of people who came back from overseas and experienced a 14-day quarantine were still confirmed as the COVID-19 pandemic patients, bringing the public's cautious emotions to the forefront.

Furthermore, the result on the psychological states during the pandemic are generally consistent with prior studies on the sentiment analysis by NLP approaches or survey results for the emotional change under the COVID-19 pandemic in Chinese. For example, Tan *et al.* (2021) investigated the influence of the pandemic on Chinese mainlanders' sentiment throughout 2020. They collected the posts from Sina Microblog automatically and analyzed sentiment based on Tecent NLP package. Their results showed that the pandemic caused long-term negative effects even during recovery. The long-term negative psychological states provide implications for the government and health departments. Zhao *et al.* (2020) also collected topics related to the COVID-19 pandemic from Sina Microblog from December 31st, 2019 to February 20th, 2020, and tracked how public attention changed to the COVID-19 related topics. They used ROST Content Mining System 6.0 to segment words, calculate word frequency, and analyze sentiment. Their findings differ from Tan *et al.*'s study (2021). In their result, the public emotional change did not center on long-term negative emotions. Rather, the negative emotion at the beginning stage changed to neutral feelings later. Based on questionnaires completed by Chinese urban residents, Shen *et al.* (2021) reported that public opinion could be heavily influenced by economic status, personal perception, and comprehension based on a mixed-effect logistic

regression model. Their survey result showed the public's big concern for the shortage of public health resources during the COVID-19 outbreak in China.

From the above, on the one hand, our assumption on psychological stages is generally correct. On the other hand, we also noticed the difference between polarity sentiment analysis based on NLP approaches and survey or content mining results. The limitation of the sentiment analysis based on the NLP approach is evident because it can only give the scores for negative, positive, or neutral tendency but cannot go deeper. For example, in Tan *et al.*'s research (2021), we only know that public emotion is generally negative during the pandemic. However, some deeper issues may not be answered by the sentiment polarity analysis, e.g., how the negative emotion among the public is specified and what specific reasons may lead to the long-term negative emotion. One recent work reflects the disadvantage of the sentiment polarity analysis. Zhao *et al.* (2020) enriched the result of the sentiment polarity analysis by adding social network analysis and visualization based on high-frequency keywords and their frequencies extracted from the Sina Microblog, as well as the topic trend analysis and sentiment analysis. Combined with the qualitative analysis based on more content mining, they found that the public emotion did not simply show long-term negative attitude during the pandemic; instead, the strong negative emotion only exists at the early stage of the pandemic. As the epidemic became under control, neutrality gradually replaced negative emotion.

Meanwhile, the sentiment polarity analysis and survey approaches may bring contradictory result. Tan *et al.* (2021) correlated the negative sentiment with newly confirmed cases at home and abroad and the post-pandemic economic recession but did not find any statistical significance. However, Shen *et al.*'s (2021) result showed that financial status, personal perception, and comprehension would all impact public opinion of the public health system

based on survey results. The above two contradictory results between NLP approaches and survey or content mining for analyzing public emotions demonstrate the necessity of going beyond the only sentiment analysis based on the limited sentiment polarity in tracking the change of the public emotion during the pandemic period. Consistent with such an appeal, our examination of the psychological states employs the emergence-based perspective of extracting the public emotions/attention from the Sina Microblogging posts at different periods. With the phrase *dài kǒuzhào* 'wear masks' as a searching phrase of the Sina Microblog website, we selected eight important dates and analyzed the co-occurring N-grams of *dài kǒuzhào* 'wear masks' in the crawled posts. Through the co-occurring frequencies with *dài kǒuzhào* 'wear masks', we deduced the change of the public attention. At the same time, the N-gram segmented texts provide richer information such as WHEN, HOW, WHO, and WHERE, thereby offering a more comprehensive picture of the change of public emotion. For example, when the phrases around *dài kǒuzhào* 'wear masks' are *kōngqì bù xīnxiān* air not fresh 'unfresh air', *bù xiǎng bèi kàndào* no want BEI see 'don't want to be seen', *dōngtiān shìhé* winter fit 'fit for winter' on December, 2019, the public's focus was on the freshness of the air and the fitness for wearing face masks in winter, as well as the superstars' frequent behavior at the airport. In contrast, when the phrases around *dài kǒuzhào* 'wear masks' changed to *chūmén* 'go outside', *qín xǐshǒu* 'wash hands regularly', *jìdé* 'remember', *yīdìng* 'must' on January, 2020, the public's focus switched quickly to the measures for protecting themselves from the pandemic. When the phrases around *dài kǒuzhào* 'wear masks' changed to *tiān tài rè* 'the weather is too hot' and *dài kǒuzhào bù shūfu* 'it is uncomfortable to wear masks' on August 2020, the public's focus is showing their negative mood of mask wearing under the high temperature. Though the public showed negative tendency at these three dates, the negative attitude differs with the negative attitude on the

unfresh air on December 2019, the fear on the emerging and fatal pandemic on January 2020, and the uncomfortable feeling of mask wearing under the hot atmosphere.

6.3.2 Inapplicability of Issue-Attention Cycle on the Public Attention during COVID-19

The other important theoretical issue that will be discussed is whether the classical Issue-Attention Cycle can be applied to psychological change under the COVID-19 pandemic. First of all, let us recall the five major stages in this classical cycle: 1) Only experts or a small number of people are aware of the issue. 2) The issue captures more attention among the public as the issue becomes more serious. However, people are optimistic that the problems will be solved at this stage. 3) People realize the issues might be far different from their expectations and out of their control. They know that the issue cannot end soon and the control of the issue will cost high financial or social benefits. 4) A steady drop occurs in public attention at the post-problem phrase. 5) The other issues coming into the public attention replaces the current issue, finally.

We noticed that the first-round year from the end of December 2019 to August 2020 shows the consistency with the classical cycle. That is, on December 21st, 2019, there has been no public realizing the pandemic, which is reflected in no mentions of the pandemic in Sina Microblog posts. However, the public attention to the pandemic increased noticeably since January 25th, 2020. Such a high degree of attention among the public has been lasting until the time after May Day in China. On May 4th, 2020 when it is the first day after the May Day holiday, the public expressed unsatisfactory emotion about wearing the mask when they are in the outdoor environment. Such unsatisfactory attitude became stronger on August 2020 when the summer holiday came. For one thing, the public who experienced a long time of quarantine at home for almost three months looked forward to traveling but was not encouraged. For another,

they were still required to wear masks even when running, which caused much discomfort. The change of the public attention was summarized as below in the first-round circle.

Little attention (2019.12.21) -- Sudden increase of attention (2020.1.25) --

Unsatisfaction (2020.5.4) -- Unsatisfaction and Distraction (2020.8.7)

However, as some researchers argued, the Issue-Attention Cycle could differ depending upon culture (Brossard, Shanahan, & McComas, 2004) and in cases of epidemic hazards (Shih *et al.*, 2008). Moreover, the classical Issue-Attention Cycle is not always fully integrated or explanatory in some health-related research, as evidenced by the "Charlie Sheen effect" phenomenon. Ayers *et al*. (2016) used results from Google's search engine data set to show the correlation between actor Charlie Sheen's disclosure of his HIV-positive status with the level of public attention to HIV and its prevention.

The second-round posting data showed the inapplicability of the classical Issue-Attention Cycle in explaining the change of public attention during different periods of the COVID-19 pandemic. Although the classical cycle announces that the other issue would finally replace the public attention, our Sina Microblog data at the end of 2020 does not show that the public attention followed such a pattern during the COVID-19 pandemic. Rather, public still had heated discussion on the Sina Microblog, wishing the pandemic will end soon. The second justification for confirming the inapplicability of the Issue-Attention Cycle to the change of public attention in the COVID-19 pandemic lies in the second-round posting data. Though January, 2021 demonstrated the peak of discussion that the public attention reached, May and August 2021 did not show unsatisfactory attitude as expected by the Issue-Attention Cycle. Comparatively, the public's awareness of self-protection and mask wearing after getting injection showed noticeable improvement. Some news says there is still a possibility of getting infected after being injected

with several doses when the public realizes that, even if the pandemic has experienced more than one year. It is uncomfortable when mask wearing in the deep summer, though mask-wearing is a necessary behavior. This may be why the awareness for self-protection replaced unsatisfactory attitudes. The following line is given to show the second-round change of public attention.

> Attention remaining (2020.12.31) -- Another peak of attention (2021.1.4) -- Awareness protection (2021.5.29) -- Awareness retaining and dissatisfaction (2021.8.4)

Under such a long-lasting disease with many different variants, the change in public attention presents a more complex pattern. Far from the argument that the classical Issue-Attention Cycle focuses on the linearly developmental pattern of the extent of attention, our thesis based on Sina Microblog posts found that more specific emotion can elaborate on the attention. Such finding is consistent with the characteristics of non-linearity in regression results and interchangeability of gain and loss framing strategies at different periods.

6.4 Chapter Summary

This chapter discusses the exciting findings in Chapter Four and Chapter Five. The qualitative analysis, regression modeling, and N-gram co-occurrence showed the COVID-19 emergent neologisms are better indicators to reflect important policy announcement, predict pandemic cases, and the change of public attention better than buzzwords including PPE names and vector names. Hence, first, we explained the advantage of the COVID-19 emergent neologisms in terms of the short word length, the high frequency, the generic semantic meaning, the closer association with the public emotion by gain and loss framing strategies, and social events. In the second part of the discussion, we explained why the regression modeling showed the emergent neologisms have better predictability than buzzwords and why the Least Angle Regression is a

better model to predict the pandemic development by the binomial expressions of all the variants in the emergent neologisms. We explained the better predictability of the COVID-19 emergent neologisms by gain and loss framing strategies and the core of emergent neologisms in the pandemic event. The better performance of the Least Angle Regression model is explained by its combining characteristics of regularization and stepwise methods. Meanwhile, the inapplicability to the classical S-curve was also discussed by our result based on the Sina Microblog data supporting a skewed S-curve proposed by Lei *et al.* (2021). The third part of discussion explains the advantage of using N-gram co-occurrence based on social media data compared with the sentiment polarity analysis based on NLP approaches in emotion extraction and responds to the inapplicability of the classical Issue-Attention Cycle to how the public attention changes under the COVID-19 pandemic.

CHAPTER SEVEN CONCLUSION

The present thesis presents a systematic and comprehensive study of how emergent neologisms interact with the collective human behavior (i.e., important policies, pandemic cases, and the change of public emotion). The overall results demonstrated the reflective, predictive, and monitoring roles of emergent neologisms in the change of collective human behavior. Specifically,

- Qualitative results: Emergent neologisms can reflect the pandemic development, relevant social events, and important policy announcement.
- Quantitative results of the correlation and regression modeling: Emergent neologisms are strongly correlated with pandemic cases and can predict the development of pandemic.
    - The Least Angle Regression is a better model to describe the mathematical relationship between emergent neologisms and pandemic cases.
    - Compared with buzzwords including vector names and PPE names, emergent neologisms following the Framing Effect under the Prospect Theory play a more critical role in the prediction of the pandemic cases.
    - Unlike the traditional S-curve, the emergent neologisms develop with binomial expressions by increasing exponentially and decreasing gently over the fifteen months after the pandemic outbreak.
- Quantitative results of N-gram co-occurrence: Emergent neologisms can monitor public emotion.

- The N-gram co-occurrence calculation based on Sina Microblog posts by the searching phrase *dài kǒuzhào* ‘wear masks’ confirms the acceptability by using emergent neologisms to monitor the change of the public emotion.
- The N-gram co-occurrence provides more contextual information than sentiment polarity analysis in NLP approaches and survey results.
- The classical Issue-Attention Cycle is not applied to the second-round-year data on Sina Microblog.

Based on the above summary of the major findings, the present thesis further proposed an interaction framework by connecting emergent neologisms and collective human behavior to demonstrate the critical role of emergent neologisms in responding to the collective human activities. The results confirmed the interpretability of the gain-and-loss theory on the relationship between neologisms and collective human behavior during the pandemic period.

There remain some directions for future studies. Firstly, future research can further verify the important role of emergent neologisms in other important and influential social events through the mixed research design of qualitative and quantitative methods used in the present thesis. Secondly, the critical theoretical issues in this thesis can be re-examined based on the data of other domains, such as the generalizability of the Framing Effect under the Prospect Theory on the public psychological states at other emerging social events, the generalizability of skewed S-curve on neologism development under other social issues, and the generalizability of the issue-attention cycle on important events. Thirdly, future research can investigate the predictability of emergent neologisms with binomial expressions on the pandemic cases based on the Least Angle Regression through longer-term internet searches and pandemic cases. Fourthly, the N-gram co-occurrence calculation on Sina Microblog posts can also be applied in future

sentiment analysis studies, thereby providing more specific information about how the public feels in addition to the result of simplified polarity sentiment.

Though the present thesis used a limited time period in the quantitative analysis of the regressions and the limited search phrase in the quantitative analysis of the N-gram co-occurrence, emergent neologisms still show their significant potential in interacting with collective human behavior.

Overall, the present thesis has twofold contributions:

- Theoretically, it proposes an interaction framework that comprehensively and systematically describes how emergent neologisms and collective human behavior interact.
- Practically, it highlights the vital role of emergent neologisms in reflecting the important policy announcements, predicting pandemic cases, and monitoring public emotional change.
    - At least at the early stage of a newly emerging pandemic event, emergent neologisms can be a good indicator in the pandemic recording system and public emotion recorders.
    - Government and the relevant department can consider internet searches as critical sources for grasping public emotional changes during different pandemic periods.

References

Ahmad, A. R., & Murad, H. R. (2020). The impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan: Online questionnaire study. *Journal Medical Internet Research*, *22*(5), e19556. DOI: 10.2196/19556

Arbab, M., Shen, M., Mok, B., Wilson, C., Matuszek, Ż., Cassa, C., et al. (2021). Determinants of base editing outcomes from target library analysis and machine learning. *Cell*.

Arendt, F., & Scherr, S. (2019). Investigating an issue-attention-action cycle: A case study on the chronology of media attention, public attention, and actual vaccination behavior during the 2019 Measles outbreak in Austria. *Journal of Health Communication, 24*(7-8), 654-662.

Artenstein, A. W. (2020). In Pursuit of PPE. *The New England Journal of Medicine*, e46(1)-e46(2).

Asif, M., Zhiyong, D., Iram, A., & Nisar, M. (2021). Linguistic analysis of neologism related to coronavirus (COVID-19). *Social sciences & humanities open, 4*(1), 100201. DOI: https://doi.org/10.1016/j.ssaho.2021.100201

Ayers, J. W., Althouse, B. M., Dredze, M., Leas, E. C., & Noar, S. M. (2016). News and internet searches about human immunodeficiency virus after Charlie Sheen's disclosure. *JAMA Internal Medicine, 176*(4), 552-554. DOI:10.1001/jamainternmed.2016.0003

Bailey, C. -J. N. (1973). *Variation and linguistic theory*. Washington, D.C.: Center for Applied Linguistics.

Baker, A. (2008). Addressing the actuation problem with quantitative models of sound change. *University of Pennsylvania Working Papers in Linguistics,* 14(1), Article 3. Available at: https://repository.upenn.edu/pwpl/vol14/iss1/3

Barry, C., McGinty, E., & Sherman, S. (2018). Language matters in combatting the opioid epidemic: safe consumption sites versus overdose prevention sites. *AJPH Perspect, 108*(9), 1157-1159.

Blythe, R., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language, 88*(2), 269-304.

Bogard, N., Linder, J., Rosenberg, A., & Seelig, G. (2019). A deep neural network for predicting and engineering alternative polyadenylation. *Cell, 1*, 91-106.

Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., et al. (2020). The psychological impact of quarantine and how to reduce it: Rapid review of the evidence. *The Lancet, 395*(10227), 912-920. DOI: 10.1016/s0140-6736(20)30460-8

Brossard, D., Shanahan, J., & McComas, K. (2004). Are issue-cycles culturally constructed? A comparison of French and American coverage of global climate change. *Mass communication & society, 7*(3), 359-377.

Brumercikova, E., & Bukova, B. (2020). The regression and correlation analysis of carried persons by means of public passenger transport of the Slovak Republic. *Transportation Research Procedia, 44*(4), 61-68. DOI: 10.1016/j.trpro.2020.02.010

Callaway, E., Cyranoski, D., Mallapaty, S., Stoye, E., & Tollefson, J. (2020). The coronavirus pandemic in five powerful charts. *Nature, 579*, 482-483 (2020). DOI: https://doi.org/10.1038/d41586-020-00758-2

Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., & Zheng, J. (2020). The psychological impact of the COVID-19 epidemic on college students in China. *Psychiatry Research, 287*, 112934.

Chambers, J.K., & Trudgill, P. (1992). *Dialectology, second edition*. Cambridge University Press, Cambridge.

Chen, P. (1999). *Modern Chinese: History and sociolinguistics*. Cambridge & New York: Cambridge University Press.

Chen, Y., Zhang, Y., Xu, Z., Wang, X., Lu, J., & Hu, W. (2019). Avian Infuenza A (H7N9) and related Internet search query data in China. *Scientific Reports, 9*, 10434. DOI: https://doi.org/10.1038/s41598-019-46898-y

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS ONE, 5*(11), 1-18.

Chou, L., & Hsieh, S. (2013). Qualia modification in Mandarin neologism: A case study on prefix "wéi 微". In: Liu, P., & Su, Q. (Eds.) *Chinese lexical semantics. CLSW 2013. Lecture Notes in Computer Science, volume 8229.* Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-45185-0_32

Colquhoun, D. (2018). The false positive risk: A proposal concerning what to do about *p* values. *The American Statistician, 73*, 192-201.

Corbett, K. S., Flynn, B., Foulds, K. E., Francica, J. R., Boyoglu-Barnum, S., Werner, A. P. et al. (2020). Evaluation of the mRNA-1273 Vaccine against SARS-CoV-2 in Nonhuman Primates. *The New England Journal of Medicine, 383*(16), 1544-1555. DOI: https://doi.org/10.1056/NEJMoa2024671

David, N. (2014). Chapter 6 - Selection of Variables and Factor Derivation. In: Kaufmann, M. (Ed.), *Commercial data mining: Processing, analysis and modeling for predictive analytics projects*, pp. 79-104. DOI: https://doi.org/10.1016/B978-0-12-416602-8.00006-6

Denison, D. (2002). Log(ist)ic and simplistic S-curves. In: Hickey, R. (Ed.), *Motives for language change,* pp. 54-70. Cambridge University Press, Cambridge.

Dinnon, K. H., Leist, S. R., Schäfer, A., Edwards, C. E., Martinez, D. R. et al. (2020). A mouse-adapted model of SARS-CoV-2 to test COVID-19 countermeasures. *Nature, 586*(7830), 560-566. DOI: 10.1038/s41586-020-2708-8.

Downs, A. (1996). Up and down with ecology: The "Issue-Attention Cycle". In: Park, M. (Ed.) *The politics of American economic policy making*. Oxfordshire: Taylor & Francis.

Dong, S., Huang, C.-R., & Ren, H. (2020). Towards a new typology of meteorological events: a study based on synchronic and diachronic data. *Lingua, 247*, 102894.

Dras, M., & Harrison, D.K. (2003). Emergent behavior in phonological pattern change. In: Standish, A., Bedau, M. A., & Abbas, H. A. (Eds.), *Artificial Life VIII, Proceedings of the 8th International Conference on Artificial Life*, pp. 390-393.

Druckman, J. N. (2001a). Using credible advice to overcome framing effects. *The Journal of Law, Economics, and Organization, 17*(1), 62-82. DOI: https://doi.org/10.1093/jleo/17.1.62

Druckman, J. N. (2001b). Evaluating framing effects. *Journal of Economic Psychology, 22*, 91-101.

Eisenstein, J., O'Connor, B., Smith, N., & Xing, E. (2014). Diffusion of lexical change in social media. *PLoS ONE, 9*(11), e113114.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407-451.

Efroymson, M. A. (1960). Multiple regression analysis. In: Ralston, A., & Wilf, H. S. (Eds.), *Mathematical Methods for Digital Computers.* New York: John Wiley.

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly, 40*, 83-107.

Ellis, N., & Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning, 59*(s1), 90-125. DOI: 10.1111/j.1467-9922.2009.00537.x

Ellis, N., & Collins, L. (2009). Input and second language acquisition: The roles of frequency, form, and function. *The Modern Language Journal, 93*(3), 329-336.

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua, 120*, 2061-2079.

Fung, I. C. -H., Fu, K. -W., Chan, C. -H. et al. (2016). Social media's initial reaction to information and misinformation on Ebola, August 2014: Facts and rumors. *Public Health Reports, 131*(3), 461-473. DOI:10.1177/003335491613100312

Fuchs, A. (2013). *Nonlinear dynamics in complex systems*. Berlin: Springer-Verlag.

Gächter, S., Orzen, H., Renner, E., & Stamer, C. (2009). Are experimental economists prone to framing effects? A natural field experiment. *Journal of Economic Behavior & Organization, 70*(3): 443-446.

Gantiva, C., Jiménez-Leal, W., & Urriago-Rayo, J. (2021). Framing messages to deal with the COVID-19 crisis: The role of loss/gain frames and content. *Frontiers in Psychology, 12,* 568212. DOI: 10.3389/fpsyg.2021.568212

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8-11. DOI: 10.1016/j.tics.2003.10.016

Gesser-Edelsburg, A., Shir-Raz, Y., Bar-Lev, O., James, J., & Green, M. (2016). Outbreak or epidemic? How Obama's language choice transformed the Ebola outbreak into an epidemic. *Disaster Med. Public Health Preparedness, 10*, 669-673.

Ghebreyesus, T. (2020). WHO Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020

Gove, P. B. (1993). *Webster's Third International Dictionary of the English Language Unabridged*. Cologne, Konemann.

Green, N. (2012). Correlation is not causation. *The Guardian*. https://www.theguardian.com/science/blog/2012/jan/06/correlation-causation

Gu, H., Chen, B., Zhu, H., Jiang, T., Wang, X., Chen, L. et al. (2014). Importance of internet surveillance in public health emergency control and prevention: Evidence from a digital epidemiologic study during Avian Influenza A H7N9 outbreaks. *Journal of Medical Internet Research, 16*(1), e20. DOI: 10.2196/jmir.2911

Gui, X., Wang, Y., Kou, Y., Reynolds, R., Chen, Y., Mei, Q. et al. (2017). Understanding the patterns of health information dissemination on social media during the Zika outbreak. Conference: AMIA 2017 Annual Symposium.

Harrison, K.D., Dras, M., & Kapicioglu, B. (2002). Agent-based modeling of the evolution of vowel harmony. *Proceedings of the Northeast Linguistic Society (NELS),* 32.

Hawryluck, L., Gold, W. L., Robinson, S., Pogorski, S., Galea, S., & Styra, R. (2004). SARS control and psychological effects of quarantine, Toronto, Canada. *Emerging Infectious Disease,10*(7),1206-1212. DOI: 10.3201/eid1007.030703

Holubnycha, L., Kostikova, I., Besarab, T., Moshtagh, Y., Lushchyk, Y., et al. (2020). Semantic and structural aspects of Donald Trump's neologisms. *Postmodern Open, 11*(2), 43-59.

Hook, G., Lester, L., Ji, M., Edney, K., Pope, C., & van der Does-Ishikawa, L. (2017). *Environmental Pollution and the Media: Political Discourses of Risk and Responsibility in Australia, China and Japan*. Abingdon: Routledge.

Househ, M. (2016). Communicating Ebola through social media and electronic news media outlets: A cross-sectional study. *Health Informatics Journal, 22*(3), 470-478. DOI: 10.1177/1460458214568037

Hoyt, R. E., Snider, D., Thompson, C., & Mantravadi, S. (2016). IBM Watson analytics: Automating visualization, descriptive, and predictive statistics. *JMIR Public Health & Surveillance, 2*(2), e157. DOI: 10.2196/publichealth.5810

Huang, C.-R., & Hsieh, S.K. (2015). Chinese lexical semantics. In: Wang, W.S.-Y., & Sun, C. (Eds.), *The Oxford Handbook of Chinese Linguistics,* pp. 290-305. Oxford University Press, (Chapter 22).

Huynh, T. L. D. (2020). The COVID-19 risk perception: A survey on socioeconomics and media attention. *Economics Bull, 40*(1): A.

IBM Cloud Education. (2021). Overfitting. https://www.ibm.com/cloud/learn/overfitting

Janssen, I., Hendriks, F., & Jucks, R. (2021). Face masks might protect you from COVID-19: The communication of scientific uncertainty by scientists versus politicians in the context of policy in the making. *Journal of Language and Social Psychology, 40*(5-6), 602-626. DOI:10.1177/0261927X211044512

Ji, Z., Yan, K., Li, W., Hu, H., & Zhu, X. (2017). Mathematical and Computational Modeling in Complex Biological Systems. *BioMed Research International, 3*, 1-16. DOI: https://doi.org/10.1155/2017/5958321

Ji, M., Bodomo, A., Xie, W., & Huang, R. (2021). Assessing communicative effectiveness of public health information in Chinese: Developing automatic decision aids for international health professionals. *International Journal of Environmental Research and Public Health, 18*(10329), 1-11.

Ji, M., Xie, W., Zhao, M., Qian, X., Chow, C., Lam, K., Yan, J., & Hao, T. (2022). Probabilistic prediction of nonadherence to psychiatric disorder medication from mental health forum data: Developing and validating Bayesian machine learning classifiers. *Computational Intelligence and Neuroscience*, 6722321.

Jiang, M., Shen, X. Y., Ahrens, K., & Huang, C.-R. (2021). Neologisms are epidemic: Modeling the life cycle of neologisms in China 2008-2016. *PLOS One, 16*(2), e0245984.

Jing-Schmidt, Z., & Hsieh, S.-K. (2019). Chinese neologisms. In: Huang, C.-R., Jing-Schmid, Z., & Meisterernst, B. (Eds.), *The Routledge Handbook of Chinese Applied Linguistics,* pp. 514-534. Routledge.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-291. DOI:10.2307/1914185.

Ke, J., Gong, T., & Wang, W. S. Y. (2008). Language change and social networks. *Communications in Computational Physics, 3*(4), 935-949.

Kim, H. K., & Tandoc, E. C. Jr. (2022). Consequences of online misinformation on COVID-19: Two potential pathways and disparity by eHealth literacy. *Frontiers in Psychology, 13*, 783909. DOI: 10.3389/fpsyg.2022.783909

Klosa-Kückelhaus, A., & Wolfer, S. (2020). Considerations on the acceptance of German neologisms from the 1990s. *International Journal of Lexicography, 33*(2), 150-167.

Kogan, N. E., Clemente, L., Liautaud, P., Kaashoek, J., Link, N. B., Nguyen, A. T. et al. (2021). An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. *Science Advances, 7*, eabd6989.

Kroch, A. S. (1989). Function and grammar in the history of English: Periphrastic 'do'. In: Fasold, R. W., & Schiffrin, D. (Eds.), *Language Variation and Change: Current Issues in Linguistic Theory, volume. 52,* pp.133-172. Philadelphia: John Benjamins.

Labov, W. (1966). The effect of social mobility on linguistic behavior. *Sociological Inquiry, 36*(2), 186-203.

Labov, W. (1994). *Principles of Linguistic Change, Volume 1: Internal Factors*. Blackwell.

Labov, W. (2001). *Principles of Linguistic Change, Volume 2: Social Factors*. Blackwell.

Lam, T. T. Y., Jia, N., Zhang, Y. W. et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature, 583*, 282-285. DOI: https://doi.org/10.1038/s41586-020-2169-0

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics, 18*(2). DOI: 10.1093/applin/18.2.141

Latinne, A., Hu, B., Olival, K. J., Zhu, G., Zhang, L., Li, H. et al. (2020). Origin and cross-species transmission of bat coronaviruses in China. *Nature Communications, 11*, 4235.

Laurini, M. P. (2018). A spatial-temporal approach to estimate patterns of climate change. *Environmetrics, 30*, e2542. DOI: https://doi.org/10.1002/env.2542

Lei, S., Yang, R., & Huang, C.-R. (2021). Emergent neologism: A study of an emerging meaning with competing forms based on the first six months of COVID-19. *Lingua, 258*, 103095.

Li, L., Huang, C.-R., & Wang, V. X. (2020). Lexical competition and change: A corpus-assisted investigation of gambling and gaming in the past centuries. *Sage Open, 10*(3), 1-14.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(18), 1-45. DOI: https://dx.doi.org/10.3390/e23010018

Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., et al. (2020). Real-time forecasting of the COVID-19 outbreak in Chinese provinces: Machine learning approach using novel digital data and estimates from mechanistic models. *Journal of Medical Internet Research, 22*(8), e20285.

Liu, W., & Liu, W. (2014). Analysis on the word-formation of English netspeak neologism. *Journal of Arts & Humanities, 3*(12), 22-30.

Mak, I. W. C., Chu, C. M., Pan, P. C., Yiu, M. G. C., & Chan, V. L. (2009). Long-term psychiatric morbidities among SARS survivors. *General Hospital Psychiatry, 31*(4), 318-326. DOI: 10.1016/j.genhosppsych.2009.03.001

Mallapaty, S. (2021). Closest known relatives of virus behind COVID-19 found in Laos. *Nature, 597*, 603. DOI: https://doi.org/10.1038/d41586-021-02596-2

Millar, N. (2009). Modal verbs in TIME: Frequency changes 1923-2006. *International Journal of Corpus Linguistics, 14*, 191-220.

Mishra, M. (2018). REGULARIZATION: An important concept in Machine Learning. Towards Data Science. https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea

Mousavizadeh, L., & Ghasemi, S. (2021). Genotype and phenotype of COVID-19: Their roles in pathogenesis. *Journal of Microbiology, Immunology and Infection, 54*(2), 159-163.

Occupational Safety and Health Council. (2021). Personal Protective Equipment (PPE). https://www.oshc.org.hk/eng/main/hot/ppe/#:~:text=Personal%20protective%20equipment%20(PPE)%20refers,%2Fher%20safety%20or%20health%22.

Oxford Dictionary. (2003). https://www.theguardian.com/books/2020/apr/15/oxford-dictionary-revised-to-record-linguistic-impact-of-covid-19

Paun, V. -P. (2021). New Perspectives in Nonlinear Dynamics of Complex Systems 2021. https://www.hindawi.com/journals/complexity/si/169896/

Peng, L., Guo, Y., & Hu, D. (2021). Information framing effect on public's intention to receive the COVID-19 vaccination in China. *Vaccines, 9*, 995.

Peng, L., Jiang, H., Guo, Y., & Hu, D. (2022). Effect of information framing on wearing masks during the COVID-19 pandemic: Interaction with social norms and information credibility. *Frontiers in Public Health, 10*, 811792. DOI: 10.3389/fpubh.2022.811792

Petersen, K. K. (2009). Revisiting Downs' issue-attention cycle: International terrorism and US public opinion. *Journal of strategic security, 2*(4), 1-16.

Plag, I. (2002). The role of selectional restrictions, phonotactics and parsing in constraining suffix ordering in English. In: Booij Geert E. & van Marle Jaap (Eds.), *Yearbook of Morphology*, pp. 285-314. Dordrecht, Boston & London, Kluwer Academic Publishers.

Plous, S. (1993). The psychology of judgement and decision making (McGraw-Hill Series in Social Psychology) 1st Edition. McGraw-Hill.

Regan, H., Griffiths, J., Culver, D., & Guy, J. (2020). Wuhan coronavirus virus spreads as China scraps New Year celebrations. CNN, Updated 0152 GMT (0952 HKT) January 24. https://edition.cnn.com/2020/01/23/china/wuhan-coronavirus-update-intl-hnk/index.html

Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). 2020. https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19)

Sakti, A. M. T., Mohamad, E., & Azlan, A. A. (2021). Mining of opinions on COVID-19 large-scale social restrictions in Indonesia: Public sentiment and emotion analysis on online media. *Journal of Medical Internet Research, 23*(8), e28249. DOI: 10.2196/28249

Santosa,F., & Symes, W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing, 7*(4), 1307-1330. DOI: http://dx.doi.org/10.1137/0907087

Seltzer, E. K., Jean, N. S., Kramer-Golinkoff, E., Asch, D. A., & Merchant, R. M. (2015). The content of social media's shared images about Ebola: A retrospective study. *Public Health, 129*, 1273-1277.

Seltzer, E. K., Horst-Martz, E., Lu, M., & Merchant, R. M. (2017). Public sentiment and discourse about Zika virus on Instagram. *Public Health, 150*,170-175. DOI: 10.1016/j.puhe.2017.07.015. Epub 2017 Aug 12. PMID: 28806618.

Shen, X., Li, J., Dong, T., Cao, H., Feng, J., Lei, Z. et al. (2021). Public opinion and expectations: Development of public health education in China after COVID-19 pandemic. *Frontiers in Public Health, 9*, 702146. DOI: 10.3389/fpubh.2021.702146

Shih, T. J., Wijaya, R., & Brossard, D. (2008). Media coverage of public health epidemics: Linking framing and issue attention cycle toward an integrated theory of print news coverage of epidemics. *Mass Communication & Society, 11*(2), 141-160.

Shou, S., Liu, M., Yang, Y., Kang, N., Song, Y., Tan, D. et al. (2021). Animal models for COVID-19: Hamsters, mouse, ferret, mink, tree shrew, and non-human primates. *Frontiers in Microbiology, 12*, 626553. DOI: 10.3389/fmicb.2021.626553

South African Government. (2021). President Cyril Ramaphosa: South Africa's response to Coronavirus COVID-19 pandemic. https://www.gov.za/speeches/president-cyril-ramaphosa-south-africas-response-coronavirus-covid-19-pandemic-27-jun-2021

Steffen, J., & Cheng, J. (2021). The influence of gain-loss framing and its interaction with political ideology on social distancing and mask wearing compliance during the COVID-19 pandemic. *Current Psychology.* DOI: https://doi.org/10.1007/s12144-021-02148-x

Steyerberg, E. W. (2009). Clinical prediction models: A practical approach to development, validation, and updating (Statistics for biology and health), pp. 11-31. Springer. DOI: 10.1007/978-0-387-77244-8

Tan, H., Peng, S. -L., Zhu, C. -P., You, Z., Miao, M. -C., & Kuai, S. -G. (2021). Long-term effects of the COVID-19 pandemic on public sentiments in mainland China: Sentiment analysis of social media posts. *Journal of Medical Internet Research, 23*(8), e29150). DOI: 10.2196/29150.

Thompson, R. R., Garfin, D. R., Holman, E. A., & Silver, R. C. (2017). Distress, worry, and functioning following a global health crisis: A national study of Americans' responses to Ebola. *Clinical Psychological Science, 5*(3):513-521. DOI: 10.1177/2167702617692030

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological), 58*(1), 267-288.

Tran, H. T. T., Lu, S. H., Tran, H. T. T. , & Nguyen, B. V. (2021). Social media insights during the COVID-19 pandemic: Infodemiology study using big data. *JMIR Medical Informatics, 9*(7), e27116. DOI: 10.2196/27116.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458. DOI: 10.1126/science.7455683

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-323.

van Eck, N.J., & Waltman, L. (2007). Bibliometric mapping of the computational intelligence field. International Journal of Uncertainty. *Fuzziness and KnowledgeBased Systems, 15*(5), 625-645.

van Eck, N. J., & Waltman, L. (2021). VOSviewer Manual. https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.17.pdf

van Eck, N. J., Waltman, L., Van den Berg, J., & Kaymak, U. (2006). Visualizing the computational intelligence field. *IEEE Computational Intelligence Magazine, 1*(4), 6-10.

Wang, W. -H., Urbina, A. N., Wu, C.-C., Lin, C.-Y., Arunee, T., Assavalapskul, W. et al. (2020). An epidemiological survey of the current status of Zika and the immune interaction between dengue and Zika infection in Southern Taiwan. *International Journal of Infectious Diseases, 93*, 151-159.

Wang, W. S. -Y., Ke, J. -Y., & Minett, J. W. (2004). Computational studies of language evolution. In: Huang, C. -R., & Lenders, W. (Eds.), *Computational linguistics and beyond* (pp. 65-108). Taipei: Institute of Linguistics, Academia Sinica.

World Health Orgnization (WHO). (2020). Vector-borne diseases. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G. et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature, 579*, 265-269 DOI: https://doi.org/10.1038/s41586-020-2008-3

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet, 395*(10225), 676. DOI: 10.1016/S0140-6736(20)30461-X

Zhao, Y., Cheng, S., Yu, X., & Xu, H. (2020). Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study. *Journal of Medical Internet Research, 22*(5), e18825

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W. et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature, 579*, 270-273.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 67*(2), 301-320.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418-1429. DOI: 10.1198/016214506000000735

Appendices

Appendix A. Scripts of Pearson Correlation and Regressions by using newly confirmed cases as a dependent variable and official names with the first degree as a predictor for examples

- **Pearson correlation**

```
proc corr data = WORK.NEOPAN;
var Official_names New_confirmed;
title 'Examination of Correlation Matrix';
run;
```

- **Linear regression**

```
proc reg data=WORK.NEOPAN alpha=0.05 plots(only)=(diagnostics residuals fitplot
                observedbypredicted);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names /;
        run;
quit;
```

- **Binomial expression**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.reg_design;
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names Official_names*Official_names / showpvalues
                selection=none;
run;
```

- **Trinomial expression**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.reg_design;
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names Official_names*Official_names
                Official_names*Official_names*Official_names / showpvalues selection=none;
run;
```

- **Forward selection**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.Glmselect_Design
                plots=(criterionpanel);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names / selection=forward
(select=adjrsq stop=adjrsq) hierarchy=single stats=(adjrsq rsquare);
run;

proc reg data=Work.Glmselect_Design plots=none;
        ods select ParameterEstimates;
        model New_confirmed=&_GLSMOD / vif;
        run;
quit;
```

- **Backward elimination**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.Glmselect_Design
                plots=(criterionpanel);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names / selection=backward
(select=adjrsq stop=adjrsq) hierarchy=single stats=(adjrsq rsquare);
run;

proc reg data=Work.Glmselect_Design plots=none;
```

```
        ods select ParameterEstimates;
        model New_confirmed=&_GLSMOD / vif;
        run;
quit;
```

- **Stepwise regression**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.Glmselect_Design
                plots=(criterionpanel);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names / selection=stepwise
(select=adjrsq stop=adjrsq) hierarchy=single stats=(adjrsq rsquare);
run;

proc reg data=Work.Glmselect_Design plots=none;
        ods select ParameterEstimates;
        model New_confirmed=&_GLSMOD / vif;
        run;
quit;
```

- **Least angle regression**

```
ods noproctitle;
ods graphics / imagemap=on;

proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.Glmselect_Design
                plots=(criterionpanel);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official /
                selection=lar
(stop=adjrsq choose=adjrsq) stats=(adjrsq rsquare);
run;

proc reg data=Work.Glmselect_Design plots=none;
        ods select ParameterEstimates;
        model New_confirmed=&_GLSMOD / vif;
        run;
quit;
```

- **LASSO regularization**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.Glmselect_Design
                plots=(criterionpanel);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names / selection=lasso
(stop=adjrsq choose=adjrsq) stats=(adjrsq rsquare);
run;

proc reg data=Work.Glmselect_Design plots=none;
        ods select ParameterEstimates;
        model New_confirmed=&_GLSMOD / vif;
        run;
quit;
```

- **Adaptive LASSO regularization**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.Glmselect_Design
                plots=(criterionpanel);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names / selection=lasso
(adaptive stop=adjrsq choose=adjrsq) stats=(adjrsq rsquare);
run;
```

```
proc reg data=Work.Glmselect_Design plots=none;
        ods select ParameterEstimates;
        model New_confirmed=&_GLSMOD / vif;
        run;
quit;
```

- **Elastic regularization**

```
proc glmselect data=WORK.NEOPAN outdesign(addinputvars)=Work.Glmselect_Design
                plots=(criterionpanel);
        partition fraction(validate=0.1 test=0.1);
        model New_confirmed=Official_names / selection=elasticnet
(choose=adjrsq) stats=(adjrsq rsquare);
run;

proc reg data=Work.Glmselect_Design plots=none;
        ods select ParameterEstimates;
        model New_confirmed=&_GLSMOD / vif;
        run;
quit;
```

- **Ridge regression**

```
proc reg data = WORK.NEOPAN outvif plots(only)=ridge(unpack VIFaxis=log)
outest=originalnames ridge=0 to 1 by 0.005;
model New_confirmed = Official_names;
plot / ridgeplot;
title 'Ridge Regression Calculation';
run;
proc print data = originalnames;
title 'Ridge Regression Results';
run;
```

Appendix B. Trinomial expressions of emergent neologisms and buzzwords

Appendix B-1 Trinomial expressions on single variants of emergent neologisms

Trinomial-IVs	DVs	R^2	RMSE
Official,Official^2,Official^3	Newly confirmed cases	0.936	161.081
Official,Official^2,Official^3	Newly suspected cases	0.964	162.004
Official,Official^2,Official^3	New deaths	0.795	11.308
Official,Official^2,Official^3	Currently suspected cases	0.939	1003.686
Preofficial, Pre-official^2, Pre-official^3	Newly confirmed cases	0.680	360.299
Preofficial, Pre-official^2, Pre-official^3	Newly suspected cases	0.844	334.874
Preofficial, Pre-official^2, Pre-official^3	New deaths	0.659	14.572
Preofficial, Pre-official^2, Pre-official^3	Currently suspected cases	0.694	2246.905
Under-specifications, Under-specifications^2, Under-specifications^3	Newly confirmed cases	0.677	362.355
Under-specifications, Under-specifications^2, Under-specifications^3	Newly suspected cases	0.779	399.431
Under-specifications, Under-specifications^2, Under-specifications^3	New deaths	0.640	14.970
Under-specifications, Under-specifications^2, Under-specifications^3	Currently suspected cases	0.617	2512.525
Stigmatizing, Stigmatizing^2, Stigmatizing^3	Newly confirmed cases	0.909	192.128
Stigmatizing, Stigmatizing^2, Stigmatizing^3	Newly suspected cases	0.814	366.080
Stigmatizing, Stigmatizing^2, Stigmatizing^3	New deaths	0.793	11.349
Stigmatizing, Stigmatizing^2, Stigmatizing^3	Currently suspected cases	0.855	1549.297
English abbreviations, English abbreviations^2, English abbreviations^3	Newly confirmed cases	0.483	458.280
English abbreviations, English abbreviations^2, English abbreviations^3	Newly suspected cases	0.501	599.724
English abbreviations, English abbreviations^2, English abbreviations^3	New deaths	0.676	14.205
English abbreviations, English abbreviations^2, English abbreviations^3	Currently suspected cases	0.369	3225.573

Appendix B-2 Trinomial expressions on single variants of vector names

Trinomial-IVs	DVs	R^2	RMSE
Animal, $Animal^2$, $Animal^3$	Newly confirmed cases	0.715	340.221
Animal, $Animal^2$, $Animal^3$	Newly suspected cases	0.826	354.450
Animal, $Animal^2$, $Animal^3$	New deaths	0.655	14.642
Animal, $Animal^2$, $Animal^3$	Currently suspected cases	0.688	2270.192
Human, $Human^2$, $Human^3$	Newly confirmed cases	0.526	438.945
Human, $Human^2$, $Human^3$	Newly suspected cases	0.691	471.922
Human, $Human^2$, $Human^3$	New deaths	0.595	15.875
Human, $Human^2$, $Human^3$	Currently suspected cases	0.493	2890.607

Appendix B-3 Trinomial expressions on single variants of PPE names

Trinomial-IVs	DVs	R^2	RMSE
Hand, $Hand^2$, $Hand^3$	Newly confirmed cases	0.842	253.582
Hand, $Hand^2$, $Hand^3$	Newly suspected cases	0.908	257.342
Hand, $Hand^2$, $Hand^3$	New deaths	0.613	15.525
Hand, $Hand^2$, $Hand^3$	Currently suspected cases	0.858	1530.951
Eye, Eye^2, Eye^3	Newly confirmed cases	0.875	225.297
Eye, Eye^2, Eye^3	Newly suspected cases	0.982	115.044
Eye, Eye^2, Eye^3	New deaths	0.617	15.441
Eye, Eye^2, Eye^3	Currently suspected cases	0.953	880.087
Face, $Face^2$, $Face^3$	Newly confirmed cases	0.626	389.645
Face, $Face^2$, $Face^3$	Newly suspected cases	0.802	377.774
Face, $Face^2$, $Face^3$	New deaths	0.649	14.775
Face, $Face^2$, $Face^3$	Currently suspected cases	0.666	2347.527
Body, $Body^2$, $Body^3$	Newly confirmed cases	0.756	314.671
Body, $Body^2$, $Body^3$	Newly suspected cases	0.870	305.696
Body, $Body^2$, $Body^3$	New deaths	0.596	15.848
Body, $Body^2$, $Body^3$	Currently suspected cases	0.757	2002.758

Appendix C. Fine-tuned regression on all variants of emergent neologisms with trinomial expressions

Regression types	IV	DV	R^2	RMSE	R^2-baseline	RMSE-baseline
Forward Selection	All emergent neologisms cubic main effects	Newly confirmed cases	0.996	46.301	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected cases	0.999	9.307	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.996	1.787	0.995	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.999	47.312	0.999	80.280
Backward Elimination	All emergent neologisms cubic main effects	Newly confirmed cases	0.991	42.035	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected cases	0.999	15.386	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.996	1.375	0.996	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.999	56.676	0.999	80.280
Stepwise Regression	All emergent neologisms cubic main effects	Newly confirmed cases	0.995	54.342	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected case	0.999	16.073	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.996	1.841	0.995	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.999	61.594	0.999	80.280
Least Angle Regression	All emergent neologisms cubic main effects	Newly confirmed cases	0.971	121.687	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected cases	1.000	4.785	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.664	16.088	0.995	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.894	1509.316	0.999	80.280

Appendix C. Continued

Regression types	IV	DV	R^2	RMSE	R^2-baseline	RMSE-baseline
LASSO Regularization	All emergent neologisms cubic main effects	Newly confirmed cases	0.797	246.867	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected cases	0.965	161.409	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.865	7.681	0.995	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.999	72.587	0.999	80.280
Adaptive LASSO Regularization	All emergent neologisms cubic main effects	Newly confirmed cases	0.993	59.846	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected case	0.993	71.040	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.936	26.062	0.995	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.817	3432.593	0.999	80.280
Elastic Net Regularization	All emergent neologisms cubic main effects	Newly confirmed cases	0.858	263.109	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected case	0.992	94.064	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.617	13.208	0.995	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.938	1084.969	0.999	80.280
Ridged Regression ($\alpha = 0.01$, the best parameter)	All emergent neologisms cubic main effects	Newly confirmed cases	0.866	236.491	0.996	45.021
	All emergent neologisms cubic main effects	Newly suspected case	0.979	124.319	0.999	17.247
	All emergent neologisms cubic main effects	New deaths	0.707	13.658	0.995	1.844
	All emergent neologisms cubic main effects	Currently suspected cases	0.942	991.692	0.999	80.280

Appendix D. Coefficients of the variant in emergent neologisms in the final Least Angle Regression

Newly confirmed cases as dependent variables:

Variable	Estimate	SE	*t*	*p*
Intercept	-136.836	28.511	-3.97	<.05
Official names	0.012	0.00258	7.60	<.05
Pre-official names	- 0.084	0.00793	-9.47	<.05
Under-specifications	-0.0003	0.000897	-2.28	<.05
Stigmitizing names	0.613	0.0249	22.11	<.05
English abbreviations	-0.061	0.041	-1.49	0.1405
Official names^2	-1.8E-7	8.2882E-9	-21.42	<.05
Pre-official names^2	1.279E-6	5.210E-8	22.29	<.05
Under-specifications^2	5.306E-9	2.686E-9	1.98	0.0518
Stigmitizing names^2	9.2E-6	4.017E-7	-20.79	<.05
English abbreviations^2	-0.0000120	0.00000610	-1.97	0.0526

Newly suspected cases as dependent variables:

Variable	Estimate	SE	*t*	*p*
Intercept	-48.4422	22.37922	-1.83	<.05
Official names	0.0148	0.00202	9.05	<.05
Pre-official names	- 0.0547	0.00622	-6.79	<.05
Under-specifications	-0.0031	0.00070438	-5.37	<.05
Stigmitizing names	0.01170	0.01957	3.88	<.05
English abbreviations	-0.0043	0.03215	-1.31	0.1956
Official names^2	3.044E-8	6.505712E-9	4.92	<.05
Pre-official names^2	1.79E-7	4.08959E-8	2.45	<.05
Under-specifications^2	1.91E-9	2.107971E-9	-0.14	0.8902
Stigmitizing names^2	1.28E-6	3.153003E-7	-2.52	<.05
English abbreviations^2	1.03E-5	0.00000479	3.45	<.05

New deaths as dependent variables:

Variable	Estimate	SE	*t*	*p*
Intercept	5.403	2.350	2.13	<.05
Official names	0.002	0.000213	10.27	<.05
Pre-official names	-0.0003	0.000654	2.13	<.05
Under-specifications	-0.0005	0.0000740	-6.80	<.05
Stigmitizing names	-0.003	0.00206	-1.06	0.294
English abbreviations	-0.015	0.00338	-6.34	<.05
Official names^2	-5.62E-9	6.83E-10	-8.03	<.05
Pre-official names^2	8.94E-9	4.30E-9	-0.15	0.883
Under-specifications^2	8,83E-10	2.214E-10	1.94	0.056
Stigmitizing names^2	-5.69E-8	3.3111E-8	0.01	0.993
English abbreviations^2	2.22E-6	5.030E-7	6.53	<.05

Currently suspected cases as dependent variables:

Variable	Estimate	SE	*t*	*p*
Intercept	143.031	18.819	-7.49	<.05
Official names	0.007	0.018	10.15	<.05
Pre-official names	-0.075	0.055	-10.00	<.05
Under-specifications	-0.005	0.006	-6.96	<.05
Stigmitizing names	-0.553	0.174	13.42	<.05
English abbreviations	0.070	0.286	5.30	<.05
Official names^2	8.03E-7	5.780E-8	-12.02	<.05
Pre-official names^2	-1.84E-6	3.633E-7	13.08	<.05
Under-specifications^2	-2.17E-8	1.873E-8	4.13	<.05
Stigmitizing names^2	1.05E-5	0.00000280	-11.58	<.05
English abbreviations^2	4.52E-5	0.00004255	-8.41	<.05