1.论文信息

论文名：Fake News Detection on Social Media: A Data Mining Perspective

2.概述

越来越多人通过网上社交媒体看新闻，因为比传统途径更及时更便宜且更容易分享和评论。但社交媒体的虚假新闻很多，因此需要检测算法。
fake newsdetection 面临的挑战：
（1）新闻内容刻意误导读者，且主题、风格、发布平台多样，还会包夹真实证据来论证其虚假理论，因此现存的手工、文本特征检测方法不能奏效。需要辅助信息比如知识基础和用户社交参与情况。
（2）fake news 通常与时事相关，已有知识网络缺乏相关的真实证据；用户参与数据庞大、不完整、松散、有噪声。因此识别诚信用户、提取有用的特征和利用网络互动的高效算法仍待开发。
本文主要从两个方面研究 fake news detection 问题：特性描述、检测。
本文贡献：
（1）讨论了 fake news 的定义
（2）概述了已有的 fake news detection 方法
（3）提出后续发展方向

3.Fake News Characterization

1）Fake News 的定义

狭义：fake news 两个性质：真实性（是假的信息）、意图（误导读者）
广义：上述两个性质占其中一个即可（e.g. 讽刺文学）
本文只研究狭义，其成果能沿用在广义 fake news 检测上

因此将 Fake News 定义为：
Fake news is a news article that is intentionally and verifiably false.

以下不属于 fake news：
讽刺类新闻（无意误导或欺骗读者）、与时事无关的谣言（各种鸡汤）、阴谋论（难分真假）、误报信息（无意的）、恶作剧（为了乐趣）

2）传统媒体上的 Fake News

Psychological Foundations of Fake News. 读者先入为主，认为自己正确，不肯接受正确信息。
Social Foundations of the Fake News Ecosystem.
publisher：short-term utility，long-term utility
consumer： information utility， psychology utility
short-term utility+psychology utility->fake news

3）社交媒体上的 Fake News

两个特性：
Malicious Accounts on Social Media for Propaganda.
恶意用户。比如用于传播 fake news 的机器人用户、恶意煽动负面情绪的真人用户。
Echo Chamber Effect. 回音室效应，不利于真相传播。

4.Fake News Detection

1）问题定义

a 代表一篇新闻，包括 Publisher 和 Content
Publisher　Pa 包括了许多信息特征，比如姓名、领域、年龄等
Content　Ca 包括标题、正文内容、图片等

E = {eit } 代表 Social News Engagements，展示了 news 在 n 个用户中的传播过程
用户列表 U = {u1 ,u2 ,...,un } ，对应的帖子列表 P = {p1 ,p2 ,...,pn } ；
eit = {ui ,pi ,t} 代表用户 ui 在时刻 t 通过帖子 pi 将新闻 a 传播了出去。
如果 a 没被传播过，则 t=Null，ui=publisher

fake news detection 是二元分类任务

$$\mathcal{F}(a) = \begin{cases} 1, & \text{if } a \text{ is a piece of fake news,} \\ 0, & \text{otherwise.} \end{cases}$$

F 是需要学习的预测函数

fake news detection 的常用数据挖掘架构一般包括特征提取和模型构建

2）特征提取

传统媒介 fake news detection 主要靠 news 的内容，而社交媒体上 fake news detection 还可以利用辅助信息。因此需要从新闻内容（news content）和社交环境（social context）上提取特征。

a) News Content Features
News Content Features 包括
Source（author、publisher），Headline，Body Text，Image/Video
新闻内容主要是基于语言（Linguistic-based）和视觉（Visual-based）。

Linguistic-based: fake news 一般使用武断、煽动性的语言，且巧妙设计标题以引诱读者点击查看。可以从不同的层次提取语言特征，比如字母、单词、句子、文

档等。
一般同时利用通用的语言特征和特定领域的语言特征。
1、通用的语言特征通常在 nlp 领域上使用。典型的通用语言特征包括
（1）词法特征（lexical features，including character-level and word-level features），比如所有单词、每个单词的字母、常见单词、独特的单词等
（2）句法特征（syntactic features，including sentence-level features），比如功能性的单词或短语的频率（使用“n-grams”和词袋方法）标点和词类标记等。
2、特定领域（新闻领域）的语言特征，包括句子引用、外部链接、图的数目和平均长度等。

Visual-based:图片、视频更具煽动性。虚假图片的识别基于使用 user-level、tweet-level 的手工特征分类架构。最近 visual and statistical features 被用于新闻分类。
visual features 包括清晰度值、相关性值、相似性分布直方图、多样性值、聚类值。
statistical features 包括数量、图像比、多图像比、热图像比、长图像比。

b)Social Context Features
Social Context 的三个主要方面： users, generated posts, and networks 这三个方面的特征提取方法：
User-based：fake news 传播者可能是机器人账户，因此需要获取用户信息和特性。用户特征可分类为不同层次：个人层次 or 群组层次。
Post-based:用户通过发帖表达对 fake news 的个人情感和观点，比如怀疑或者煽情。特征提取分为三个层次：post level, group level, and temporal level
Network-based: 通过在发布相关 post 的用户之中构建特定网络以提取 Network-based 特征。

3）模型构建

a) News Content Models
依靠提取到的 news content features 和已证实的知识网络来检测。
当前有 Knowledge-based 和 Style-based 方法。
Knowledge-based:
检查 news 的主要观点的真实性的 Fact-checking 方法主要有 expert-oriented（人工检查，不实际）, crowdsourcing-oriented（读者投票，人多力量大）, computational-oriented（算法，先提取观点，后根据开放网络或知识图谱判别观点虚实，即判断该观点是否能由知识图谱中的已知事实推出）.
Style-based:
从 news content 中获取其写作风格，主要方法包括
Deception-oriented and Objectivity-oriented

Deception-oriented：检测欺骗性短语，主要有 Deep syntax（PCFG 方法）和 Rhetorical structure（CNN 算法检测真假新闻的区别）

Objectivity-oriented：检测新闻的客观性，比如 hyperpartisan styles（极端地支持

某一个政党，可用语言特征提取方法检测）和 yellow-journalism（黄色新闻一般是标题党）。

b) Social Context Models
包括 Stance-based and Propagation-based

Stance-based:根据相关 post 的立场推断 news 真实性。post 的立场有 explicitly（明确，如赞，踩）和 implicitly（含蓄，要依靠特征学习方法）两种。
stance 分类方法可参考 LDA（潜在狄利克雷分布），其中 stance 类别有支持、中立、反对三种。

Propagation-based:根据相关 post 的相互关系预测 news 可信度。假设 news 的可信度与相关 post 的可信度关联性高。可构建两种网络模拟传播过程：
homogeneous credibility networks （单一实体，比如 post 或 event）
heterogeneous credibility networks（多种实体，比如 posts、sub-events 和 events）

4.效率评估

1）Datasets

四个公开可用的数据集：BuzzFeedNews, LIAR, BS Detector, CREDBANK

这些数据集各有缺点，比如 BuzzFeedNews 只包含标题和正文内容且来自数目有限的几个新闻机构（9 个）；LIAR 只有个人言论而不是正规的新闻，即言论来自个人而不是新闻出版社；BS Detector 的标签是模型预测得出，而不是专家评估，因此准确性存疑；CRED-BANK 是有关推特的数据集而不是新闻。

对这四个数据集进行对比，标出能提取出来的特征种类，可见没有一个数据集能提供所有种类的特征。

Table 1: Comparison of Fake News Detection Datasets.

Features / Dataset	News Content		Social Context		
	Linguistic	Visual	User	Post	Network
BuzzFeedNews	✓				
LIAR	✓				
BS Detector	✓				
CREDBANK	✓		✓	✓	✓

此外作者正在收集合适的数据集 FakeNewsNet

2）指标

fake news detection 可看做二元分类任务，其四个经典的评估标准有：

- **True Positive (TP):** when predicted fake news pieces are actually annotated as fake news;（被检测为 fake news 且真的是 fake news 的实例）
- **True Negative (TN):** when predicted true news pieces are actually annotated as true news;（被检测为 true news 且真的是 true news 的实例）
- **False Negative (FN):** when predicted true news pieces are actually annotated as fake news;（被检测为 true news 但实际上是 fake news 的实例）
- **False Positive (FP):** when predicted fake news pieces are actually annotated as true news.（被检测为 fake news 但实际上是 true news 的实例）

$$
\begin{aligned}
Precision &= \frac{|TP|}{|TP| + |FP|} \\
Recall &= \frac{|TP|}{|TP| + |FN|} \\
F1 &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \\
Accuracy &= \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}
\end{aligned}
$$

精确率(Precision):被检测为 fake news 的实例中，真的是 fake news 的实例的比率。Precision=1 时代表检测结果完全正确。
（精确率的一般解释：针对模型判断出的所有正例(TP+FP)而言,其中真正例(TP)占的比例.）

召回率(Recall): 被正确检测为 fake news 的实例占所有 fake news 实例的比率。Recall=1 时代表所有 fake news 都被检测出来了。
（召回率的一般解释：针对数据集中的所有正例(TP+FN)而言,模型正确判断出的正例(TP)的比例.）
*召回率针对的是数据集中的所有正例,精确率针对的是模型判断出的所有正例

F1 值是精确值和召回率的调和均值
Accuracy 是检测正确的样本数与总样本数之比。

还可以使用 ROC 曲线与 AUC 值做评估
ROC（Receiver Operating Characteristic），其主要分析工具是一个画在二维平面上的曲线 ROC curve。横坐标是 FPR(false positive rate)，纵坐标是 TPR(true positive rate)。分类结果可映射成一个点。

TPR：在所有实际为 Positive 的样本中，被正确地判断为 Positive 的比率（与 Recall 相同）。TPR=TP/(TP+FN)
FPR：在所有实际为 Negative 的样本中，被错误地判断为 Positive 的比率。FPR=FP/(FP+TN)。
可见应该是 TPR 越高越好，FPR 越低越好；

如果分类器只会将样本分类为阳性，那么 TPR=1，FPR=1，分类效果不好。
如果分类器只会将样本分类为阴性，那么 TPR=0，FPR=0，分类效果不好。
在（0,0）于（1,1）之间画一条直线，位于直线上方的点则是 TPR 高，FPR 低，分类效果较好，且越靠近左上角越好；位于直线下方的点则是 TPR 低，FPR 高，分类效果较差。
ROC 曲线：将阈值设置从 0 遍历到 1，将得到的分类结果点连接起来即得到 ROC 曲线。

AUC（Area Under the Curve）值为 ROC 曲线所覆盖的区域面积，显然，AUC 越大，分类器分类效果越好。
AUC = 1，是分类完全准确
0.5 < AUC < 1，AUC 越大，分类效果较好，有预测价值。
AUC = 0.5，跟随机分类一样，模型没有预测价值。
AUC < 0.5，比随机分类还差；但只要将分类结果调转，则优于随机分类。
AUC 常用于不平衡的分类问题上，因此适用于 fake news classification

5.相关领域

1）Rumor Classification
谣言的状态有：true,false,unverified
谣言分析集中于四个子任务：rumor detection（判断是否是谣言）, rumor tracking（跟踪那些谈及谣言的 post）, stance classification（确定相关 post 的立场）, veracity classification（判断谣言真实度，这个子任务最类似于 fake news detection）。
谣言有长期谣言和短期谣言，fake news 一般与时事挂钩

2）Truth Discovery
从多个相互矛盾的信息源中分辨出正确的信息源。但依赖于多角度信息源的采集，如果 fake news 发布时间早，其他信息源及相关 post 过少，则难以检测。

3）Clickbait Detection
标题党，诱发读者好奇心点击，通过提高点击率获得广告营收。但往往文不对题，迁移到 fake news detection 可以使用语言特征学习来发现标题与新闻内容的不一致性进行检测。

4）Spammer and Bot Detection
垃圾邮件检测已有算法主要依赖于从用户和社交网络信息中提取特征，社交机器人检测已有算法则是基于社交网络信息、差异特征等。因此垃圾邮件检测和社交机器人检测都可以迁移到 fake news detection 中用以识别发布 fake news 的恶意用户。

6. 未来研究方向

将未来研究方向分为四类：
Data-oriented, Feature-oriented, Model-oriented and Application-oriented.

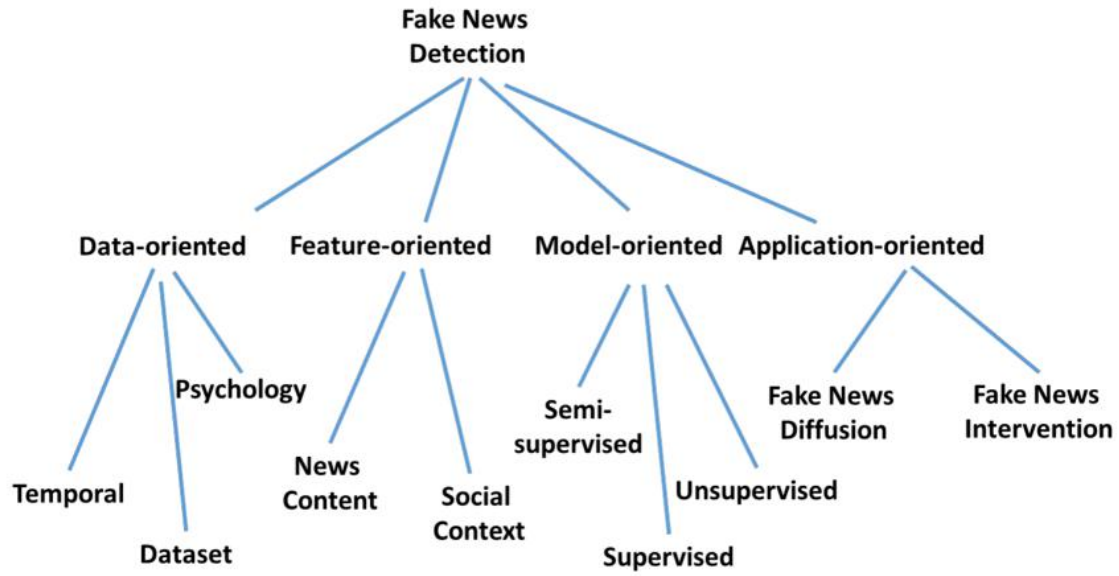

Figure 2: Future directions and open issues for fake news detection on social media.

Data-oriented:

三个方面：dataset, temporal and psychological

dataset：上文已经论证过不存在能提取所有类型的特征的数据集，因此正确的方向应该是找一个易于理解和大规模的数据集即可。

temporal：发展 early fake news detection，在 fake news 传播的早期就检测并警告

psychological：相关心理学理论很成熟但未得到充分利用，比如回音室效应作用很大，可以进一步研究如何利用。此外许多已有算法仅关注新闻的真实性，忽视了新闻发布者的意图，因此可以探究如何使用数据挖掘方法获取新闻的意图。

Feature-oriented：前文提到特征学习的两个信息源： news content，social context.

news content 方面介绍了 linguistic-based and visual-based techniques 提取文本特征。

linguistic-based 方面特征提取技术在 nlp 领域已经得到很大发展，词向量技术、深度神经网络的应用也有一定前景

visual-based 方面很少人研究，且挑战很大，急需先进的特征学习方法。

social context 方面 user-based, post-based, and network-based features

user-based 现有特征集中在普通的用户档案上，而不是按类型区分用户并提取指定用户的特征。

post-based 使用 CNN 算法能更好的获取 post 的观点和反应，同时要把 post 的贴图也考虑进去。

network-based 特征用于表现不同类型网络的构建过程。发展方向有：（1）探究其他网络是如何构建 user 与 post 之间不同类型的关系的，（2）其他更能代表网络的方法，比如网络嵌入方法。

Model-oriented:此前的方法集中在提取各种特征用于有监督分类模型中，比如朴素贝叶斯、决策树、逻辑回归、k 近邻（KNN）和支持向量机（SVM），然后选出表现最好的分类器。

可以用更复杂和高效的模型以利用提出来的特征，比如聚合方法（取各特征加权和，最优化权值）、概率方法（预测结果不是二元分类，而是概率）、集成方法（将多个弱分类器集成为一个强分类器）、投影方法（从原特征空间投影到潜在的特征空间）等。

此外已有算法大都是有监督的，需要人工给部分 fake news 上标签，因此需要发展半监督甚至无监督的算法。

Application-oriented: 可将 fake news detection 拓展到 fake news diffusion and fake news intervention 领域。

fake news diffusion 研究 fake news 传播的路径和模式，需要考虑的特性有 social dimensions, life cycle，spreader identification。

fake news intervention 研究如何通过及早截停 fake news 以降低其影响。包括隔离恶意用户，推送相关 true news 使用户免疫于 fake news

6.评价

1）优点

a）原以为 fake news detection 就是分析 news 的内容进行检测，但论文中提出将用户的 social engagements 作为辅助信息加以利用，包括检测恶意用户、利用相关 post 的立场等，有创新性的同时也具有说服力。

b）条理清晰，分类型分层次，对 fake news detection 进行了较全面的分析。

2）缺点

a）对 Social Context 进行特征学习时，把 user 和 post 的特征学习分裂开来，user 信息也可以纳入 post level 的特征学习中，反之亦然。

b）论文中否定了人工识别 fake news 的可行性，认为 news 数量太多，工作量大且准确率低，但可以缩小人工审查范围，只审查那些相关 post 大多持反对立场的 news，审查出 fake news 后加入数据库中作为有标签数据备用。

3）改进方向

a）将 user 和 post 的特征学习结合起来，比如进行 post 的特征学习时，可以认为用户位置与 news 发生地点的距离越近，其发布的 post 可信度越高；对 user 进行特征学习时，其发布的 post 越少，则可信度越低。

b）假设真新闻的传播主要是靠用户从单个可靠消息源的直接分享，而假新闻的传播则主要依托用户间的分享。则可以对 news 的传播方式进行特征学习，对传播方式进行二元分类，有助于 fake news detection。

c）true news 一般由于客观真实、证据充实，因此检测难度更小，可以逆向思维，检测真实的新闻，然后提高真实新闻在推送中的优先级，间接的降低 fake news 的曝光度。