

## 阻击疫情, 抵制谣言

## 复旦95后学子用大数据“捉”谣

复旦大学新闻系大三学生温瑞琪与8位同学一起爬取了946条辟谣微博后,用数据可视化的方式提炼出谣言的辨别方法。“保持冷静,保持怀疑,小心求证。谣言,便能在铁一般的事实面前溃不成军!”这群爬得了数据、绘得出词云图的“95后”说,新闻人的使命感促使他们行动起来,变身“科学引导员”,辟除谣言,传递正音,也以此为抗疫贡献自己的青春力量。

青年报记者 刘昕璐

## 组团 谣言让人不安,一起粉碎它!

“刚接通知,定于今晚12点对全市主干道进行大面积消杀、消毒工作。因此次消毒液水浓度较大,请各位今晚10点后务必待在家中……”“水厂的朋友来电告知,非常时期,自来水中在允许的范围內加大了氯气的注入,水静置两小时以上再用……”

疫情爆发之后,温瑞琪一直在想,不在一线的自己能在这次疫情做些什么。每天在查看疫情相关信息的过程中,温瑞琪发现,有很多谣言给公众造成了恐慌,产生了不必要的负面情绪。

于是,温瑞琪主动加入了复旦新闻学院徐笛老师牵头的一个谣言相关可视化作品项目里面来,想要发挥的学科特长,为公众提供一些分辨谣言的专业建议。“我们的初衷就是对公众做一些启发,与数据新闻结合的方式,直观分析谣言背后的特征,继而提供辨识方法。”

在上海、宁波、绵阳、合肥、泉州、六安各地的同学们迅速在线上集结,微信群起名为“事实核查”。徐笛老师此时正在日本访学,也毅然投入了项目的指导。

据了解,团队成员主要是7位复旦新闻学院的学生,并邀请了2位华东师范大学通信与工程学院的学长学姐当外援,帮助项目做情感分析。团队2月12日开始组建,17日项目结束,爬取数据、讨论编码规则,查阅文献,每一样都不能疏忽,每一天也都在争分夺秒。

不爬不知道,原来,谣言真的是在“日长夜大”!“微博辟谣”是新浪微博虚假信息辟谣官方账号,通常转载其他认证账号发布的“辟谣信息”,每日汇总为合集置顶推送。团队就此爬取了“微博辟谣”平台于1月20日至2月11日间发布的与新型冠状病毒肺炎相关的全部微博共804条,删除重复、拆分合集后共946条。



复旦大学新闻系大三学生温瑞琪。

受访者供图

## 分析 为捉“谣”忙到凌晨4点收工

基于相关文献回溯,研究从谣言内容特征、呈现形式及发布主体等维度进行分析。其中,内容特征包括内容类别、信源交代方式、信源是否具名、是否含时间、是否含地点等维度。呈现形式包括纯文字、图文、文字加视频等。发布主体主要分为普通个人、媒体、政务机关等。

看似简单清晰的编码规则,恰是处理数据过程中最具挑战、分歧最大的部分。“我们都是人工编码。大家在确定编码表的时候,很难考虑到每条谣言的具体情况,有的时候会很迷惑。比如,一条谣言是微信聊天的截图,那争议究竟是纯文字还是图片?类似的争议,大家讨论了很久。”温瑞琪说道。

作为编码组组长的复旦新闻系大三学生朱月萌深有感受。她和另外三名同学共同负责编码工作,她觉得比较坎坷的是正式编码前的培训,“我们为了保证大家的互相信度比较高,编码工作可靠,光是编码培训就做了两天。最后才是每人200多条的工作量完成编码。”

在朱月萌看来,编码培训的过程真的需要耐心,需要不断调整编码规则以求得统一,确保最终编码的科学性。“我记得,最后一次编码培训完成,信度达到指导老师的要求后,已经是下午4点多了。大家都泛出了疲惫感,但是为了能够让可视化和分析小组在我们之后尽快开工,当天晚上,大家就自觉熬夜,做完自己部分的200多条数据编码。”

通过此前预编码的时间,朱月萌当时就预判出,最终的工作

要八九个小时才能完成。于是,在信度合格后,这名“95后”姑娘选择先去睡了两个小时,好好地吃了一顿晚饭,打足了精神,在晚上7点,开始投入正式编码工作。

为了给自己打气加油,朱月萌零时还开了一条微博在评论区记录自己的进度。“凌晨12点其实对于熬夜做项目来说,也不算太晚,但那天降温,风很大,夜里还蛮可怕的。当我完成100条数据时,我还去吃了碗泡面!”

凌晨2点多的时候,编码组陆续有同学往群里上交上完成后的表格。朱月萌此前有所休息,因此是最后一个完成。加上是组长,在统筹合并全部表格后,她看了一眼收工时间,已是凌晨4点多。

“做新闻赶工很正常,在时效性面前,通宵也并不算什么吧。”朱月萌说,由于数据爬取已经是几天前的事情,因而,编码、可视化的工作就更应强调效率,不容耽误。

## Tips

## 面对信息保持冷静保持怀疑小心求证

移花接木是谣言生成里惯用的一个伎俩。团队们给出了一个识别照片真假的“彩蛋”——比如,三年前拍摄的图片,重新成为热点,使用TinEye等图像反向搜索引擎即可帮助公众鉴别图片真伪。用户可以在TinEye主页面中上传提交图像文件或输入图像地址,随后TinEye会匹配寻找该图像的大量编辑版本,在搜索结果中提供各编辑版本的图像网址和发布时间,将搜索结果“按最旧排序”则可寻找到最接近源图像的编辑版本,根据其网页内容辨别图片真伪。

当然,面对信息的海洋,这群“科学引导员”给出的最重要的提醒还是:保持冷静,保持怀疑,小心求证,假信息的逻辑终将不堪一击。

## 提炼 利用大数据找到辨别谣言方法

编码员熬夜完工的成果不会被辜负。几小时后,可视化和分析小组接着热乎的数据表格继续开工。做词云图、数据可视化、数据分析,视觉美化,齐头并进,最终写成一篇文章《看了946条辟谣信息,我们提炼出一些信息辨别方法》,以“复数实验室”的账号,公开发表。

团队通过数据分析发现,“微博辟谣”账号转发的相关微博,其主题集中在“疫情传播情况”与“政府政策措施”上,呈现形式以文字为主,近7成被判定为谣言的信息中未交代信源。

温瑞琪绘制的词云图揭示,除“肺炎”“疫情”“武汉”“新型冠状病毒”等与事件直接相关的词

汇外,“车辆”“口罩”“消毒”“人员”“指挥部”等是主要的高频词。与人们生活贴近的“小区”“门店”“超市”“外出”等也频频被提及。“转发”“转告”等高频词映射出相关内容的传播需求。

通过分析相关微博,团队还发现,被判定为谣言的内容中,有30.2%写明了信源,假借官方发布内容,是最常见的方式,例如以“市政府办”“公安局网络中心”等落款。其次是假借其他“知情人”消息,例如有些内容来自内部工作人员透露。还有部分内容假借了媒体和专家的名义,如通过PS合成电视节目截图,或“钟南山院士称盐水漱口可防病毒”等。

学生团队据此采访了复旦大学新闻学院周葆华教授。他指出,带有时间和地点的信息,使得内容看上去比较具体、准确,对传播有作用;诉诸权威信源的策略则反映了来自权威部门的信息容易获得信任和传播,这也提醒政府部门应当及时公开与疫情有关的重要公共信息。

通过数据分析,团队还发现,绝大部分被判定为谣言的内容,都是由普通个人发布的,也有极少数出自媒体、政务机关、自媒体等。而发布辟谣信息的则多为政务机关,媒体也是较为重要的辟谣主体。少数辟谣信息由专家、普通个人、自媒体等发布。

## 呼吁 公共部门要及时回应公众关切

周葆华教授认为,疫情事关每个人的生命健康与安全,充满风险与不确定性,公众对疫情信息的需求度很高。同时,我们身处信息高速流动的移动互联网时代,加速了信息的传播;与疫情有关重要信息的公开、透明不够及时、充分,也是影响部分传言或谣言产生的可能因素。

通过参与这一项目,学生们直言,自身收获成长了很多。朱月萌认为,和身在前线的记者前辈们相比,他们在后方的数据整理工作并不算什么。个人感受比较深的是,在这种公共卫生事件面前,掌握准确的信息意味着掌握守护自己、他人和社会健康的权利。因此,在信息冗杂的互联网时代,研究“谣言”显得非常重要。

“这次经历,让我们洞察数据背后的规律和原由,可视化让

这些规律和原由能更直观地呈现出来。”温瑞琪专门提及,这次经历教会自己如何去尊重事实。她最初对于谣言就是愤怒,但是通过阅读文献、采访教授,如今的她越来越体会到,由于社会环境、社会心理等因素,谣言作为一种面对重要事件、存在模糊性情境下必然会产生产生的“即兴”信息,不可能完全消除,但可以减少其规模与负面影响。

因此,首先,在信息的供给方面,他们共同呼吁,政府等掌握公共信息的部门应当坚持公开透明,及时发布信息、回应公众关切,减少公众疑虑;在公众对信息的接受方面,则需要增强对纷繁复杂信息的鉴别力,例如对信息、信源的核实查证能力,通过多个渠道来校验与核实信息,提高自己的信息素养。

