

社交网络谣言检测



中国科学院 信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING, CAS

汇报人：王伟平

wangweiping@iie.ac.cn

目录

1

网络谣言的危害

2

谣言检测的挑战

3

谣言检测的方法

4

谣言检测研究展望

社交媒体网络谣言

社交媒体是人们用来创作、分享、交流意见、观点及经验的网络平台，成为信息传播和维系社会关系的重要渠道。根据新浪微博2019年发布的年度报告显示，三分之一的谣言始发于社交网络。



社交媒体中的谣言

- 社交媒体谣言危害国家安全，引发社会恐慌

2019网络十大谣言@中国互联网联合辟谣平台

1 美中情局向本·拉登家族道歉求谅解

真相：网传报道是一篇“洋葱新闻”。从美国各大新闻媒体的报道来看，中情局从未就“9·11”事件向本·拉登家族致歉。



2 华为与袁隆平合作培养水稻

真相：华为发布声明称，此消息不属实，华为没有任何从事水稻种植业务的计划，不会进入不擅长且没有优势的行业。



3 玉兔二号因重大故障7天无法移动

真相：上述内容是谣言。1月4日至10日，玉兔二号确实在月球背面一动不动，但并非出现重大故障，而是进入了月昼休眠模式。此后玉兔二号被唤醒并成功与嫦娥四号着陆器进行互拍。



4 中国人口负增长人口红利消失

真相：国家统计局局长宁吉喆称，2018年中国人口保持正增长，全年出生人口1523万人，其中相当比例是二胎，劳动年龄人口仍有约9亿人，人口红利仍然存在。



5 拼多多重大BUG致一晚损失200亿

真相：拼多多方面回应称，损失并不是市传的200亿，而是千万元优惠券。



6 炮弹果爆炸威力类似小型手榴弹

真相：中国科学院网站曾刊文科普，炮弹果个大皮硬，成熟后掉下来可能砸伤人，但不会因爆炸而伤人。



7 杨利伟因涉嫌军中腐败被降职

真相：中国载人航天工程网官方微博称，杨利伟因达到领导干部任职最高年限，不再担任中国载人航天工程办公室主任，根据组



航天

8 普华永道宣布将全员在家办公

真相：普华永道中国官方微博称，网传文章标题误导了员工和读者，公司提的是“灵活工作”制度，指员工工作满八小时，可灵活安排上下班时间，并非全员将在家办公。



9 私立幼儿园将退出历史舞台

真相：专家表示，国务院下发的有关通知，特指小区配套幼儿园不得办成营利性幼儿园，而不是其他场所的幼儿园也不能办成营利性幼儿园。网传说这是对国家政策的断章取义。



10 春节期间快递停运

真相：国家邮政局12日辟谣称，“春节期间快递停运”系假消息。据悉，邮政、顺丰、“三通一达”、百世等主要快递企业均表示不会停运。



网络谣言的危害（1）

1. 网络谣言加剧社会恐慌

“人人都有麦克风、时时刻刻都发言”的自媒体、微信息时代已经到来。现在**网络引燃现实矛盾的燃点更低**，网上一个不负责任的谣言，很容易成为引发社会恐慌的爆点。



2011年3月，日本发生地震核泄漏事故后，网民“渔翁”在QQ群上发消息称中国食盐将受核污染，后经大量转发扩散，中国部分地区开始疯狂抢购食盐，市场秩序一片混乱……

网络谣言的危害（2）

2. 网络谣言引发社会信任危机

我国正处在经济转轨与社会转型的特殊时期，也是社会矛盾的高发期。网络造谣者迎合转型期一些人内心的不安全感 and 不确定性，利用网民心理弱点，编造散布公益谣言，引发社会信任危机。



10月11日晚上6点50分左右在广东省佛山市汽车站附近走失，走失时上穿黑蓝皮衣，下身穿黑色裤子。黑色皮鞋，孩子家长非常着急，提供准确线索有重谢。
联系电话13710893270（父亲张文杰）
13725125415（母亲王俊丽）
13710893291（小姨张晓晓）

警察已经确认被人贩子拐跑了，希望大家不要再提供线索，不要冷漠，请转发分享给朋友，让更多人看到。希望大家都能伸出援助之手，让骗子无处可去，让更多的人看到。看的人多，找到的机会就会更大，求求你们，请转发，分享出去，让警察早日破案。抓住人贩子，不要再让人贩子危害其他的家庭。



在一系列寻人启事和网络募捐的谣言与恶作剧背后，严重透支了公众的爱心和信任，亵渎了这种民间自发的公益精神和公益力量。

网络谣言的危害（3）

3. 网络谣言损害国家形象

一起起网络谣言的背后，有的是个人无意识之举，但更多的是有组织有预谋而为之。在西方敌对势力指使下，一些别有用心者经常将网络谣言的矛头指向党和政府，恶意捏造事实，产生巨大的眼球效应，污染网络生态，损害国家形象



谣言自古以来就是个厉害角色，在我国古代，每逢社会混乱尤其是改朝换代的战乱年代，就盛产谣言。



有网民在互联网上特别是微博中编造、传播所谓“军车进京、北京出事”等谣言，产生了恶劣的社会影响。

网络谣言的危害（4）

4. 网络谣言扭曲意识形态

很多谣言最终导致的是民众对国家、对政府、对社会的失望和不满，西方国家也正不遗余力的对我们国家进行意识形态输出，网络成了他们最拿手的舆论工具。而对敏感事件情有独钟的“愤青”则不自觉成了他们输出意识形态绑架的工具。



“秦火火”和“立二拆四”创立了北京尔玛公司，伙同少数所谓的意见领袖，组织网络水军长期在网上兴风作浪。二人称，网络炒作必须要“忽悠”网民，只有反社会、反体制、反国家，才能宣泄对现实不满情绪，并叫嚣称，“谣言并非止于智者，而是止于下一个谣言”。

自动化谣言检测的意义



个人：上当受骗

商业：影响品牌形象

社会：引发民众恐慌

国家：危害国家安全



主要依靠人力

成本高、效率低

难以满足现实需要

迫切需要技术手段自动化检测信息内容的真实性

目录

1

网络谣言的危害

2

谣言检测的挑战

3

谣言检测的方法

4

谣言检测研究展望

谣言检测的挑战（1）

谣言内容的角度

- 真假难辨
- 普通人很难鉴别，机器更难

谣言

为什么相信谣言

他人原因

- 看似可靠的信息
- 似乎是真的消息

自己原因

- 认知非全知全能
- 心理有情感



新型冠状病毒基因序列与SARS相似度高达80%

正解

病毒的基因相似**不等于致病能力相似**。此外，新型冠状病毒与SARS和MERS分属不同的亚群分支，它们的病毒基因序列差异比较大

谣言检测的挑战（2）

谣言传播的角度

- 谣言传播迅速
- 传播过程中会多次发生反转
- 不断产生新的谣言，生命力仍然会很强

谎言跨越半个地球的时候，真相还在穿鞋
——马克吐温



一旦开始有人相信
谣言，相信的人就
会越来越多

信息
流瀑

群体
极化

谣言经群体讨论后
更容易走极端

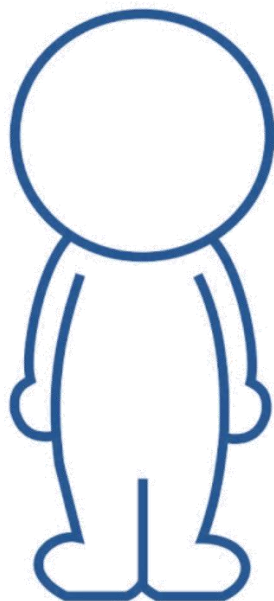
谣言检测的挑战（3）

信息源头的角度

- 谣言制造者
 - 普通网民的无心之过，缺乏“历史前科”
 - 非法组织的恶意为之，擅长移花接木、无中生有和身份隐藏



处于困境中
无恶意者
未经证实的言论
误解



谋求一己私利者
哗众取宠者
追求政治利益者
恶意中伤者

谣言检测的挑战（4）

谣言动机的角度

- 欺骗、误导具有知识盲点的网民以污染网络环境
- 炒作、夸张、冲击等眼球效应以牟取商业利益
- 煽动不满情绪，内容涉及公共安全引发社会恐慌



煽动不满
引发恐慌



欺骗、误导
炒作、夸张



意见领袖（个体）
伙同网络水军（群体）

谣言检测的挑战（5）

模型学习的角度

- 不同谣言之间缺乏相似性，很难总结规律
- 已有辟谣数据对今后新产生谣言的识别指导有限，不适合监督学习



规律难以总结



监督学习能力受限

目录

1

网络谣言的危害

2

谣言检测的挑战

3

谣言检测的方法

4

谣言检测研究展望

谣言检测的主要方法

- 基于内容特征的谣言检测
- 基于用户特征的谣言检测
- 基于传播特征的谣言检测

基于内容特征的谣言检测

1 基于语言特征的检测

上世纪50年代，美国社会心理学家奥尔波特和波斯特曼给出了一个决定谣言的公式：
谣言 = (事件的) 重要性 × (事件的) 模糊性，即谣言的产生与事件的重要性和模糊性成正比



2-19 20:50 已编辑

+关注

🔔 网传牛蛙列入禁食名录 国家林草局：从未发布相关名单

啥？以后！没有！牛蛙！吃了！？

(全国各地内情况有所不同，但已经有坐标长沙的网友亲身经历说他们那禁止售卖已经大半月了，也许是不给个人卖但是不影响餐厅供货？归根结底一句话，我现在特别想吃美蛙鱼头😋)

转发这条微博，请十个关注我的朋友吃牛蛙！

语言具有煽动性：含有大量标点、语气词、表情符号、诱导转发、细节模糊



2-19 14:50 来自微博 weibo.com

+关注

【蛙类养殖委称野生动物产品是刚需 中国野生动物保护协会：#撤销蛙类养殖专业委员会#】
2月16日，蛙类养殖委在其微信公号上发布原创文章《野生动物养殖是人类祖先的伟大创举》，文章称“对于人类而言，对野生动物产品的需求从未停止，某种意义上说已经成为‘刚性需求’。”2月18日晚，蛙类养殖委所属中国野生动物保护协会发声明致歉，称上述文章给社会造成不良影响，决定撤销蛙类养殖委。📺新京报动新闻的秒拍视频 @新京报动新闻

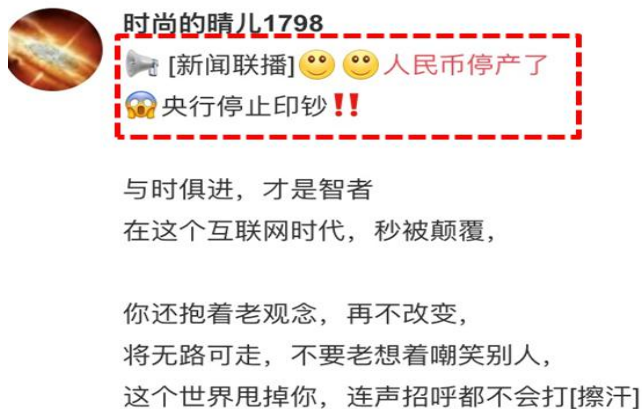
措辞中立、细节丰富、时间/机构等具体

语言风格可以作为自动谣言检测的特征之一

基于内容特征的谣言检测

2 基于群智的谣言检测

从群体智慧的角度看，网评集中反映了网民的观点立场，这些立场信息有助于判别和辨识原始内容。例如，谣言微博与真实的新闻微博相比，在转发和评论内容中表现出更多的质疑。

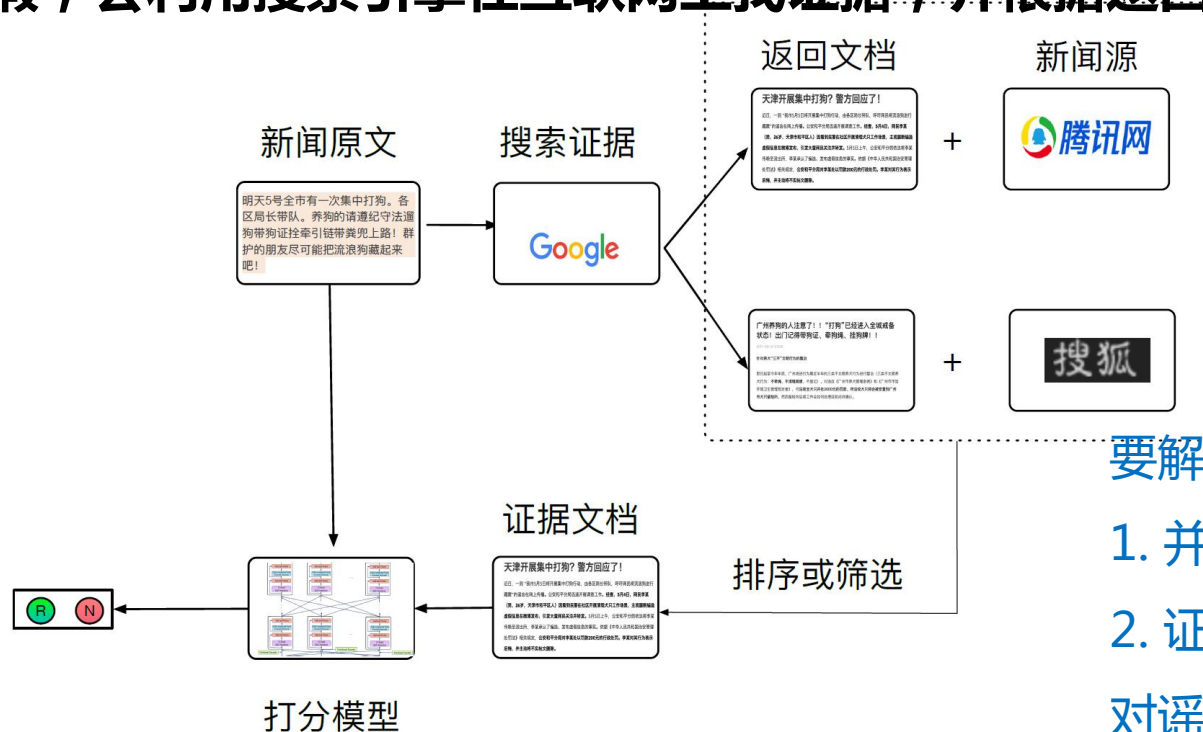


网民评论中的立场信息可以作为自动谣言检测的特征之一

基于内容特征的谣言检测

3 基于搜索引擎的谣言检测

搜索引擎可以为谣言检测提供更多外部证据，就像人类一样，为了鉴别真假，会利用搜索引擎在互联网上找证据，并根据返回结果进行下一步判断。



要解决的问题：

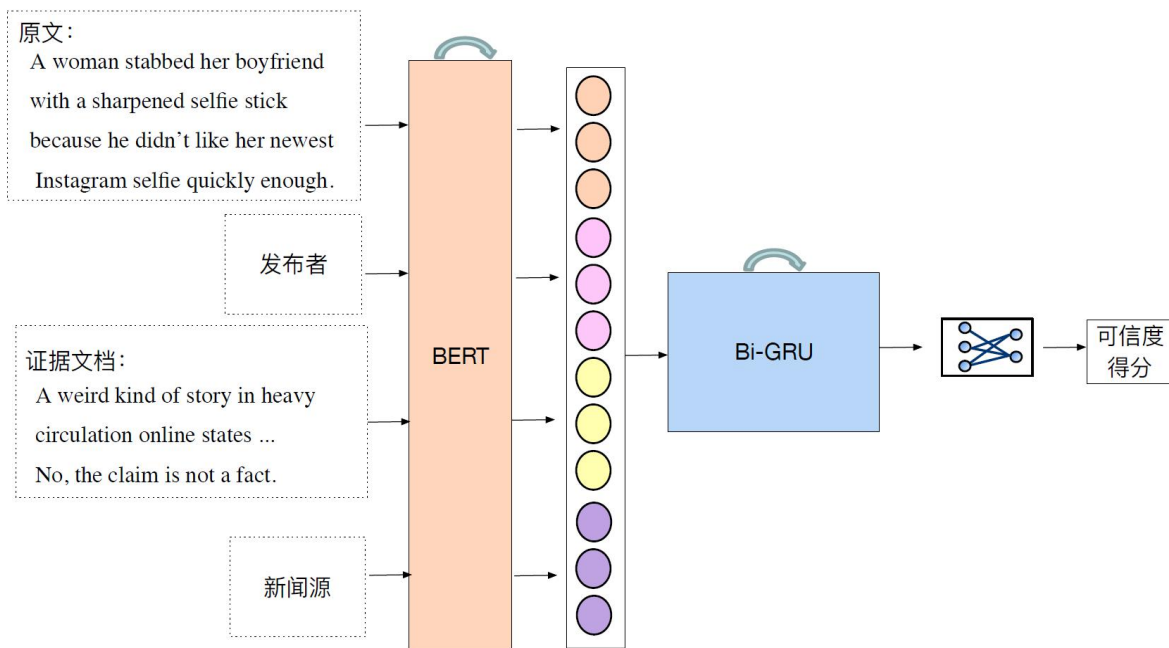
1. 并不是所有证据文档都有用
2. 证据文档内并不是每句话都对谣言检测有帮助

搜索引擎可以为谣言甄别提供有价值的证据信息

基于内容特征的谣言检测

3 基于搜索引擎的谣言检测

搜索引擎返回的相关证据文档，还可以结合新闻的发布者、新闻源等多个不同视角的信息，共同用于网络谣言检测。除此之外，利用搜索引擎还可以发现旧谣新传。



基于内容特征的谣言检测

4 基于多模态融合的谣言检测

网络谣言可能包含虚假内容文本，以及篡改伪造、重复使用的图片，从而误导读者。文本角度提供的信息有限，因此可以结合图片、视频等多模态信息进行辅助判断。



(a) Looking out my window to this is amazing!!
#newyork #usa #sandy
#hurricane #amazing
#powerful #picoftheday #

图文不一致



(b) Woman, 36, gives birth to 14 children from 14 different fathers.

文是伪造的



(c) an IAF F-16 deploying a flare over Southern Lebanon.

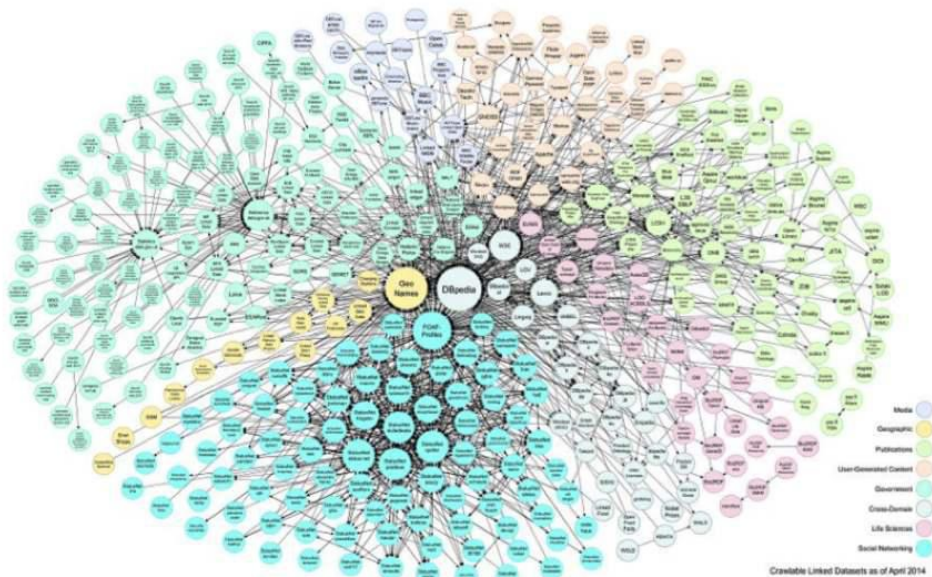
图是伪造的

结合多模态信息有助于提高网络谣言检测能力

基于内容特征的谣言检测

5 基于常识知识的谣言检测

每个人的认知都非全知全能，但如果能构建一个大而全的常识知识库，那么机器就可以在“全能大脑”的记忆指导下，自动鉴别那些违反常识类的网络谣言。



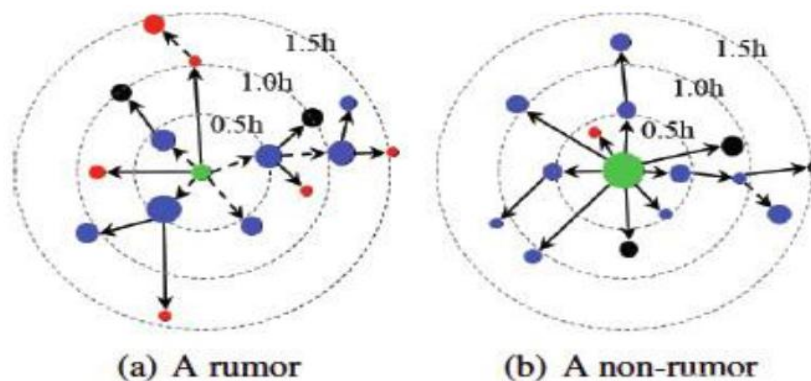
基于常识知识可以让机器模拟人类鉴别谣言

基于用户特征的谣言检测

- **谣言传播用户的特征可以有效辅助谣言鉴别**
 - 性格：性格开发性更有可能传播谣言
 - 情感：人民对于物理位置相关或者情感认同的事件更愿意传播
 - 信用：发布特定目的谣言的用户往往注册时间不长，信用度不高
- **用于谣言检测的用户特征选择**
 - 用户资料是否完整：用户名、头像、注册时间、注册地
 - 用户历史特征：发布信息数量、历史信用
 - 网络特征：粉丝数、朋友数
 - 系统特征：是否认证、用户类型

基于传播规律的谣言检测

- 谣言传播具有一定的独特特点，挖掘谣言传播与正常信息传播的异同，可以有效辅助谣言检测
 - 时间特征：源发时间、转发时间
 - 扩散方式：不同类型的谣言有特殊的传播结构
- 基于信息传播树的谣言检测



目录

1

网络谣言的危害

2

谣言检测的挑战

3

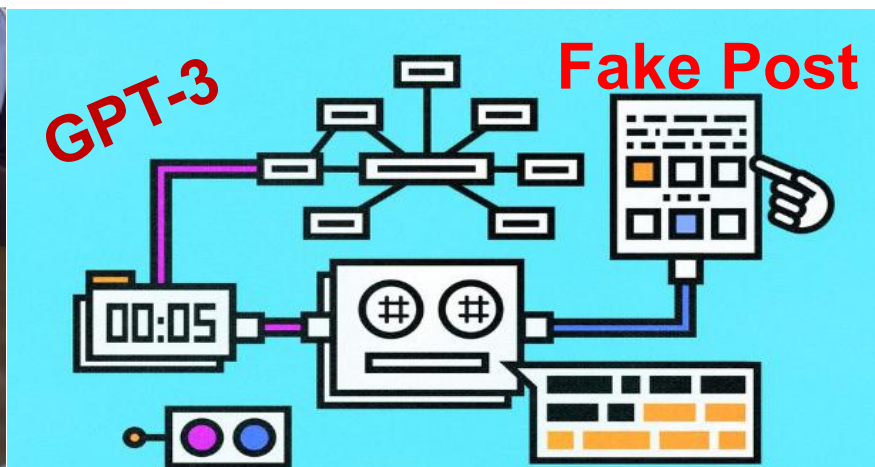
谣言检测的方法

4

谣言检测研究展望

人工智能带来新的风险

- **Deepfake最初用于换脸技术，后被滥用催生各种假视频**
 - 深度伪造视频层出不穷，让有心人轻松制造假新闻、假影片
- **GPT-3第三代生成预训练语言模型，具有1700亿参数**
 - 不仅能够答题、写文章，做翻译，还能生成代码、做数学推理、数据分析、画图表制作简历
 - 最强假新闻生成器，开发者OpenAI因担心技术滥用而未公开完整模型



技术本无罪，善恶在人心

谣言检测研究方向展望

- 新媒体时代谣言传播规律挖掘
- 深度伪造检测识别
- 实时谣言检测鉴别
- 众智谣言识别

谢谢，请批评指正！



中国科学院 信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING, CAS