互联网不实信息系统的流程和技术点

盛强

2020/5/18

第一部分 概念

	Authenticity	Intention	News?
Maliciously false news	False	Bad	Yes
False news	False	Unknown	Yes
Satire news	Unknown	Not bad	Yes
Disinformation	False	Bad	Unknown
Misinformation	False	Unknown	Unknown
Rumor	Unknown	Unknown	Unknown

For example, disinformation is false information [news or non-news] with a bad intention aiming to mislead the public.

新闻：具有时效性和事件性，其拥有大部分新闻要素。
非新闻：如历史内容、名人言论、科学健康常识。

一个事件发生时，新闻和非新闻不实信息都会出现。

各类信息的载体是多种多样的，包括文字、图片、视频、语音等或它们的混合体。

第二部分 不实信息判别的各种层次

- **表现层次：**How it is presented and how others says about it. 包括原文的风格、情感、痕迹，发布者、发布平台、发布时间等特征，也包括社交互动中其它用户的对应特征。
- **语义层次：**What it says and what others says about it. 包括原文讲述的内容和社交互动中其它人讲述的内容。
- **意图层次：**What it targets and what it can lead to on its targets. 包括发布者的意图对象和潜在后果。

第三部分 不实信息检测系统的生命周期

第一步 获取原始信息

1.人工方法：利用举报机制。
2.自动方法：

- For newly–emerging: 在社交上下文中的疑问句/判别句：对特定意图对象进行监控。
- For existing: 在已有的数据库中寻找高频或周期性假新闻的关键表达，进行监控。

第二步 真实性预测或判别

1.预测（Prediction）：使用非核查类信息

- **痕迹：对自然场景的视觉内容，检查其是否在成像后经过修改或压缩；对非自然场景的视觉内容（如截图），检测其压缩痕迹**
  - **优点：**假设较弱，可早期识别，在合适的设定下置信度高，解释性在一定情况下尚可
  - **缺点：**修改痕迹查全率低，压缩痕迹存在误报风险且敏感度高
  - **技术手段：**包括两种思路，一是针对自然场景内容，建模原始图像采集的固有特征（如相机指纹），针对非自然场景内容，提取其固有特征（如网页样式、截图分辨率、元数据等）；二是JPEG压缩步骤的学习（如量化表、DCT特征）、高频成分（噪声）的抽取和表示；GAN生成步骤的学习（如上采样引入的相关性、Loss设置带来的特殊统计规律）
  - **难点：**技术层面，如何将特定工具造成的痕迹考虑其中，这包括两个角度，是以通用化模型，二是对不同方法训练分类器分类，之后分别处理；实用层面，如何识别修改或压缩的目的，以排除无效信息
  - **BTW：**对文字，也可以检查其是否是机器生成的，但机器写稿本身应用范围还很有限，由于文本创作的成本很低，所以目前使用机器社撰假新闻还没有实际的市场，但它现在可以考虑被用来制造虚假评论等短小、含义清晰的文段，以作为水军。
- **风格：检查文字或视觉内容是否具有特定风格，如煽动性、冲击力，也可能是值任感**
  - **优点：**可早期识别
  - **缺点：**查全率低；风格存在不稳定风险（网络语言和媒体语言的变迁）；风格的描述方法没有可操作的共识；解释性有限
  - **技术手段：**提取统计特征进行评估，利用其它任务的预训练网络迁移（纹理学习）
  - **难点：**如何建立不同风格与不同类别假新闻的联系；对统计特征，如何将离散的统计特征空间融入连续的学习空间中。
- **情绪：检查文字或视觉内容是否传达了特定情绪或引起了其它用户的特定情绪**
  - **优点：**固有特点较强（公众情绪反应是相对稳定的），有较多先行研究可供借鉴
  - **缺点：**与真实性的相关度较弱
  - **技术手段：**利用情感词典作为先验，识别单条内容的情绪，表征并进行融合
  - **难点：**如何捕捉网络语言的复杂情感（对符号、表情的处理）
- **元数据：如发布者、平台、时间等数据**
  - **优点：**易获取，准确度高
  - **缺点：**几乎无解释性，几乎必须与其它方法结合
  - **技术手段：**大部分针对离散特征，使用经典ML方法
  - **难点：**如何与Fake Account Detection产生联系
- **内容：使用正文内、上下文内与正文与上下文的（一般是词级别）的共现关系**
  - **优点：**对单个帖子，信息量较为丰富
  - **缺点：**共现关系的可解释性弱，学习到的信号混杂了风格、情绪等不同方面
  - **难点：**如何建模丰富的上下文结构；如何针对视觉内容中的词进行识别考虑（OCR技术的可靠性）
- **上述信息的联系：如图文的相关程度、作者–文章–关注者的图结构、传播网络等**
  - **优点：**关注于网络特征而不是单纯的点特征，可以建模更复杂的交互信息
  - **缺点：**实际中数据噪声较多，要更多地考虑噪声（如表情包图片、图结构的边界、因删帖断掉的传播链路）
  - **难点：**对于成熟的表示结构（如传播网络），如何更好地学习表征；对于不成熟的表示结构（如图文相关性），要考虑图文相关有哪些类型，以及该问题与图文互译问题、视觉关系推理的关系。

2.判别（Judgement）：使用事实核查（Fact–check）

2.1 获取claim

- **人工获取：**FEVER任务中的Claim经过了人工提取和改写；Snopes和较真等平台的编辑会对待查证的说法进行总结。部分网页拥有ClaimReview这种Schema，可以采集。
- **自动抽取：**给定一条待查证消息，抽取其中的待查证claim
  - 现有方法倾向于言论类文本，相对简单

    Fact-check: ... The Facebook post says that "D.A.R.E. removed cannabis from its list of gateway drugs." That claim seems to stem from ... We rate this Facebook post False.

    Claim: "D.A.R.E. removed cannabis from its list of gateway drugs."
    Claimant: Viral image
    Verdict: False
  - 真实情况下，需要面对复杂的抽取任务（短文/长文，非正式语言和正式语言等）

2.2 搜寻证据

- **主要途径：**受限环境（维基百科、百度百科、知识图谱、知识库）；非受限环境（搜索引擎）。
- **流程：**获取候选页面→寻找相关性高的句子，形成候选证据集合；有时也会使用隐性知识（如BERT中蕴含的知识）直接上。

2.3 推理验证

- **主要技术点：**自然语言推理（判断蕴含、反对、没有充足信息等关系）、机器阅读理解（Q&A）
- **技术支撑：**比较前沿的包括BERT–based methods和Graph–based methods

优点：能够提供解读，对用户的友好型大大增加
缺点：对证据库要求很高，对模型推理能力也要求很高。
技术问题：现有评测针对的环境是封闭的，要实用化需要考虑若干实际问题，包括

- 如何从待查证文本中抽取合适的Claim？（其它在预测时有效的风格化文本，此时可能是一种噪声）
- 如何获得合理的查证文档？考虑对待查证信息进行分类，判断其属于静态知识类型（可以从百科中找到答案），还是周期性类型（可以从辟谣知识库中寻找答案），还是时效型（需要从搜索引擎获取证据），之后分流操作。
- 如何有效整理证据库？考虑半结构化证据（如使用Claim+Review或其它结构）。
- 如何面对多证据问题，多证据之间存在矛盾（噪声）如何处理？多证据的信服度（convincing or not）怎么衡量？

扩展：事实验证的任务与很多其它任务相关，如知识图谱补全（知识库收集），基于文章或表格的问答等。

3. **预测与判别的区别：**预测可以作为核查数据的前置任务，核查具有一定的可靠性（前提是证据库可靠），并可以提供更为丰富的解读信息。
4. **所有努力面临的共同问题：**

- **泛化性**
  - **跨语言：**中文、英文及其它语种（模型迁移性，不同语言互补扩充信息源）
  - **跨平台：**表现形式差异的影响（长文 vs 短文，营销号 vs 个人）；文化差异（大陆 vs 港澳台 vs 海外华语人士）
  - **对抗能力：**基于文本信号的方法很容易被对抗，两个方向考虑：用对抗学习使模型更鲁棒；思考抛弃vulnerable的方面，考虑什么东西无法被对抗。
- **效率**
  - **实时化：**对大数据量的问题，如候选新闻预测，如何降低检索获取成本
  - **轻量化：**对于问题输入完整性较好的问题，如何降低模型大小

第三步 解读和展示

1. **可视化预测信号：**如图片中使用CAM，文字中使用attention，但它们的可读性还比较有限。
2. **证据展示：**提供证据库中的证据句子与原文的关联性。可能存在一些人际交互的问题：如挑选更令人信服的证据进行展示。
3. **意图分析：**对于fake news，发布者在发布时必然存在一定的目的。
4. **意义：**用于fake news筛选（因为fake news的破坏力可能是更大的）；为核查中的证据搜集提供方向；为判罚提供参考。
5. **任务建模：**
   - 抽取分类问题：对给定内容，抽取意图对象，并输出针对该对象的正负面影响，或一些具体的类别（挑拨群体关系？损害名誉？营销？）
   - 抽取总结问题：对给定内容，抽取意图对象，并总结具体的影响（意图）（But it is too hard），如果参考WebConf '20的最佳论文《Open Intent Extraction from Natural Language Interactions》，那么具体的影响可以是某些动词。

   举例：

Input Text Utterance	Intents
Is it possible to navigate back … to previous page after save processing? … I have a page where I click on a link and use navigateURL … how to to obtain picture ID of previous calling page and complete the processing of the save.	navigate previous page, complete processing save
The "four tweets retweeted" page … find out all the users who retweeted a tweet of mine? … have retweeted a tweet and what their Twitter IDsare?	find retweeted Twitter IDs
Is there a WordPress plugin that will … have when a scheduled post is posted? … will tweet when you publish a post, but none I have tried will do it on a scheduled post	tweet when publish scheduled post
How can I keep my phone from just falling over when watching videos? … I also want to have my hands free to do other things …	keep phone from falling, have hands free
I'm starting a micro-school… I want to manage sick notes and absences … How can I synchronize one central Google Calendar … Parents should be able to schedule future absences and excuse past absences…	manage sick notes, manage absences, synchronize central calendar

   - **任务难点：**
     - 数据来源：意图是主观的，也是难以描摩的高层概念。可行的思路有：
       - 众包式人工标注（需要机器在其中参与抽取实体，考虑人在回路的标注系统）
       - 从辟谣帖中分析中抽取（仅一部分可行）
       - 从社交上下文中获取信号（比如假新闻是教育部每年给留学生很多资金，它的评论区就可能是骂教育部的更多，所以意图对象就找到了）
   - **相关任务：**
     - 方面层次的情感分析任务，但其一般情况下方面是固定的（价格、位置、服务…）；
     - 信息抽取任务，但其一般是只抽取不总结，本任务在某些情况下可能会无文本可抽取；
     - 仇恨言论识别，仇恨言论也是一种比较高层概念的文本类型，这类问题一般是先识别trigger word，之后观察其上下文是否是abusive的。
   - **技术流程中一些问题：**
     - 预处理：对于文本短小的情况，不得不抽取出图片中的文字；
     - 信号完整度差异：有的帖子找不到合适的辟谣帖（或辟谣帖不提供有效信息），有的帖子没有评论转发，得不到社交上下文，如何预测？
     - 输出不对齐:空情况/多情况：如何处理空意图对象的情况？如何处理多个潜在意图对象的情况？
     - 隐晦表示：网络语言中存在需要隐晦的表示，如何捕捉这种信号？
     - 模态扩展：多模态内容的恶意程度如何衡量？
     - 如何构建对意图对象的正负面影响和意图之间的关系？

第四步 治理

1. **澄清草稿自动撰写：**对于没有辟谣帖的情况，可以自动生成澄清草稿；
2. **切断传播路径：**寻找传播网络中的重要传播节点并屏蔽；
3. ……

第四部分 未来的发展方向

1. **更广的适用场景：**我们目前局限于微博平台（考虑到他们开放性和信号的丰富性），但有希望派生到其它类型的平台（如资讯聚合平台和短视频平台）
2. **更高的认知层次：**我们对信号和相关关注较多，但是现有CV/NLP的思索方向倾向于更高的认知层次，他们希望能够模拟人们的认知和推理过程。