参赛队号：（由大赛官网自动生成）

2022 年（第八届）全国大学生统计建模大赛

参赛学校：华中科技大学

论文题目：基于机器学习的公共卫生事件次生衍生事件识别、风险预测和演化机制研究

参赛队员：马世拓，张飏扬，李鹏坤

指导老师：王然，邹猛

基于机器学习的公共卫生事件次生衍生事件识别、风险预测和演化机制研究

摘要

突发性公共卫生事件在发展过程中具有突发性、强扩散性和不确定性，所带来的次生衍生事件种类并不局限于疫情的传播，也会带来很强的社会影响。2022年初上海市新冠疫情爆发以来所造成的社会影响和舆论传播为突发公共卫生事件下的应急管理与风险预警研究提供了一个典型的案例，以上海疫情为实证案例，综合利用机器学习方法对事件链展开分析。利用文本相似度模型和 BERT 分类器实现了次生衍生事件的自动识别与分类；利用 BERT-CRF 模型将次生衍生事件链条的风险预测抽象为序列标注问题并得到了一系列有意义的风险演化模式；利用聚类与话题模型分析在疫情演化的不同时间段网民的关注热点；利用 BERT-BP 网络实现了情感分类分析网民情绪随时间的演化。得出的一些结论能够在一定程度上提供在突发公共卫生事件中管理和预防的相关意见。

关键词：公共卫生事件，机器学习，次生衍生事件，BERT-CRF

Research on Identification, Risk Prediction and Evolution Mechanism of Secondary Derived Accidents of Public Health Emergencies Based on Machine Learning

Abstract

Public health emergencies are sudden, highly diffuse and uncertain in the development process, and the types of secondary derivative accidents brought about are not limited to the spread of the epidemic, but also have strong social impacts. The social impact and public opinion spread since the outbreak of the new crown epidemic in Shanghai in early 2022 provides a typical case for emergency management and risk early warning research under public health emergencies. Event chain expansion analysis. Using text similarity model and BERT classifier to realize automatic identification and classification of secondary derivative accidents; using BERT-CRF model to abstract the risk prediction of secondary derivative accident chain into sequence labeling problem and obtain a series of meaningful risk evolution mode; using clustering and topic models to analyze the hotspots of netizens' attention in different time periods of the epidemic evolution; using BERT-BP network to achieve sentiment classification to analyze the evolution of netizens' emotions over time. Some of the conclusions drawn can provide relevant advice for the management and prevention of public health emergencies to a certain extent.

Keywords: Public Health Emergencies, Machine Learning, Secondary Derived Accidents, BERT-CRF

图表目录

目录

引言

在现实生活中，一个突发事件的出现往往会引发其他事件从而造成一定社会影响与危害。而在 2020 年新冠肺炎爆发以后，对突发性公共卫生事件的应急策略与传媒预警同样是一项重要的研究问题。如何对突发性公共卫生事件进行次生衍生事件的识别预测和传媒预警对疫情防控工作有重要意义。

然而，现有研究对突发性公共卫生事件的次生衍生事件分析仍然存在一定局限性。现有研究对次生衍生事件识别尚没有一个明确的界限定义，采用量化方法对次生衍生事件展开分析的研究暂时还比较少。随着大数据时代的来临，一个重要的数据源就是网络与社交媒体数据，而将大数据挖掘与分析方法应用到突发公共卫生事件的次生衍生事件分析中的研究则更少。为进一步深入研究，本文基于机器学习方法与大数据挖掘手段对问题进行了一定探究。

2022 年上海的新冠疫情自爆发以来形势较为严峻，其发展动向牵动着全国人民的心，也在网络上引起了多个热点话题事件的讨论。本文以上海疫情为例，收集自 2022 年 3 月 1 日以来与上海疫情有关微博进行次生衍生事件的识别与分析，从大数据角度对这一案例进行分析。

本文研究数据来源新且详实，通过数据挖掘与自然语言处理的方法对问题展开深入探究，并能从社会科学角度对结果进行良好解释，对后期中国突发性公共卫生事件的管理和疏导有一定参考价值。

国内外研究综述

一. 突发事件次生衍生事件研究综述

次生衍生事件带来的危害和损失有时甚至超过原生事件，易引发更复杂的社会安全问题[1]。但多数研究却未能明确次生事件和衍生事件的概念，更未提出两者的差异和分类标准[2-6]。次生事件和衍生事件最重要区别在于与原生事件的关系，次生事件往往与原生事件具有相似的事件类型和产生机理且具有连带性或延续性，衍生事件与原生事件的事件类型和产生机理具有较大差异[7]。上述研究为本文提供了分类基础，但仍未建立筛选次生事件和衍生事件的合理标准。另外，传统的文本分析手段局限于样本数量和定性研究，无法适应大数据时代在内容挖掘上对广度和深度的要求[8]。

社会流动性和复杂性空前提高导致突发事件的次生衍生事件日益增加，原生事件与各级此次生衍生事件逐步构成链式效应，某些情况下，其造成的危害程度已经超过原生事件本身[9]。不少学者从不同学科视角分析事件演化特点，构建模型预测演化[10-15]，但研究多针对单一代表性事件，具有典型性却难以在普适性上实现较强的说服力。

二. 突发公共卫生事件研究综述

突发性公共卫生事件具有爆发前的不可知性、快速传染性和传播期内人群的高速流动性，给防控工作带来巨大挑战，也带来极大的社会影响[16]。而对突发事件的高效应对策略需要有庞大的技术支持，但如何将大数据方法应用于公共管理领域是一项重大的挑战。

而政府在公共卫生事件的应急策略中起到关键作用，祝哲等将突发公共卫生事件中的政府角色归结为四类: 风险沟通者、应急主导者、资源协调者和创新促进者，并指出现阶段社会环境发生了巨大变化，对政府在“新冠肺炎”等突发公共卫生事件中所承担的角色提出了新挑战，政府应根据社会环境的变化，进一步发挥作为政府在突发公共卫生事件应急管理中的应有作用[17]。但作为主导者应该如何进行合理的应急管理仍然缺乏一定经验。突发事件的应急管理可以分为三个阶段：事前的应急准备和风险预警，事中的应急决策、沟通协调和应急处置，事后的经验学习与事件调查等。目前我国在对风险预警、危机沟通和应急决策等方面仍然存在不足[18]，

在危机监测和预警过程中，信息发挥着越来越重要的作用, 传统的信息传递渠道和方式效率低下，不利于对灾害和危机及早预警和有效应对[19]。随着网络与社交媒体的迅速发展，突发性公共卫生事件的应对与治理应该变为线上与线下的双重治理模式，针对网络舆情的监测和疏导对于互联网时代的应急管理尤为重要。文献[20]将媒体大数据下的突发公共卫生事件网络舆情总结为五个要素：疫情，医情，政情，民情，媒情。文献[21]对政府在应急管理的过程中表现构建了合理的评价指标体系，但并未明确揭示如何对指标进行量化。危机情境中的政策响应速度,是公共管理研究的重要议题[22]，而社交媒体大数据的舆情分析与传媒

预警，则有利于政府合理做出决策。

对于突发性公共卫生事件的应急响应和风险预警，国外的一些经验可供我们参考。例如，美国在 SARS 大流行期间对原有的公共卫生医疗系统做出调整，逐步形成今天的“CDC (联邦) 疾病控制与预防系统 ——HRSA(地区/州) 医院应急准备系统——MMRS(地方)城市医疗应急系统”三级架构[23]。而在大数据治理方面，美国疾控中心和谷歌公司的研究人员利用 2003 年到 2008 年美国实际流感病例数据和谷歌搜索数据的挖掘，证明通过社交媒体大数据挖掘的方式进行疫情的风险预警时可行的[24]。但国外经验受到自身国情和经济发展水平影响，能否对中国适用还并不明确，能否在其他领域迁移应用同样不明确。

三. 序列标注研究综述

传统的序列标注方法采用隐马尔可夫模型，但由于隐马尔可夫模型为生成式模型，McCallum 认为在很多序列标注任务中，需要用大量的特征来刻画观察序列，并且很多问题是在已知观察序列的情况下求解状态序列，于是提出可以使用结合最大熵模型的马尔可夫模型进行序列标注[24]。但最大熵马尔科夫模型虽然结合了隐马尔科夫模型和最大熵模型的最大特点，但是仍然忽略了标签之间的约束关系，只求在当前时刻的最大条件概率。于是，开始广泛应用条件随机场对序列进行标注并取得重要成就[25-27]。但条件随机场特征稀疏庞大导致训练困难，并且不同领域有不同约束，难以进行迁移。

随着深度学习的发展，也有更多研究开始采用深度学习进行序列标注。文献 Ma 等人首次利用 CNN 进行序列标注并取得了良好效果[28]。而相比于 CNN，RNN 模型例如 BiLSTM 能保留到远端的上下文信息，百度团队基于 BiLSTM 实现了非领域特定的序列标注模型[29]。将二者融合的 BiLSTM-CNN-CRF 完全摆脱了人工特征的构造，实现了真正的 end-to-end 的序列标注模型[30]。而继 BERT、GPT-3 等一系列大规模预训练模型诞生以后，也出现了利用 BERT、BERT-CRF、BERT-BiLSTM-CRF 等模型进行序列标注的方法[31-33]。

四. 数据挖掘在突发事件风险演化中研究综述

有关网络上突发事件的传播数据挖掘目前主要是一些基础统计学方法。Chen 等人基于时空序列的统计学特征与文本情感识别方法，对美国 Harvey 飓风事件的影响与信息传播进行了挖掘[34]。Yao 等人发现转发行为表现出聚合特征，具有不同属性的用户在新浪微博上有特殊的转发习惯[35]。钟盛涛、王然等从特征工程的角度，对比支持向量机、随机森林、CatBoost 和 LightGBM 四种算法，从文本以及用户行为等多个角度挖掘了突发事件在微博传播中的重要特征[36,37,38]。王腾等通过对社交媒体大数据中突发事件的文本识别揭示了旅游类突发事件的时空分布规律[39]。

这些研究都从数据挖掘的角度对网络突发事件进行了建模，但大部分事件局限性比较强或者不够具象化，而且对灾害相互引发的机理与特性并未作出讨论。

五. 本文的主要工作

此前的一系列研究都具有其重要的前沿和先驱意义,它们为我们的研究提供了许多具有建设性的分析方法,以及一些在实践中具有重要的参考价值的结论。但此前的一些研究仍然存在着三个主要的限制:

1. 研究方法的局限性，传统研究中对网络舆情数据关注较少，且多为定性分析，缺乏量化方法与统计测度。

2. 研究案例的局限性，以往的研究案例大都距今较远，其研究方法不一定适用于实时案例的挖掘分析。

3. 数据体量的局限性，以往的研究案例不仅时间长周期长而且数据体量小，无法充分反映其在海量数据上的适用性。

鉴于前人的工作成果，我们在这项研究中主要的工作如下：

1. 收集海量微博数据，以上海疫情为话题筛选出事件并对其进行统计意义上的分析与可视化，并结合统计学与自然语言处理方法发现其统计规律。

2. 从序列标注与条件随机场的角度，利用自然语言处理的手段进行次生衍生事件的演化机制，在一定程度上可以视作因果关系。

3. 从舆情传播角度，分析不同阶段网民关注的热点话题和情绪演化，从而发现相应的传播意义的规律。

我们的工作成果从统计学意义上揭示了在上海疫情这一案例中公共卫生事件的次生衍生事件爆发与演化规律，并对事件的引发因果关系进行了一定程度上的阐释，对于公共卫生事件的防控管理与风险预警有一定借鉴意义。

变量符号与意义

表 1. 变量符号与定义

符号名称	含义
X	原始事件的文本序列
Y	事件链标签
w, v	条件随机场中的节点
λ	转移特征因子
μ	状态特征因子
t	转移特征函数
s	状态特征函数
Z	规范化因子

案例描述与统计特性

一. 数据的获取

自 2022 年春上海的新冠疫情爆发以来，有关话题在网络上引起了广泛讨论。我们以“上海疫情”为关键词，爬取了从 2022 年 3 月 1 日到 2022 年 5 月 11 日的微博 47.7 万条到我们的数据库中进行挖掘分析。在对突发事件分类与识别的基础上，为了保证结果的合理性，我们将上海疫情事件链中包含的六类次生衍生事件划分为如图 1 所示的层级。

图 1. 事件关系的分类层级示意图

由于数据主要构成为文本模态且数据体量较大，传统的纯粹统计方法难以对这类数据进行分析处理。基于此，我们使用机器学习的方法对此进行分析。

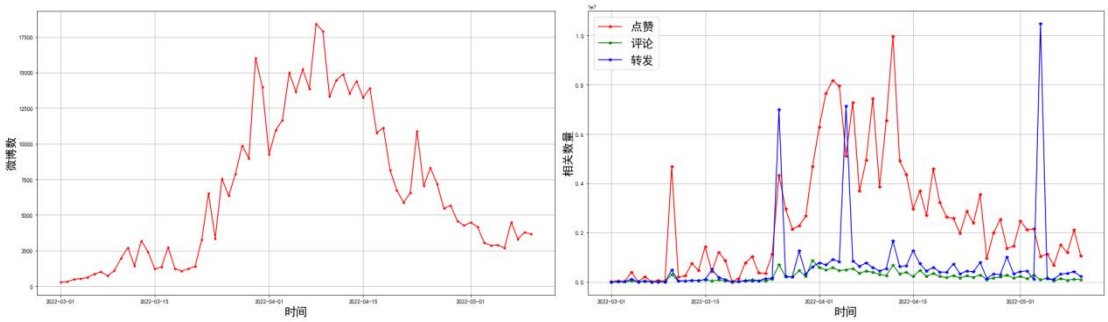
二. 数据的处理

我们基于 NLP 领域常用的 BERT 模型对文本进行了突发事件的分类。为了获得突发事件分类的预训练模型，我们基于我们自己的 200G 微博数据库上训练的突发事件分类模型对本文使用数据进行了分类，提取出了各类突发事件与非突发事件。由于爬取过程中容易混入一些噪声数据，我们也根据时间和地点约束将所研究时间段外的微博和非上海地区微博排除在外。文本中的噪声我们也进行了剔除，所得到的文本内容有利于后续文本分析。

另外为了方便后续操作，我们基于白化后的 BERT 对微博文本形成了对应的嵌入式表示，将其映射到句向量空间中方便模型进行计算。

三. 数据的统计特性与可视化

我们将预处理后的数据经过整理将相关统计特性描述如图 2 所示：



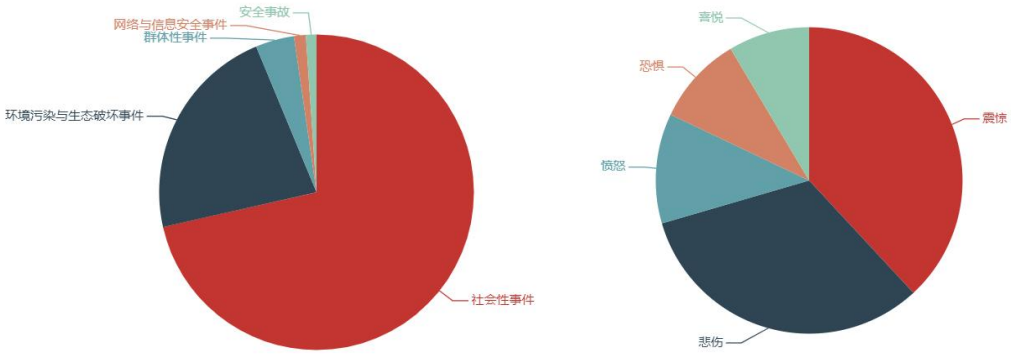
(a). 微博数量随时间的变化　　(b). 点赞评论转发随时间的变化

图 2. 话题的传播规模分析

从图 2(a)中可以发现，“上海疫情”这一话题下的微博大致呈现先迅速增高后逐渐降低的走势。借鉴 Steven 提出的“危机传播四阶段理论”，根据突发事件不同时期的演化特点将事件演化过程划分为四个阶段：潜伏期、初始期、爆发期和衰退期。在 2022 年 3 月 1 日之前上海疫情在未爆发时就已经处于潜伏期，而从 3 月 1 日到 4 月 1 日微博数量迅速蔓延，处于初始期，这一阶段下涉及事件却十分复杂，是次生衍生事件扎堆出现的时期，单位时间内如此大量的信息涌入对于有关部分的应急管理将是极大的挑战，而这一时期的突发性更是让事件影响变得变化莫测，多重因素叠加之下，起始阶段必然成为了突发事件演化过程中影响最大也是最需被重点注意的阶段。从 4 月 1 日到 4 月 15 日话题热度持续保持一个高水平，这一时期定义为爆发期，虽话题热度较高但总体保持稳定。从 4 月 15 日到 5 月 11 日开始逐渐下降，这一时期我们称为衰退期。通常情况下，该时期不再有新的次生衍生事件出现，仅有少量次生衍生事件仍会收到少数人关注，但整体热度、数量少。需注意的是，少量特殊事件收到原生事件或前期次生衍生事件的余波的影响，会在衰退期出现小高峰。

一个有趣的现象是，在观察图 2(b)时发现有几天出现传播小高峰。这几日代表的典型案例包括：3 月 10 日吉林农业科技学院学生发文引起同样身处疫情中心的上海网民转发；3 月 26 日专家指出由于上海的重要国际地位不能轻易封城引发网民情绪；4 月 5 日爆发式的新增病例与无症状感染让网民情绪陷入恐慌；5 月 4 日在合理部署下上海疫情防控效果开始突显，情况有所好转。这几波传播的高峰也揭示了疫情发展和网民情绪变化过程。

在进行分类后我们也将除公共卫生事件以外其他类型的次生衍生事件和网民情绪进行分类，得到如图 3 所示的分类图：



(a). 次生衍生事件类别占比　　(b). 网民情绪占比

图 3. 次生衍生事件类别和网民情绪

从图 3(a)中可以看到，除公共卫生事件以外，最容易引发的次生衍生事件为

社会性事件，占比 72%，其次是环境污染与生态破坏事件占比 19%。环境污染和生态破坏的主要原因是由于某些群体随意处理防疫物资，比如浪费防疫提供蔬菜等造成环境污染。而网民情绪演化造成的群体极化则由容易造成群体性事件例如网络抗议等。对于网民情绪，图 3(b)揭示了网民的情绪构成。在整个过程中，网民所展现的负面情绪是远远高于证明情绪的，仅有 10%的网民还能在整个过程中表现出乐观情绪。但整体而言，疫情爆发与隔离措施给居民带来的震惊感和悲伤情绪占比超过 70%，说明从舆情上来讲政策确实是引发了居民的不满情绪。其中，约有 12%的居民对政府的防疫政策表示愤怒，而约 9%的居民感到了恐慌感。

我们还将微博的热点高频词绘制在图 4 所示的词云图中：

图 4. 词云图

从词云图中可以看到，“上海”、“疫情”、“求助”是最大的三个热点词。说明在疫情的中心居民除了防疫抗疫以外还迫切需要获得帮助。核酸检测变得频繁，无症状感染者频发也引起居民恐慌。

得益于这些数据基本特征的挖掘，我们利用统计方法和自然语言处理手段对问题展开进一步探究。

事件识别与风险演化

一. 基于文本相似度的事件识别

安璐、李倩等对次生衍生事件有初步定义，认为次生事件和衍生事件的区别在于与原生事件相似度不同[7]。而事件的相似度我们选择使用话题文本的相似度来衡量，但多条微博可以对应一个话题，一条微博也可以对应多个话题，对微博文本进行相似度衡量并不利于我们分析。为了使问题得到简化，我们将微博的话题标签进行抽取，并按照时间排序获得原生事件和其他事件，利用余弦相似度进行相似度计算。若相似度较高则被判定为次生事件，相似度较低则为衍生事件。

$$sim_{\cos}(u,v)=\frac{u\cdot v}{|u|\cdot|v|} \tag{1}$$

图 5 展示了将句向量映射到欧几里得空间中的结果。可以看到，句向量之间存在一定相似性，向量夹角的余弦值可以衡量相似度。

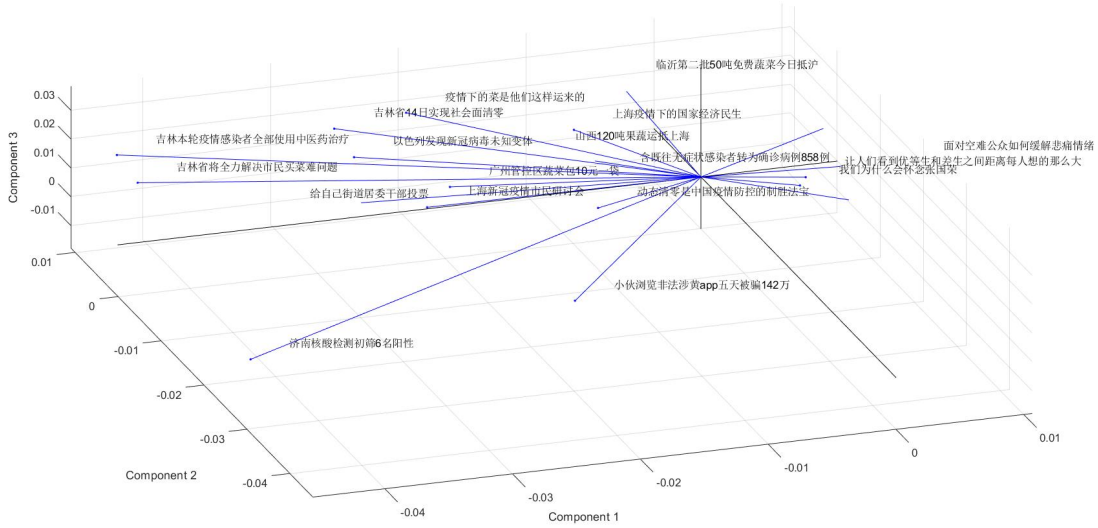


图 5. 句向量空间

所得余弦相似度均值为 0.05，标准差为 0.03。根据相似度，我们基于统计学常用的 3σ原则提取了 153 个次生事件和 862 项衍生事件，有 7013 项事件与疫情并无太大关联，被识别为无关事件。可以发现，事实上衍生事件的数量远远多于次生事件，但经分析发现，衍生事件中不少都同样属于公共卫生事件。这说明次生衍生事件并不能单纯从事件类型上判断，即使是同一类突发事件，在不同的场合不同的时间段不同的对象主体下所造成的事件因果也是不同的，需要具体情况具体分析。

基于相似度的标注，我们对微博文本对应的话题标签进行修正，将微博文本对应为次生事件、衍生事件和非次生衍生事件三个类别，并基于预处理得到的 BERT 嵌入表示构建 BERT-BP 神经网络模型对文本进行事件识别与分类，得到结果的准确率为 93.81%，合理完成了对次生衍生事件的分类与识别。

综上，所得到的结果均表明事件识别效果是非常好的，初步完成预期。

二. 基于 BERT-CRF 模型的风险演化

而对于不同类型次生衍生事件的演化，我们将其抽象为一个事件序列的标注。原始特征为事件对应的文本序列，而标签为事件类型。我们基于 BERT-CRF 模型对问题进行建模，这一模型包括 BERT 和 CRF 两个部分。

2018 年底微软提出的 BERT 相较于 Elmo[40]和 GPT-2[41]取得了更好的表现，目前也是应用最广泛的文本向量化方法之一,因为在 BERT 中，特征提取器也是使用的 Transformer,且 BERT 模型是真正在双向上深度融合特征的语言模型[42]。

BERT 架构如图 6 所示，不同于 GPT、ELMo 模型，BERT 采用的是 Transformer Encoder，也就是说每个时刻的 Attention 计算都能够得到全部时刻的输入。BERT 预处理进行下游任务的输入是三个嵌入表示叠加。

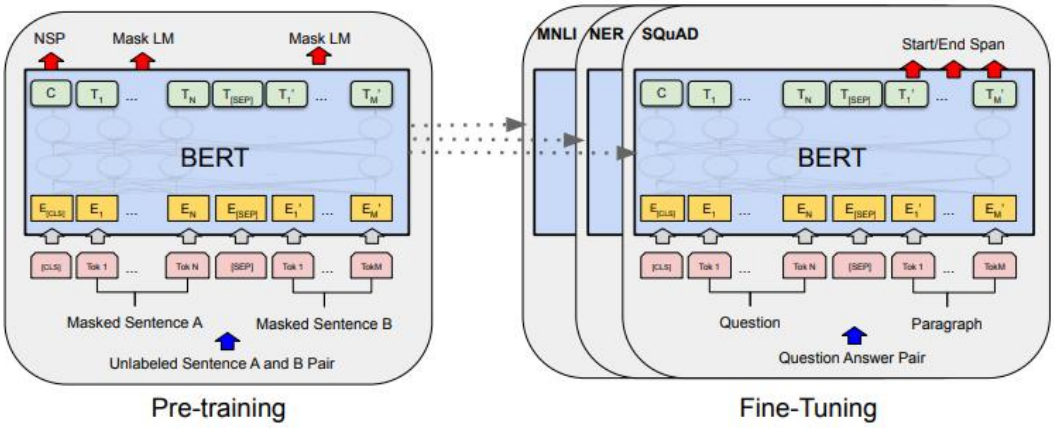


图 6. BERT 架构图

条件随机场(Conditional Random Fields,CRF)是给定一组输入序列的条件下，另一组输出序列的条件概率分布模型[43]。随机变量的集合称为随机过程。由一个空间变量索引的随机过程，称为随机场，也就是说，一组随机变量按照某种概率分布随机赋值到某个空间的一组位置上时，这些赋予了随机变量的位置就是一个随机场[44]。而条件随机场，就是给定了一组观测状态下的马尔可夫随机场，这个随机场满足马尔可夫性。也就是说 CRF 考虑到了观测状态这个先验条件，这也是条件随机场中的条件一词的含义。条件随机场的数学模型为：

$$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v) \tag{2}$$

这一表达式充分体现了条件随机场所带有的马尔可夫性。

CRF 中有两类特征函数，分别是状态特征和转移特征，状态特征用当前节点的状态分数表示，转移特征用上一个节点到当前节点的转移分数表示[47]。CRF 损失函数的计算，需要用到真实路径分数和其他所有可能的路径的分数[48]。在给定某个状态序列时，某个特定的标记序列概率为：

$$P(Y|X)=\frac{1}{Z}\exp\left(\sum_j\sum_{i=1}^{n-1}\lambda_j t_j(y_{i+1},y_i,x,i)+\sum_k\sum_{i=1}^{n}\mu_k s_k(y_i,x,i)\right) \tag{3}$$

其中，t 和 s 分别为转移特征函数和状态特征函数，Z 为规范化因子。

我们训练的 BERT-CRF 模型以白化后的 BERT 做为微博文本的句向量嵌入，

而 CRF 作为事件的状态转移。训练得到的指标如表 2 所示：

表 2. 不同类别事件的评估指标

Evaluation	安全事故	群体性事件	环境污染与生态破坏	网络与信息安全	社会性事件	公共卫生事件
Recall	91.22%	95.43%	95.32%	95.72%	95.65%	100.00%
Precision	91.49%	95.43%	95.37%	95.74%	95.69%	92.86%
F1-score	91.30%	95.42%	95.32%	95.70%	95.64%	96.30%

可以看到，基本上每个类的查准率和召回率都在 0.9 以上，我们认为，我们的结果已经获得了比较好的效果。整体的准确率能够保持在 95.11%。

通过 CRF 获得的转移矩阵如图 7 所示：

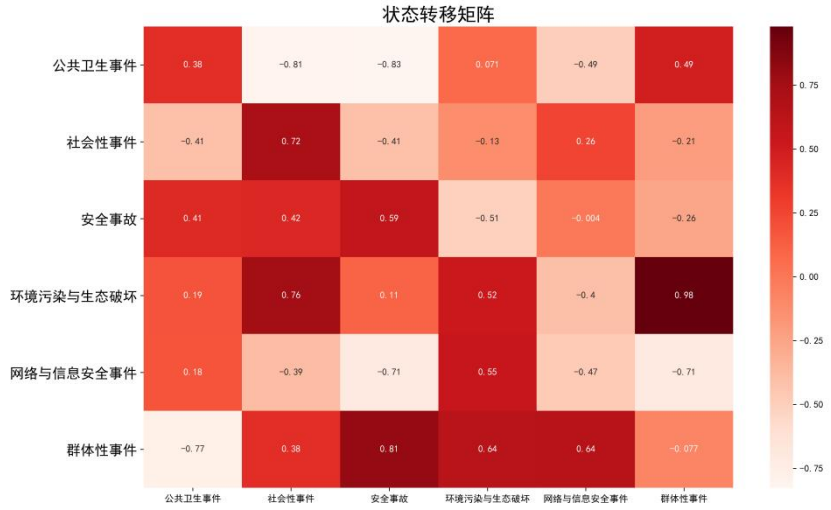


图 7. 转移矩阵形成的热力图

在图 7 中，颜色深浅表示转移得分大小，而转移得分的大小能够反映从一类事件转移到另一类事件的概率。公共卫生事件最容易引发的除了自身的扩散与传播之外，最多的一类其实并不是占比最高的社会性事件而是群体性事件。这也映证了在应急管理中居民由于封控等措施会带来情绪上的激化，若此时管理或引导不当则容易爆发群体规模的事件。而社会性事件容易自印发以外还容易引发网络与信息安全事件，那么当事件链条中出现社会热点时应马上关注网络舆情及时向公众传递真实可靠的信息。环境污染与生态破坏事件对群体性事件是一条重要的导火索，此前已经有研究对环境污染造成的群体极化与邻避冲突进行了探讨分析[45,46]，我们认为我们的转移矩阵也是对其的补充与验证。对于网络与信息安全事件，我们认为这一事件的引发链条结果 并不具备太强的可解释性，原因可能是由于这些网络安全事件在时空上分布较为零散。而群体性事件则容易引起安全事故，也符合我们的常理认知。

三. 基于主题聚类的话题识别

为对不同阶段的话题热点进行合理识别，我们在起始期、爆发期和衰退期分别利用 KMeans 对时期内的微博文本进行聚类找到最佳话题数。而聚类过程中 SSE 较大，在肘部图中不利于问题分析，我们以如下统计测度作为评估指标：

$$L(n)=\frac{\log SSE(n)}{\log SSE(2)} \tag{3.5}$$

绘制了改进肘部图如图 8 所示

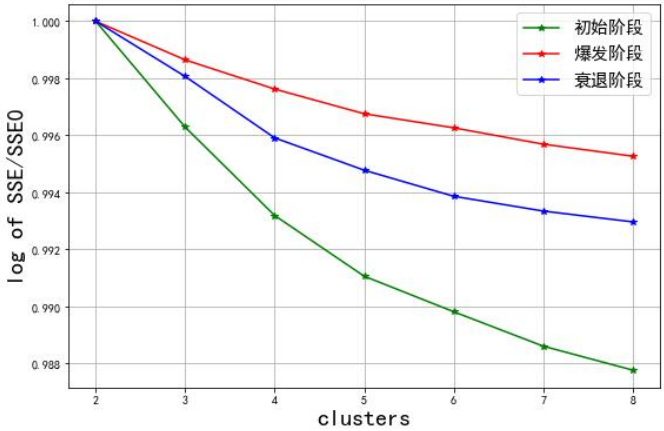


图 8. 改进肘部图

根据肘部图的拐点原则可以发现，起始阶段可以归结出五个热点话题；爆发阶段可以归结出五个热点话题；而衰退阶段则可以归结出四个热点话题。我们通过 LDA 模型对热点话题的关键词进行归纳得到时序图如图 9 所示：

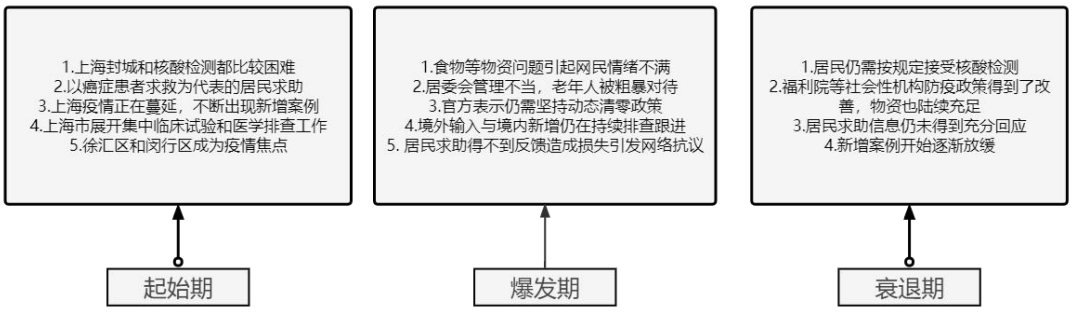


图 9. 话题时序图

在起始阶段，上海封城和核酸检测都是两大难题，而这两大难题的现状下，上海疫情蔓延的同时，以某位癌症患者及其家属的求救为导火索，居民生活问题的求助需求与管理资源分配不足的矛盾激化。与此同时，上海市也紧锣密鼓开展临床试验与医学排查，其中，徐汇区和闵行区成为焦点。在爆发期，仍然存在新增案例，但官方表态仍需坚持动态清零。在这一时期，真正引发网民情绪的是物资问题和部分基层管理组织过机械化的管理策略和态度，居民的诉求难以得到反馈。最后在衰退期，新增案例开始放缓，但居民仍需按时接受核酸检测，基层管理策略有所转变，但如何合理回应居民需求仍然是一项热点问题。

四. 基于文本分类的舆情传播

为了深入分析网民情绪随时间的变化，我们在我们团队此前整理的 200G 大规模微博数据集上训练了情感分类模型，并将其迁移到本文使用的上海疫情微博数据集上来。所得情感分类整体情况如图 2(b)所示。情感分类的准确率并不算太高，仅为 85.82%，但这样的分数对整体舆情的把握是足够的，因此我们仍然可以应用这一网络对微博文本进行情感分类。

情感识别使用的模型架构仍然为 BERT-BP 神经网络，仍然以白化后的 BERT 作为句向量的嵌入表示并将其送入全连接神经网络中进行训练得到结果。

我们基于预测的情感将每一日的情感占比随时间变化绘制在图 10 中：

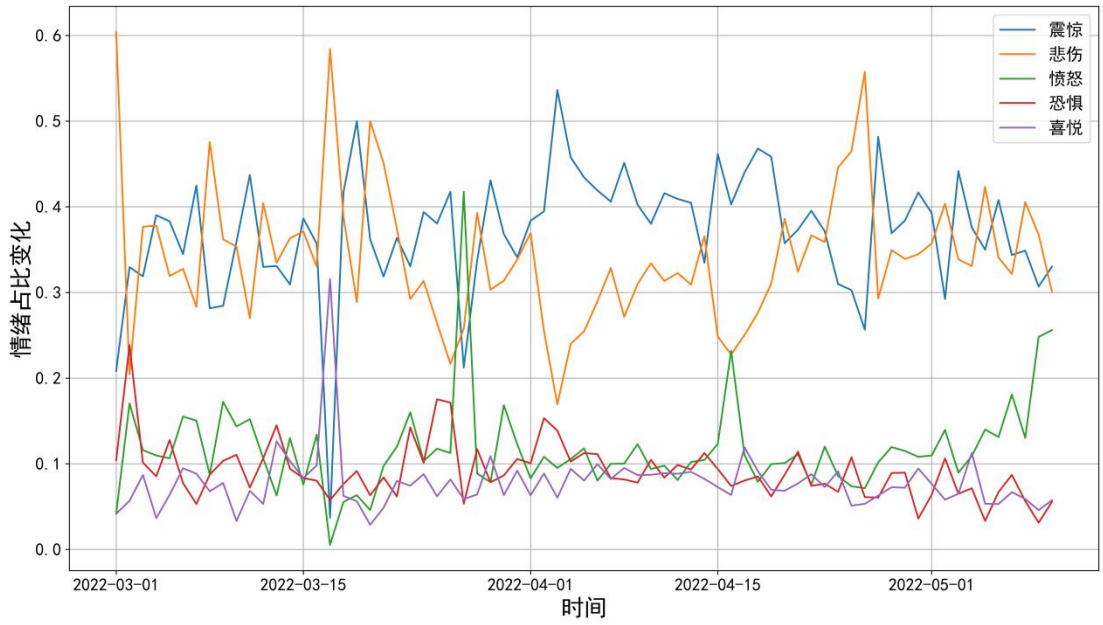


图 10. 每日情感占比随时间的变化

从图 10 中可以发现，喜悦情感在整体中占比较低，并且呈现前期密集后期趋于稀疏的分布状态。震惊情绪和悲伤情绪占比较高，并且方差也比较大，反映在疫情中后期网民情绪的不稳定性，需要得到政府等管理方的及时疏导与沟通。愤怒情绪在 3 月 28 日左右达到高潮，而在这一天，基层管理组织的管理失误与不当策略引发群体情绪，说明管理人员的措施如果引起民众愤怒需要及时疏导。而愤怒情绪在后期的缓慢升高也反映即使在后期管理人员的措施仍然存在漏洞，至少在网民视域下仍然会被视作漏洞，需要及时调查公众态度并合理调整措施，做好公关工作。

对结果的讨论与结论

一. 事件的分布规律

（1）从事件种类上来说，公共卫生事件会引起较大程度的社会性事件，当舆论舆情控制不当时会引发群体性事件，也要注意防疫物资造成的环境污染。

（2）随着时间的推移，公共卫生事件链的发展可以分为起始期、爆发期和衰退期，每个时期有不同的特点，但在起始期引发次生衍生事件种类多且变化快，是最需要加以控制与预警的时期。

（3）基于微博文本形成对次生衍生事件的自动识别能够有效进行风险预警，有利于管理者对事件进行控制。

二. 风险的演化规律

通过训练 BERT-CRF 模型得到的转移矩阵找出了一些事件之间的转移模式。出现这一现象的原因是不同类型的突发事件之间相互联系，但通过单一案例与类别不平衡的链条得到的结果并不全部具有可解释性。尽管如此，在这一案例当中我们通过数据驱动的策略仍然能够对问题进行合理的解答与分析，所得到的转移矩阵与转移模式大多也是具备良好可解释性的。

三. 舆情的传播规律

（1）从话题的热度来看，热点话题随不同时期也有所不同，但不同时期的话题也存在一定共性。管理者应当密切关注相关舆情，随实际情况调整策略。

（2）从网民情绪上来看，乐观情感在整体中占比较低，悲观的情绪占了大多数。震惊情绪和悲伤情绪占比较高，，愤怒情绪不仅方差大，在后期竟然还缓慢升高也反映管理人员的措施仍然存在一定问题，需要及时调查公众态度并合理调整措施，做好公关工作。

四. 展望与总结

在公共卫生事件的发展与管理过程中，管理者应该采取线上线下相结合的方式，合理结合网络上突发事件的热点话题与舆情演化，根据居民诉求合理调整管理策略而非简单的机械化管理甚至限制居民提出自身诉求。随着事件的进一步演化，事件会在不同时期呈现不同规律，在对应时期进行合理的风险预测和传媒预警对防控管理工作有着重大意义。本文以近期上海疫情为案例，从量化手段剖析了事件演化和舆情传播的相关问题，对后续公共卫生事件控制有一定参考价值。

参考文献

[1] 袁宏永,付成伟,疏学明,陈涛.论事件链、预案链在应急管理中的角色与应用[J].中国应急管理,2008(01):28-31.
[2] 罗成琳,李向阳.突发性群体事件及其演化机理分析[J].中国软科学,2009(06):163-171+177.；
[3] 张海蛟,邵荃,贾萌,朱燕.航班延误下的机场次生衍生事件安全管理研究[J].航空计算技术,2015,45(04):66-70.；
[4] 游上院. 基于情景-任务-能力的电力系统应急管理能力评估研究[D].广东工业大学,2019.DOI:10.27029/d.cnki.ggdgu.2019.001527.；
[5] 李嘉莉. 台风灾害下广东省沿海城市生命线系统安全规划研究[D].华南理工大学,2017.
[6] 吴国斌. 突发公共事件扩散机理研究[D].武汉理工大学,2006.
[7] 安璐,李倩.基于热点主题识别的突发事件次生衍生事件探测[J].情报资料工作,2020,41(06):26-35.
[8] 喻国明,李慧娟.大数据时代传播研究中语料库分析方法的价值[J].传媒,2014(02):64-66.
[9] 季学伟,翁文国,赵前胜.突发事件链的定量风险分析方法[J].清华大学学报(自然科学版),2009,49(11):1749-1752+1756.DOI:10.16511/j.cnki.qhdxxb.2009.11.02.
[10] 陈长坤,孙云凤,李智.冰雪灾害危机事件演化及衍生链特征分析[J].灾害学,2009,24(01):18-21.
[11]佘廉,吴国斌.突发事件演化与应急决策研究[J].交通企业管理,2005(12):4-5.
[12] 李藐,陈建国,陈涛,袁宏永.突发事件的事件链概率模型[J].清华大学学报(自然科学版),2010,50(08):1173-1177.DOI:10.16511/j.cnki.qhdxxb.2010.08.001.
[13]裘江南,刘丽丽,董磊磊.基于贝叶斯网络的突发事件链建模方法与应用[J].系统工程学报,2012,27(06):739-750.
[14]王建伟,荣莉莉.突发事件的连锁反应网络模型研究[J].计算机应用研究，2008(11):3288-3291.
[15] 马骁霏,仲秋雁,曲毅,王宁,王延章.基于情景的突发事件链构建方法[J].情报杂志,2013,32(08):155-158+149.
[16] 渠慎宁,杨丹辉.突发公共卫生事件的智能化应对：理论追溯与趋向研判[J].改革,2020(03):14-21.突发公共卫生事件中的 政府角色厘定: 挑战和对策 祝 哲 彭宗超
[17] 童星,丁翔.风险灾害危机管理与研究中的大数据分析[J].学海,2018(02):28-35.DOI:10.16091/j.cnki.cn32-1308/c.2018.02.004.
[18] 邵东珂、吴进进、彭宗超: 《应急管理领域的大数据研究: 西方研究进展与启示》,《国外社会科学》，2015 年第 6 期。
[19] 彭宗超,黄昊,吴洪涛,谢起慧.新冠肺炎疫情前期应急防控的“五情”大数据分析[J].治理研究,2020,36(02):6-20.DOI:10.15944/j.cnki.33-1010/d.2020.02.001.
[20] 王晓东,吴群红,郝艳华,康正,梁立波,陈海平.突发公共卫生事件应急能力评价指标体系构建研究[J].中国卫生经济,2013,32(06):47-50.
[21] 吴克昌,吴楚泓.重大突发公共卫生事件背景下政策响应速度差异研究——

基于 283 个城市复工复产政策的事件史分析[J/OL].北京工业大学学报(社会科学版):1-15[2022-05-26].http://kns.cnki.net/kcms/detail/11.4558.G.20220524.1357.002.html
[22] 薛澜,朱琴.危机管理的国际借鉴:以美国突发公共卫生事件应对体系为例[J].中国行政管理,2003(08):51-56.
[23] Ginsberg J.，Mohebbi M．H.，Patel Ｒ．S.，et al.． Detecting Influenza Epidemics Using Search Engine Query Data． Nature，2009，457(7232) ．
[24] 何炎祥,罗楚威,胡彬尧.基于 CRF 和规则相结合的地理命名实体识别方法[J].计算机应用与软件,2015,32（01）:179-185+202.
[25] 周俊生,戴新宇,尹存燕,陈家骏.基于层叠条件随机场模型的中文机构名自动识别[J].电子学报,2006(05):804-809.
[26] 洪铭材,张阔,唐杰,李涓子.基于条件随机场(CRFs)的中文词性标注方法[J].计算机科学,2006(10):148-151+155.
[27] Natural Language Processing (Almost) from Scratch
[28] Huang Z， Wei X， Kai Y . Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.
[29] Ma X， Hovy E . End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.
[30] 王子牛,姜猛,高建瓴,陈娅先.基于 BERT 的中文命名实体识别方法[J].计算机科学,2019,46(S2):138-142.
[31] 田梓函,李欣.基于 BERT-CRF 模型的中文事件检测方法研究[J].计算机工程与应用,2021,57(11):135-139.
[32] 张秋颖,傅洛伊,王新兵.基于 BERT-BiLSTM-CRF 的学者主页信息抽取[J].计算机应用研究,2020,37(S1):47-49.
[34] Chen S， Mao J， Li G , et al. Uncovering Sentiment and Retweet Patterns of Disaster-related Tweets from a Spatiotemporal Perspective—A Case Study of Hurricane Harvey[J]. Telematics and Informatics, 2019, 47:101326.
[35] Yao,W., Jiao, P., Wang, W.and Sun, Y.,“Understanding human reposting patterns on Sina Weibo from a global perspective”,Physica A: Statical Mechanics and its Applications 518, 374-383(2019)
[36] Prokhorenkova L， Gusev G， Vorobev A , et al. CatBoost: unbiased boosting with categorical features[J]. 2017.
[37] Meng Q . LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2018.
[38] Shengtao Zhong,Rui Sheng,Ran Wang,Yiyao Li. Prediction of the propagation effect of emergencies microblog[P]. International Symposium on Multispectral Image Processing and Pattern Recognition,2020.
[39] 王艳东,李昊,王腾,朱建奇.基于社交媒体的突发事件应急信息挖掘与分析[J].武汉大学学报(信息科学版),2016,41(03):290-297.DOI:10.13203/j.whugis20140804.
[40]Jozefowicz R， Vinyals O， Schuster M , et al. Exploring the Limits of Language Modeling[J]. 2016.
[41]Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving Language Understanding by Generative Pre-Training[J]. arXiv, 2018.
[42] Devlin J， Chang M W， Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[43]Lafferty J ， Mccallum A ， Pereira F . Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proc. 18th International Conf. on Machine Learning. 2001.
[44]Imry Y ， Ma S K . Random-Field Instability of the Ordered State of Continuous Symmetry[J]. Physical Review Letters, 1975, 35（21）:1399-1401.
[45]朱清海,宋涛.环境正义视角下的邻避冲突与治理机制[J].湖北省社会主义学院学报,2013(04):70-74.
[46]孟卫东,佟林杰.“邻避冲突”引发群体性事件的演化机理与应对策略研究[J].吉林师范大学学报(人文社会科学版),2013,41(04):68-70.
[47]陈晴. 基于条件随机场的自动分词技术的研究[D]. 东北大学.
[48]张鹏. Markov 随机场在图像处理中应用的研究[D]. 华中科技大学.

附录

本文的研究使用 Python 3.7.6 作为编程语言，本地 GPU 环境为 NVIDIA GTX 1650，CPU 为 Intel i7，操作系统 Windows 10。服务器环境为 Ubuntu 10，GPU 为双 GTX 2080Ti。

使用的数据集为 2022.3.1 日到 2022.5.11 日从微博爬取，以上海疫情为关键词，共 47.7 万条，经筛选，筛选出与突发事件有关微博 57432 万条，但即使微博文本中没有包含突发事件，我们认为，仍然能够在话题和舆情上对网民情绪进行探讨。

而针对突发事件的分类 BERT 模型和情感识别 BERT 模型，我们应用的是我们团队在先前两年内的研究中在我们自己的 200G 微博数据上预训练得到的突发事件分类模型和情感模型。

数据来源真实可靠，且时间非常近。

致谢

在本文编写的过程中，首先我非常感谢我的导师王然老师对我平日里的悉心指导,另外数学与统计学院的邹猛老师也提供了很多有价值的参考意见。

本文是对我一年前一项研究作品的一个延伸与补充，通过更 machine learning 的方式对问题进行了更深入探究。在这一年里，我首先要感谢我的师兄钟盛涛和贾彬师兄协助我完成实验任务，感谢张席斌师兄、李逸尧师兄、王诗月师姐、孙楚蕾师姐在社会科学与经济学领域为我补充的一些 domain knowledge。然后，我还应该感谢我实验室的十二名优秀的后辈：陈奥威，李鹏坤，申伯阳，叶奔航，王俊帅，赵睿，张飏扬，朱悦婷，樊翀宇，李天蔚，陶诗婷，朱幼衡。大家在各个项目组与我并肩作战，也非常感谢大家长久以来的努力。

其实我还应该感谢我的前女友，从我们认识第一天到现在已经快 400 天了，这是我第一段恋爱经历，她给我带来了人间第一抹浪漫与温柔，时至今日我想到这里，仍然会湿了眼眶。

自我读本科到现在已经三年了，三年之后又三年，不知我下一个三年会经历些什么。人生代代无穷已，江月年年照相似，生命非常短暂，但在短暂的生命中，作为一个旅客我能够从很多人身边路过。我寻找着我光与影的去向，并试着将光带给更多人，从前是这样，今后也会是这样。人生几十年，如梦亦如幻，我仍将揣着一份纯粹的心继续走下去。

最后，我需要感谢数模协会的孩子们，其实你们才是我的老师。