ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202310.00935

Postprint: Technological Approaches and Related Models for Fake News Governance from an International Perspective

Authors: Li Jing

Date: 2023-10-08T00:00:00+00:00

Abstract

With the further development of artificial intelligence, social media, big data, VR/AR, and other technologies, fake news has infiltrated the vast ocean of information, becoming pervasive and difficult to distinguish, triggering a global communication crisis and social crisis, and has become a pressing issue that cannot be ignored. This article systematically reviews current technological means for combating false news, summarizes international experience, aiming to provide references and lessons for domestic efforts to crack down on false news, purify the media environment, and foster a clean and clear media space.

Full Text

Preamble

International Perspectives on Technical Approaches and Models for Governing Fake News

(China Media Technology Magazine, Beijing 100031)

Abstract: With the further development of artificial intelligence, social media, big data, VR/AR, and other technologies, fake news has infiltrated the vast ocean of information, becoming pervasive and difficult to identify. This has triggered a global communication crisis and social crisis, making it an issue that cannot be ignored. This paper reviews current technical means for governing false news and summarizes international experiences, aiming to provide references and lessons for combating fake news, purifying the media environment, and creating a clean media space in China.

Keywords: fake news governance; fact-checking; artificial intelligence; fake news detection technology; deepfake

ChinaRxiv 中X

CLC Number: G210
Document Code: A
Article ID: 1671-0134(2021)08-017-05
DOI: 10.19483/j.cnki.11-4653/n.2021.08.003

Citation Format: Li Jing. Technical Approaches and Related Models for Governing Fake News from an International Perspective [J]. China Media Technology, 2021(08): 17-21.

1. Early Methods for Combating Fake News

Early approaches to combating fake news included fact-checking, source verification, comment verification, analysis of news content characteristics, user engagement metrics, and examination of relationships between user profiles, articles, readers, and publishers. Among these, fact-checking and source verification remain the most prevalent practices. In 2010, the UK' s Channel 4 established a dedicated anti-fake news unit called "FactCheck," followed by The Guardian' s launch of the fact-checking blog "Reality Check" in 2011. The BBC created its own fact-checking team, "Reality Check," in 2015, and in 2016, the fact-checking organization Full Fact introduced its own monitoring system to track the lifecycle of each rumor. The current standard practice is to label content after verification. For instance, social platforms like Facebook and Twitter, as well as search engines such as Google News, mark suspicious content with labels like "Being fact-checked by third-party organizations," and such content is not prioritized by the platforms. Facebook has also implemented a 24-hour reporting function to identify misinformation as early as possible.

Le Monde launched the "Décodex" fact-checking database, which uses a color-coding system to help readers assess the reliability of websites. Green indicates high reliability, yellow suggests cautious reading, and red signifies that the site publishes false information or completely fabricated stories. Satirical websites are marked blue, while unverifiable sites are marked gray. Italy established the fact-checking website Pagella Politica in 2012. While these verification efforts have effectively curbed the spread of fake news, the exponential growth in information volume demands greater technical support.

2. Recent Research Directions in Fake News Governance

Statistics indicate that one-third to two-thirds of accounts spreading fake news are social bot accounts. The primary technologies employed in fake news production and dissemination include algorithmic techniques, deepfake technology, and machine learning. Algorithmic recommendations guide the flow of traffic, while user-generated misinformation remains a persistent problem. Deepfake technology, combined with machine deep learning and automated decision-making, enables the automatic generation of images and videos, making news produc-

tion and distribution increasingly mechanized and intelligent, which enhances the manipulability of information. As artificial intelligence becomes more sophisticated, news created and distributed by bots has become a hotbed for fake news. Manual verification, already stretched thin, can no longer effectively combat the massive and increasingly covert wave of fake news, prompting research to shift toward automated technologies.

Le Monde' s fact-checking team has been collaborating with data scientists to explore how automated technologies can detect fake news in real time. In 2019, researchers from the University of Washington and the Allen Institute for Artificial Intelligence proposed a detection model called Grover, which learned from 120 gigabytes of real news articles written by 5,000 different media outlets on Google News. Grover achieved 92% accuracy in distinguishing between human-written and AI-generated stories, surpassing the previous best fake news detector' s accuracy of 73%. Grover' s effectiveness stems from its own proficiency in content generation, embodying the principle of using technology to combat "technology-made" fakes—currently our best defense.

3. Technical Approaches to Fake News Governance

The specific technical path to governing fake news lies in better understanding patterns, establishing rules, and developing models that provide correct guidance to algorithms, enabling artificial intelligence to more effectively identify false and vulgar content and control its spread. False news detection technology leads this effort, employing intelligent processing methods such as natural language processing, social mining, and cross-modal analysis to discover and utilize the intrinsic characteristics, generation mechanisms, and propagation patterns of information to identify and intervene in fake news dissemination. The technologies employed include traditional machine learning algorithms (such as logistic regression, support vector machines, and random forest), deep learning (including convolutional and recurrent neural networks), and other models (matrix factorization and Bayesian inference). The following sections introduce the latest detection models from the perspectives of news content features, predictive detection, and explainable detection.

3.1 Content-Based Fake News Detection Technology

Based on different content features, false information detection can be categorized into text-based, image-based, and multimodal detection.

3.1.1 Text-Based Detection Text-based detection is the most commonly used method for false information detection. Most studies utilize both the textual content itself and social context generated during propagation (based on user behavior credibility, propagation networks, or semantic and sentiment metrics). Research has focused on the specific linguistic styles of fake news; for example, the paper *Capturing the Style of Fake News* uses document style to

classify data from different sources and score news credibility for detection purposes. Early approaches extracted linguistic and topic features, while recent methods employ deep models to automatically learn high-level data features. Social context-based methods primarily include approaches based on user behavior credibility and propagation network structures. The paper "A Joint Topic and Sentiment Pre-training Method for Fake Review Detection" extracts semantic and emotional context features from reviews for joint training and optimization.

3.1.2 Image-Based Detection Research on graph-based anomaly detection first emerged in 2003. Graph anomaly detection involves finding structures (including nodes, edges, or subgraphs) containing unfamiliar or anomalous patterns in large graphs or massive graph databases. Research objects can be divided into static and dynamic graphs. Anomalies in static graphs typically refer to nodes, edges, or subgraphs with significant deviations. Dynamic graphs, which evolve over time with the addition and deletion of nodes and edges, typically exhibit anomalies as the top-k nodes, edges, or subgraphs that cause changes or events. Graph anomaly detection is often based on structural information (including node-to-node, node-to-subgraph, and subgraph-to-subgraph anomalies), subspace selection to detect anomalies in subsets of node features, and statistical analysis based on probability statistics to obtain graph statistics. Dynamic graphs often first obtain graph summaries, then use clustering and anomaly detection to identify anomalies in the summaries. Meanwhile, deep learning methods for image detection continue to be explored.

In industry, Microsoft employs Face X-ray technology for image detection, proposing a universal method for detecting synthetic images generated by different models. The core is learning the boundaries of face-swapping, demonstrating excellent generalization performance. The University of Chicago' s Fawkes technology provides pixel-level protection for private photos that is invisible to the human eye, preventing users from being detected and tracked by unknown third-party facial recognition models. Facebook launched an AI system called "Rosetta" to help computers understand and analyze the massive number of images posted on its platform daily.

3.1.3 Audio-Based Detection Common voice forgery methods include voice imitation, recording replay, voice synthesis, and voice conversion. Detection primarily employs convolutional neural networks and their variants. Research focuses on two directions: (1) Utilizing acoustic features that can distinguish real from forged speech, such as Todisco et al.' s application of constant Q cepstral coefficients (CQCC) to voice anti-spoofing, and Sahidullah et al.' s proposal of linear frequency cepstral coefficients (LFCC), which uses linear filtering instead of mel-scale filtering to focus more on high-frequency features. (2) Designing classification models that can learn discriminative representations of genuine and fake speech. The Gaussian mixture model (GMM) was previously the most commonly used classification model. With the advancement of deep learning,

convolutional neural networks (CNN) perform better than GMM alone. For example, the lightweight convolutional neural network (LCNN) with max feature map (MFM) activation uses competitive learning to separate not only noise from information signals but also to perform feature selection. Both approaches have proven effective, demonstrating the importance of using appropriate front-end acoustic features and deep learning models.

3.1.4 Multimodal Detection Multimodal false news detection is more challenging and represents a current research hotspot. Text, images, and video together constitute multimodal information that mutually supports and corroborates, increasing credibility while also enhancing the difficulty of authenticity verification. Examples include Deepfakes for face-swapping, real-time motion capture technology developed by German and American researchers that can convert any actor' s facial expressions into video clips of political figures like Trump and Putin, and NVIDIA' s new StyleGAN, whose ability to disentangle image attributes (styles) has spawned numerous image editing applications such as the popular creative tool Artbreeder (https://www.artbreeder.com) and HKUST' s InterFaceGAN, which proposes a latent space structure GAN generation space method that can be generalized to all GAN-generated face sample spaces, including attribute editing and style conversion. The adoption of these new technologies has increased the difficulty of multimodal fake news verification.

Current multimodal detection methods generally employ generic recurrent neural networks (RNN) and convolutional neural networks (CNN) to capture the characteristics of false news in textual and visual modalities. Specific detection model research includes: using text and visual feature extractors with joint training; employing variational autoencoders to learn shared representations of text and images, reconstructing text and images from shared latent representations to capture associations between modalities; calculating similarity between textual and visual representations after separate learning to identify "mismatches" between patterns; adding three types of modules (text, image, and user profile) to models and incorporating user comment sentiment to determine article/post authenticity; and utilizing external knowledge graphs to learn foundational knowledge combined with textual and visual features for detection. In industry, the San Francisco-based AI Foundation, established in 2017, developed the Reality Defender system, which scans images, videos, and other media content using AI-driven analysis to help people identify AI algorithm-generated content and detect potential fake news.

3.2 Explainable Detection

Explainable detection is an emerging area in false news detection research. The public questions whether certain information is flagged as suspicious or false for political or economic purposes, and why it receives particular labels and on what basis. Therefore, explaining why an article receives a certain label has become

a new research direction. The paper *dEFEND: Explainable Fake News Detection* proposes an explainable fake news detection model that separately encodes news content and user comment components, using hierarchical attention mechanisms and sentence-comment co-attention subnetworks to capture associations between content and comments. The model ultimately identifies explainable top-k sentences and user comments worth examining, demonstrating good detection performance. Other papers such as "Propagation2Vec: Embedding Partial Propagation Networks for Explainable Early Fake News Detection" have also conducted related work, indicating broad research space in this area.

3.3 Predictive Detection

Rumor propagation patterns often differ significantly from those of real information. Achieving predictable rumor detection can yield twice the result with half the effort. Soroush et al.' s analysis of rumor propagation patterns on Twitter revealed that rumors spread much wider than real messages. While real messages might have up to 1,000 participants at any single level of forwarding, rumors can reach tens of thousands. Debunking simply cannot keep pace with propagation speed. Therefore, blocking fake news at the source and in the early stages is also a current research direction.

The paper *Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks* recently proposed a "Bi-GCN" model suitable for "early rumor detection." Unlike previous methods, this model considers both top-down rumor propagation structures and bottom-up rumor dispersion structures from different communities. It also employs "enhancement of root post features," specifically by concatenating the root post' s hidden feature representation from the previous layer with each node' s hidden representation at the current layer in each graph convolutional layer (GCL) as the node' s final hidden representation. This approach enhances the influence of the rumor' s root post on learning other post node representations, helping the model learn more informative node representations for rumor detection. This is also the "first" model to use "GCN-based methods" for rumor detection tasks.

Additionally, propagation structure-based detection methods represent another hot research direction. UK-based tech company Fabula AI uses geometric deep learning to focus on how information spreads across social networks and who is spreading it, classifying content credibility and providing scores based on the degree of truthfulness. This enables faster and more accurate fake news detection within a short time after content publication. Furthermore, while AI is currently mainstream for detecting fake news, blockchain-based models such as Userfeeds and PressCoin are also worth exploring.

4. Current Dilemmas in Fake News Governance

Reviewing the current state of technical development for fake news governance, we can see that significant progress has been made in building models to detect

fake news using artificial intelligence. However, most existing research does not use real information propagation data from actual social networks. The datasets required by models are relatively scarce. Although some studies are based on real social network topologies, the specific information propagation and suppression processes are entirely simulated, lacking authenticity. This has resulted in academic models remaining at the conceptual stage, limited to publication in papers without connection to practice. It is hoped that those interested in this field, whether from media, technology companies, or investment firms, will increase their attention to outstanding papers and strive to translate academic achievements into industry practice to achieve greater impact.

5. Future Directions and Trends in Fake News Governance

The truth and intent of any information cannot be assessed by computers alone. While technical means are effective in the short term, they treat symptoms rather than root causes and represent temporary measures. Fake news will inevitably coexist with technology and humanity in the long term. To achieve long-term effective governance of fake news, we must still rely on cooperation between humans and technology, combined with legal constraints and the cultivation of citizens' moral character and media literacy. This is the ultimate path to governing fake news.

Specifically, in improving citizens' media literacy, education from primary school through university must take responsibility for helping young people develop the ability to identify fake news, offering relevant media literacy courses to cultivate critical thinking. Universities should also make news verification and fact-checking foundational disciplines, moving from skill-based training in information verification to improving self-awareness and overcoming cognitive biases, ultimately achieving the goal of "rumors stop with the wise." For everyone, the task of governing fake news remains arduous and the road ahead is long.

References

[1] Amrita, Bhattacharjee, Shu Kai, Gao Min, Liu Huan. Disinformation in Online Information Ecosystems: Detection, Mitigation, and Challenges [J]. Journal of Computer Research and Development, 2021(7): 1353-1365.

[2] Zhang Jianzhong. Combating Fake News: Innovations and Practices in European Countries in the "Post-Truth" Era [J]. Journalism and Mass Communication, 2017(06): 95-101.

[3] GILL S. Why Combatting Fake News Requires People And Technology—Working Together [EB/OL]. (2019-09-15). https://www.knightfoundation.org/articles/why-combatting-fake-news-requires-people-and-technology-working-together.

[4] Huashang Network. Artificial Intelligence is the "Nemesis" of Fake News [EB/OL]. https://www.hsw.cn/a/129892845_{119659}.

[5] Jiang Mengting. Application of Fake News Detection Technology [J]. Network Security Technology and Application, 2021(04): 54-55.

[6] Guo Bin, Ding Yasan, Yao Lina, et al. The Future of False Information Detection on Social Media: New Perspectives and Trends [J]. ACM Computing Surveys, 2020(4): 68.

[7] Castillo C, Mendoza M, Poblete B. Information Credibility on Twitter [C]//Proc of the WebConf 2020. New York: ACM, 2011: 675-684.

[8] Qazvinian V, Rosengren E, Radev D, et al. Rumor has it: Identifying Misinformation in Microblogs [C]//Proc of the 2011 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1589-1599.

[9] Pérez-Rosas V, Kleinberg B, Lefevre A, et al. Automatic Detection of Fake News [EB/OL]. (2017-08-23) [2020-10-08]. https://arxiv.org/pdf/1708.07104.pdf.

[10] Ma Jing, Gao Wei, Mitra P, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks [C]//Proc of the 25th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 3818-3824.

[11] Liu Bo, Li Yang, Meng Qing, et al. Analysis and Evaluation of Social Media Content Credibility [J]. Journal of Computer Research and Development, 2019(9): 1849-1858.

[12] Jin Zhiwei, Cao Juan, Jiang Yugang, et al. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model [C]//Proc of the 2014 IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2014: 230-239.

[13] Jin Zhiwei, Cao Juan, Zhang Yongdong, et al. News Certification by Exploiting Conflicting Social Viewpoints in Microblogs [C]//Proc of the 30th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 2972-2978.

[14] Shu Kai, Wang Suhang, Liu Huan. Beyond News Contents: The Role of Social Context for Fake News Detection [C]//Proc of the 20th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2019: 312-320.

[15] Ma Jing, Gao Wei, Wong KF. Rumor Detection on Twitter with Tree-Structured Recursive Neural Networks [C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 1980-1989.

[16] Hu Songlin, Zhao Jun, Tang Jie, Qin Bing, Shi Chuan, Yan Shuicheng. Preface [J]. Journal of Computer Research and Development, 2021, 58(07): 1351-1352.

[17] Noble CC, Cook DJ. Graph-Based Anomaly Detection [C]//Proc of the 9th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2003: 631-636.

[18] Seo J, Mendelevitch O. Identifying Frauds and Anomalies in Medicare-B dataset [C]//Proc of the 39th Annual Int Conf of the IEEE Engineering in

Medicine and Biology Society. Piscataway, NJ: IEEE, 2017: 3664-3667.

[19] Colladon AF, Remondi E. Using Social Network Analysis to Prevent Money Laundering [J]. Expert Systems with Applications, 2017, 67: 49-58.

[20] Manjunatha HC, Mohanasundaram R. BRNADS: Big Data Real-Time Node Anomaly Detection in Social Networks [C]//Proc of the 2nd Int Conf on Inventive Systems and Control. Piscataway, NJ: IEEE, 2018: 929-932.

[21] Sánchez PI, Müller E, Laforet F, et al. Statistical Selection of Congruent Subspaces for Mining Attributed Graphs [C]//Proc of the 13th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2013: 647-656.

[22] Sánchez PI, Müller E, Irmler O, et al. Local Context Selection for Outlier Ranking in Graphs with Multiple Numeric Node Attributes [C]//Proc of the 26th Int Conf on Scientific and Statistical Database Management. Piscataway, NJ: IEEE, 2014: 16.

[23] Perozzi B, Akoglu L, Sánchez PI, et al. Focused Clustering and Outlier Detection in Large Attributed Graphs [C]//Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1346-1355.

[24] Dai Hanbo, Zhu Feida, Lim EP, et al. Detecting Anomalies in Bipartite Graphs with Mutual Dependency Principles [C]//Proc of the IEEE 12th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2012: 171-180.

[25] Tsang S, Koh Y S, Dobbie G, et al. SPAN: Finding Collaborative Frauds in Online Auctions [J]. Knowledge-Based Systems, 2014, 71: 389-408.

[26] Shehnepoor S, Salehi M, Farahbakhsh R, et al. Netspam: A Network-Based Spam Detection Framework for Reviews in Online Social Media [J]. IEEE Transactions on Information Forensics and Security, 2017(7): 1585-1595.

[27] Carvalho LFM, Teixeira CHC, Meira W, et al. Provider-Consumer Anomaly Detection for Healthcare Systems [C]//Proc of the IEEE Int Conf on Healthcare Informatics. Piscataway, NJ: IEEE, 2017: 229-238.

[28] Ranshous S, Shen Shitian, Koutra D, et al. Anomaly Detection in Dynamic Networks: A Survey [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2015, 7(3): 223-247.

[29] Manzoor E, Milajerdi SM, Akoglu L. Fast Memory-Efficient Anomaly Detection in Streaming Heterogeneous Graphs [C]//Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1265-1274.

[30] Chen Bofeng, Li Jingdong, Lu Xingjian, Sha Chaofeng, Wang Xiaoling, Zhang Ji. A Survey of Deep Learning-Based Graph Anomaly Detection Techniques [J]. Journal of Computer Research and Development, 2021(7): 1436-1455.

[31] Wu Zhizheng, Evans N, Kinnunen T, et al. Spoofing and Countermeasures for Speaker Verification: A Survey [J]. Speech Communication, 2015: 130-153.

[32] Wang Chenglong, Yi Jiangyan, Tao Jianhua, Ma Haoxin, Tian Zhengkun, Fu Ruibo. Voice Forgery Detection Based on Global-Time-Frequency Attention Networks [J]. Journal of Computer Research and Development, 2021(7): 1466-1475.

[33] Qi Peng, Cao Juan, Sheng Qiang. Semantic-Enhanced Multimodal Fake News Detection [J]. Journal of Computer Research and Development, 2021(7): 1456-1465.

[34] Amrita, Bhattacharjee, Shu Kai, Gao Min, Liu Huan. Disinformation in Online Information Ecosystems: Detection, Mitigation, and Challenges [J]. Journal of Computer Research and Development, 2021(7): 1353-1365.

[35] Wu Guozhong, Li Tairu. Combating Fake News with Blockchain Technology —An Introduction to Userfeeds and PressCoin Models [J]. News Front, 2018(13): 71-73.

Author Bio: Li Jing (1983-), female, from Nanyang, Henan, is the Editorial Director of *China Media Technology* Magazine. Research interests: news communication.

(Responsible Editor: Chen Xuguan)

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.