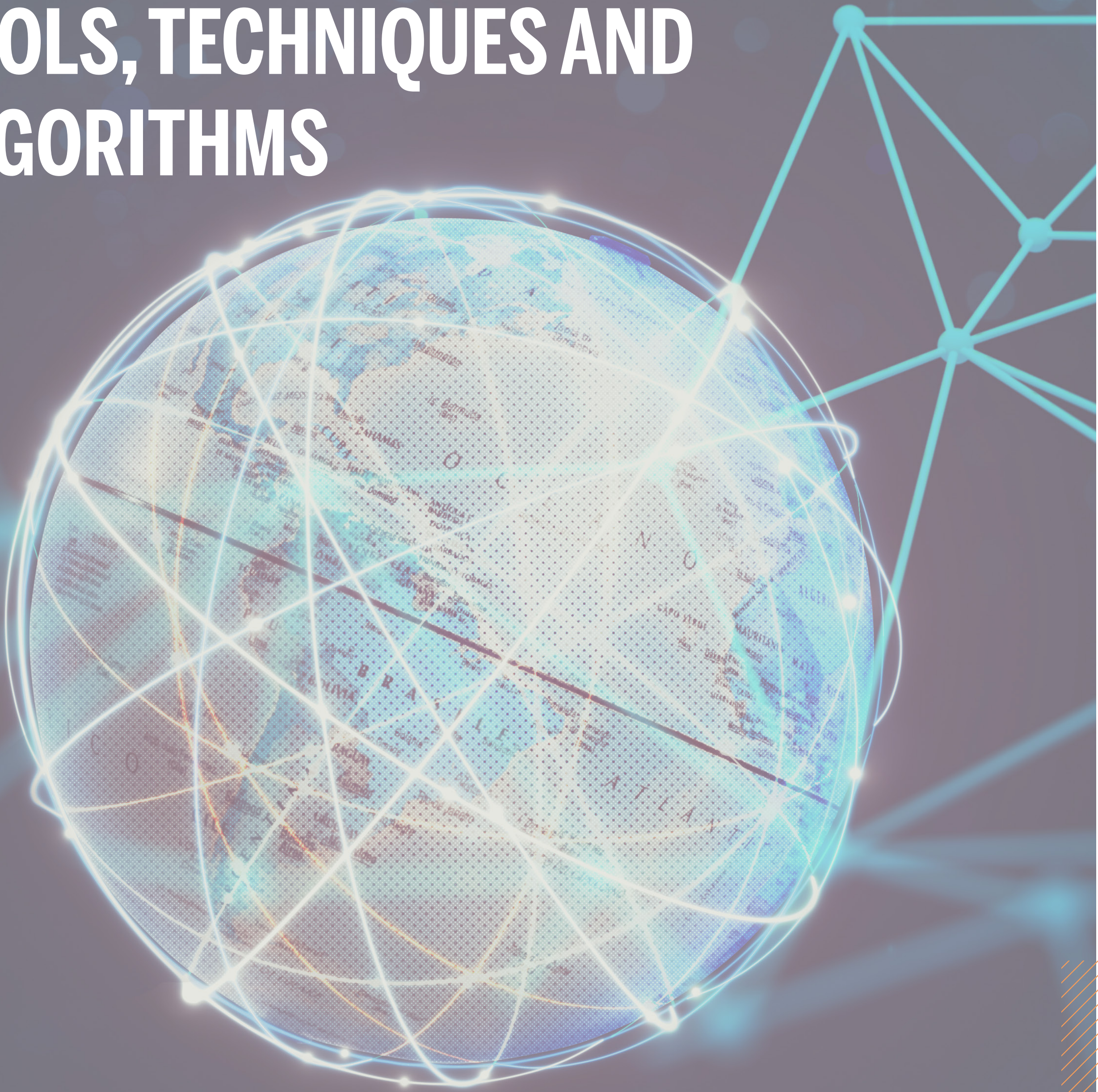
edam
Centre for Economics
and Foreign Policy Studies

CYBER GOVERNANCE & DIGITAL DEMOCRACY PROGRAMME

AUGUST 2023

EMERGING TECHNOLOGIES AND AUTOMATED FACT-CHECKING: TOOLS, TECHNIQUES AND ALGORITHMS



Akın Ünver, Associate Professor, Ozyegin University

INTRODUCTION

With the rise of digital platforms and the exponential growth of online information, fact-checking has undergone a significant transformation, relying heavily on technological advancements. Artificial Intelligence (AI) and Natural Language Processing (NLP) have revolutionized fact-checking processes, enabling the analysis of vast amounts of textual data in real-time. Increasingly more complex algorithms can identify patterns of misinformation, detect misleading claims, and streamline fact-checking efforts, allowing fact-checkers to keep up with the sheer volume of information circulating online. Automated fact-checking tools that are becoming increasingly more popular among fact-checkers often rely on the very technologies that aid in the creation and dissemination of disinformation and information manipulation. These tools use machine learning algorithms to assess the veracity of claims and provide fact-checkers with near-instant results, although the accuracy of most methods have long been imperfect, and defined in terms of confidence intervals. By automating the fact-checking process, these tools promise greater efficiency and scalability, facilitating the dissemination of accurate information to counter falsehoods more effectively.

Blockchain technology has also entered the fact-checking arena, offering a means of enhancing trust and transparency in information verification. By timestamping and securing verified information, blockchain ensures an immutable record of claims and their accuracy. The decentralized nature of blockchain provides an opportunity to counter information manipulation and create a reliable source of truth. Deepfake detection technologies, while originally developed to combat the rise of deepfake content, have found relevance in fact-checking as well. Fact-checkers are increasingly concerned about the use of manipulated audio and video content to deceive audiences, making deepfake detection a crucial component of their toolkit.

Despite the potential benefits of new technologies in fact-checking, debates abound concerning its efficacy, challenges, and potential biases. One major debate revolves around the ability of fact-checkers to keep pace with the rapid spread of misinformation in the digital age. The sheer volume of content generated daily poses significant challenges to fact-checkers, as some false claims can gain traction before they can be thoroughly debunked. Another contentious issue relates to the perceived biases of fact-checkers and fact-checking organizations. Some critics argue that fact-checkers may have their own ideological leanings, potentially influencing their assessments of claims. Debates also emerge regarding the selection of claims to fact-check, as prioritization may inadvertently reflect the fact-checker's own biases or interests.

The tension between fact-checking and free speech is a persistent debate that lies at the heart of information verification efforts. While fact-checking aims to counter misinformation, some argue that overly aggressive fact-checking may lead to accusations of censorship and infringement on free expression. Striking a balance between countering disinformation and upholding democratic values is a complex challenge faced by fact-checkers. Moreover, the global nature of misinformation and disinformation campaigns raises questions about cross-border collaboration among fact-checkers. With information manipulation often originating from state-sponsored actors spanning multiple countries, collaborative efforts between fact-checking organizations worldwide are vital in effectively combating foreign information interference.

NEW TECHNOLGIES OF FACT-CHECKING ADVANCED DATA ANALYSIS AND ARTIFICIAL INTELLIGENCE

Several pioneering organizations are reshaping the landscape of fact-checking through the integration of advanced technologies. Full Fact, based in the UK, stands at the vanguard, blending the capabilities of natural language processing and machine learning to discern truth from falsehood in news reporting. Their ongoing efforts to automate parts of the fact-checking process underscore their commitment to expanding their reach and efficacy. Snopes, one of the trailblazers in the realm of fact-checking, is also bolstering their meticulous verification process with the insights offered by machine learning. This infusion of technology amplifies their capacity to sift through social media noise and unmask fabrications. Simultaneously, FactCheck.org, a project of the Annenberg Public Policy Center, is employing sophisticated data analytics to track the dissemination of misinformation across social media networks. Their work provides a valuable understanding of the dynamics of false information spread.

The Agence France-Presse has mobilized its own fact-checking division, AFP Fact Check, that harnesses digital forensics. Utilizing tools like reverse image searches and geolocation, they scrutinize the authenticity of circulating images and videos. The Associated Press also brings digital tools into play, validating social media content and visual media. They employ AI-driven analytics to study social media patterns, providing a radar for potential false claims that require a closer examination. In Argentina, Chequeado leverages machine learning and automation to keep pace with live speeches, cross-referencing statements with their pre-existing fact-checked database. This real-time verification process exemplifies the potential of technology in challenging inaccuracies instantly. While these organizations are harnessing the power of technology in their fact-checking efforts, the essence of human discernment continues to be integral. They illustrate how technology, while powerful, serves as a complementary tool in the relentless pursuit of truth, rather than a replacement for human judgement.

Here are some of the most-frequently used emerging technologies within the global fact-checking ecosystem:

1. Automated Fact-Checking Tools:

AI-powered automated fact-checking tools are becoming increasingly prevalent in the fact-checking ecosystem. These tools use natural language processing (NLP) algorithms to analyze textual content, identify claims that require verification, and cross-reference them with reputable sources and databases. As a result, fact-checkers can streamline their workflow, identify potentially misleading claims more

[1] Nic Newman et al., "Reuters Institute Digital News Report 2023," 2023. 11.

efficiently, and focus their efforts on the most critical information. Developing and employing automated fact-checking tools can significantly enhance the speed and accuracy of fact-checking processes. Natural Language Processing (NLP) algorithms can help fact-checkers analyze textual content and identify factual inaccuracies or false claims. NLP is a subfield of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. By leveraging NLP technologies, fact-checkers can process large volumes of textual data, identify claims that require verification, and assess the accuracy of information with greater efficiency and accuracy. This extended analysis delves into the technicalities of NLP and its impact on fact-checking and information verification.

The essential stages of the automated fact-checking pipeline, namely claim detection and extraction, are deeply intertwined with the use of artificial intelligence and natural language processing (NLP). They are a combination of multiple techniques each playing a unique role in parsing through text and identifying significant elements. One such technique is Information Extraction, which pulls structured information from unstructured text data. This process contains two crucial components. First, there's Named Entity Recognition (NER), which identifies distinct entities like people, organizations, and dates within the text. Second, Relationship Extraction (RE) is applied to determine the relationships between the identified entities.[1] Together, these processes provide a rich foundation for extracting claims.

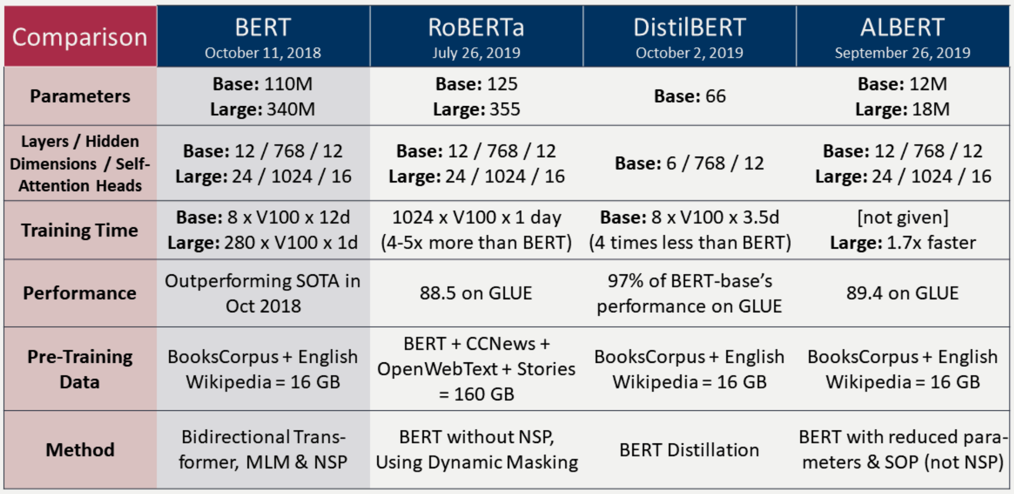
On the other hand, machine learning finds its use in text classification, where it scans and labels sentences or sections of text as factual claims requiring verification. This task usually involves supervised learning models, trained on manually annotated data that already identifies factual claims, and then these models are deployed to detect similar claims in new data sets. The domain of NLP presents a specialized subfield, argument mining, that focuses on identifying argumentative structures within the text, such as the claims and evidence we aim to extract. Techniques derived from this subfield are crucial in pinpointing and extracting these claims from a body of text. Deep learning models, especially transformers like BERT, GPT, and RoBERTa, have shown promising results in NLP tasks and can be effectively used for claim detection and extraction.[2]

Complementing these sophisticated techniques, there are more straightforward approaches such as rule-based systems, which are manually programmed to follow linguistic rules or heuristics to identify claims. These systems scan for specific keywords, signaling the presence of factual claims. The complex landscape of NLP also incorporates tasks like Semantic Role Labeling (SRL), which identifies the semantic roles and relationships of words in a sentence, assisting in the

[1] Bose, Priyankar, Sriram Srinivasan, William C. Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. "A survey on recent named entity recognition and relationship extraction techniques on clinical texts." Applied Sciences 11, no. 18 (2021): 8319.

[2] Casillas, Ramón, Helena Gómez-Adorno, Victor Lomas-Barrie, and Orlando Ramos-Flores. "Automatic Fact Checking Using an Interpretable Bert-Based Architecture on COVID-19 Claims." Applied Sciences 12, no. 20 (2022): 10644.

recognition of claim structures.[3] Researchers have made strides in this area, as exemplified by the development of ClaimBuster, a unique algorithm using supervised machine learning to detect check-worthy factual claims.[4] It inspects features such as part of speech tags, the presence of numbers, and named entities. Despite the advancements in these methods, human intervention is indispensable, especially for model training, supervision, and validation. This necessity arises from the complexity and variability of human language. The goal of developing a fully automated and foolproof claim detection system is still a work in progress, warranting continued research efforts.



Comparison	BERT October 11, 2018	RoBERTa July 26, 2019	DistilBERT October 2, 2019	ALBERT September 26, 2019
Parameters	**Base:** 110M **Large:** 340M	**Base:** 125 **Large:** 355	**Base:** 66	**Base:** 12M **Large:** 18M
Layers / Hidden Dimensions / Self-Attention Heads	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 6 / 768 / 12	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16
Training Time	**Base:** 8 x V100 x 12d **Large:** 280 x V100 x 1d	1024 x V100 x 1 day (4-5x more than BERT)	**Base:** 8 x V100 x 3.5d (4 times less than BERT)	[not given] **Large:** 1.7x faster
Performance	Outperforming SOTA in Oct 2018	88.5 on GLUE	97% of BERT-base's performance on GLUE	89.4 on GLUE
Pre-Training Data	BooksCorpus + English Wikipedia = 16 GB	BERT + CCNews + OpenWebText + Stories = 160 GB	BooksCorpus + English Wikipedia = 16 GB	BooksCorpus + English Wikipedia = 16 GB
Method	Bidirectional Transformer, MLM & NSP	BERT without NSP, Using Dynamic Masking	BERT Distillation	BERT with reduced parameters & SOP (not NSP)

A comparison of main text transformer models. Source: Siddharth Godbole, Karolina Grubinska & Olivia Kelnreiter "Economic Uncertainty Identification Using Transformers - Improving Current Methods" (https://humboldt-wi.github.io/blog/research/information_systems_1920/uncertainty_identification_transformers/)

2. Sentiment Analysis and Stance Detection:

Sentiment analysis and stance detection play important roles in automated fact-checking by helping to understand the attitude, emotions, and perspectives expressed in textual content. Here's an overview of some advanced methods and techniques in these areas:

Analyzing sentiments and detecting stances in a textual context relies heavily on sophisticated techniques ranging from deep learning models to ensemble methods. These innovative procedures have significantly improved the potential of these tasks, particularly when paired with high-quality, diverse training data and domain-specific knowledge. Deep Learning Models, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and notably, Transformer-based models such as BERT, have demonstrated their efficacy in

[3] Màrquez, Lluís, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. "Semantic role labeling: an introduction to the special issue." Computational linguistics 34, no. 2 (2008): 145-159.

[4] Hassan, Naeemul, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan et al. "Claimbuster: The first-ever end-to-end fact-checking system." Proceedings of the VLDB Endowment 10, no. 12 (2017): 1945-1948.

sentiment analysis.[5] These models can be customized using sentiment-tagged datasets, enabling them to accurately predict the sentiment of any given text.

Contextual Embeddings also play an essential role in sentiment analysis. While pre-trained word embeddings like Word2Vec and GloVe have seen extensive use, contextual embeddings from models like ELMo, GPT, and BERT have proven to be even more effective due to their ability to capture the contextual meaning of words.[6] Aspect-Based Sentiment Analysis (ABSA) is a more specialized approach that targets sentiment at the aspect or entity level, rather than just at the overall document level. ABSA has proven particularly useful when analyzing feedback or reviews that express opinions about specific aspects of a product or service.[7]

Multimodal Sentiment Analysis allows for a comprehensive understanding of sentiment by considering not only the text but also other modalities like images, videos, or audio.[8] Attention Mechanisms, on the other hand, enable models to concentrate on the most pertinent parts of a text when predicting sentiment, enhancing accuracy, particularly for longer documents. Stance detection, a task that can be structured as a classification task, utilizes machine learning models such as Support Vector Machines, Random Forests, or deep learning models. These models are trained on labeled datasets to predict stances.[9]

Transfer learning, a technique similar to sentiment analysis, can also be applied to stance detection tasks using pre-trained language models like BERT. Fine-tuning these models on stance-labeled data often results in better performance. Similarly, neural network architectures like attention mechanisms and recurrent neural networks are employed in stance detection to gain a deeper understanding of the context and relationships between words. When it comes to fact-checking, identifying the stance of a claim is crucial. Advanced techniques are employed, leveraging fact-checking datasets and domain-specific knowledge, to classify the stance of the claims. For instance, in the 2023 Indian General Elections, these techniques were used to fact-check various political claims.[10]

Finally, ensemble methods are used to improve the overall performance and robustness of stance detection. By combining the predictions of multiple stance detection models using techniques such as voting ensemble or weighted averaging, the results are further optimized.[11] It's crucial to remember that while these methods have shown promising results, automated sentiment analysis and

[5] Balakrishnan, Vimala, Zhongliang Shi, Chuan Liang Law, Regine Lim, Lee Leng Teh, Yue Fan, and Jeyarani Periasamy. "A Comprehensive Analysis of Transformer-Deep Neural Network Models in Twitter Disaster Detection." Mathematics 10, no. 24 (2022): 4664.

[6] Dharma, Eddy Muntina, F. Lumban Gaol, H. Leslie, H. S. Warnars, and B. Soewito. "The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification." J Theor Appl Inf Technol 100, no. 2 (2022): 349-359.

[7] Wang, Jie, Bingxin Xu, and Yujie Zu. "Deep learning for aspect-based sentiment analysis." In 2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), pp. 267-271. IEEE, 2021.

[8] Zadeh, Amir, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. "Tensor fusion network for multimodal sentiment analysis." arXiv preprint arXiv:1707.07250 (2017).

[9] Al Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine El Kadiri. "Random forest and support vector machine based hybrid approach to sentiment analysis." Procedia Computer Science 127 (2018): 511-520.

[10] Santos, Fátima C. Carrilho. 2023. "Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis" Journalism and Media 4, no. 2: 679-687. https://doi.org/10.3390/journalmedia4020043

[11] Siddiqua, Umme Aymun, Abu Nowshed Chy, and Masaki Aono. "Tweet stance detection using an attention based neural ensemble model." In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers), pp. 1868-1873. 2019.

stance detection remain challenging tasks due to the complexities of language, context, and subjectivity. Much of their performance is tied to the quality and diversity of the training data and the domain specificity of the fact-checking task. Like any AI system, these techniques require validation and human oversight to ensure reliable results.

3. Real-Time Verification and Debunking:

The swift and dynamic nature of online information dissemination necessitates the need for real-time verification and debunking as core components of automated fact-checking. A diverse range of advanced methods and techniques have been developed to aid this important work. Automated systems have the ability to crawl and monitor news websites, social media platforms, and other online sources continuously, identifying new claims and information being shared in real time. APIs and web scraping tools are employed to efficiently collect data from various sources. For instance, during the 2022 Australian bushfires, Australian Associated Press deployed such an API-based system to sift through the plethora of information and misinformation being circulated in real time.[12]

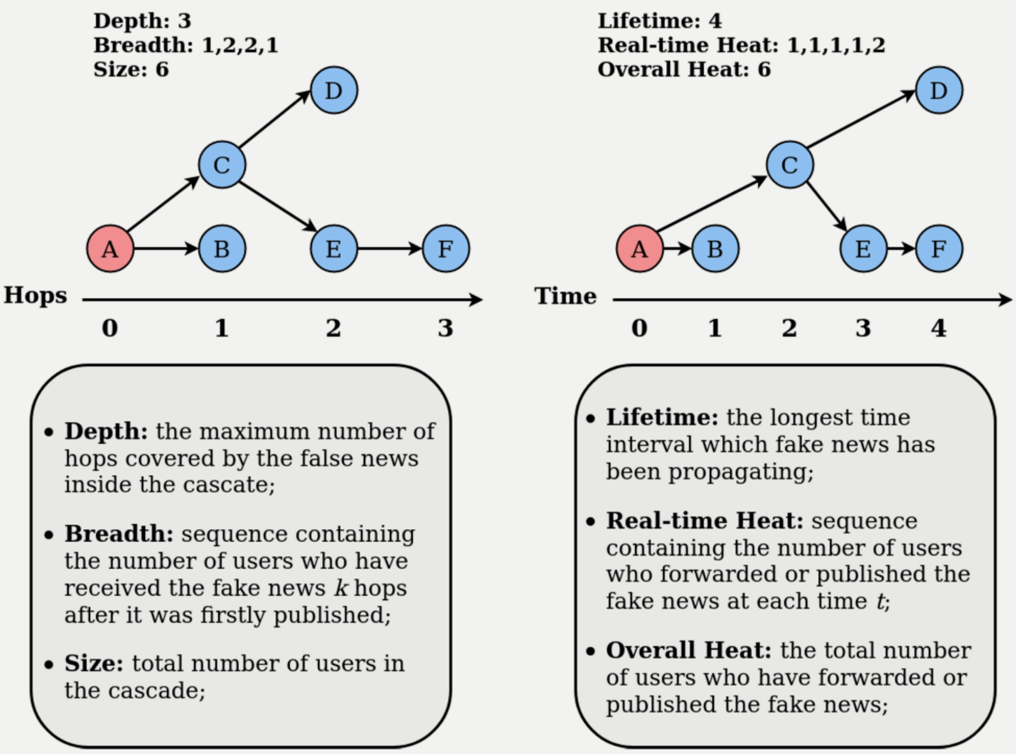
Fact-checking also involves prioritizing claims for verification. Advanced algorithms decide this based on factors such as virality, potential impact, and the credibility of the source. Machine learning models use historical data to determine which claims are more likely to be misinformation or require urgent fact-checking. This process involves analyzing the source's reputation, past accuracy, and bias. Sources are assigned trustworthiness scores, which can significantly assist fact-checking decisions. For instance, in Japan, an AI-powered fact-checking tool was used during the 2023 Tokyo earthquake to assess the credibility of various online sources reporting on the incident.[13] In the aftermath of the devastating Tokyo earthquake in 2023, there was a surge of online information as various sources began reporting on the event. This inundation of news ranged from accurate updates to rampant speculation, misinformation, and even malicious disinformation. Consequently, it became challenging for individuals, both locally and globally, to discern fact from fiction. In response to this predicament, Japanese technology experts deployed an AI-powered fact-checking tool designed to assess the credibility of the plethora of online sources reporting on the incident.[14] The tool had been developed by a team of researchers at the University of Tokyo and was still in its testing phase when the earthquake hit. It employs a deep learning algorithm that cross-references the information from online sources with multiple credible databases, including verified

[12] Brookes, Stephanie, and Lisa Waller. "Communities of practice in the production and resourcing of fact-checking." Journalism (2022): 14648849221078465.

[13] False rumors spread again following recent earthquake off northeast Japan. The Mainichi. 15 February 2023. https://mainichi.jp/english/articles/20210215/p2a/00m/0na/016000c

[14] Bazarkina, Darya, Yury Kolotaev, Evgeny Pashentsev, and Daria Matyashova. "Current and Potential Malicious Use of Artificial Intelligence Threats in the Psychological Domain: The Case of Japan." In The Palgrave Handbook of Malicious Use of AI and Psychological Security, pp. 419-451. Cham: Springer International Publishing, 2023.

news outlets, official government statements, and seismic research institutions. By doing so, it can analyze the probability of a piece of information being true.



- **Depth:** the maximum number of hops covered by the false news inside the cascade;
- **Breadth:** sequence containing the number of users who have received the fake news *k* hops after it was firstly published;
- **Size:** total number of users in the cascade;

- **Lifetime:** the longest time interval which fake news has been propagating;
- **Real-time Heat:** sequence containing the number of users who forwarded or published the fake news at each time *t*;
- **Overall Heat:** the total number of users who have forwarded or published the fake news;

Cascades of fake news in a hop-based (left) and time-based (right) perspective. The root node A, in both perspectives, represents the first user to publish or create the fake news, while the remaining nodes represent users that actively forward or share the fake content. Source: de Oliveira, Nicollas R., Pedro S. Pisa, Martin Andreoni Lopez, Dianne Scherly V. de Medeiros, and Diogo M. F. Mattos. 2021. "Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges" Information 12, no. 1: 38. https://doi.org/10.3390/info12010038

In 2023, during a series of protests in Hong Kong, the world witnessed the application of sophisticated AI techniques in fact-checking multimedia content. At the core of this technology were knowledge graphs - vast databases of verified information. These databases were continuously updated, capturing a vast range of topics, events, and previously assessed claims. Fact-checking AI systems leveraged these knowledge graphs to cross-reference new claims swiftly. They were able to rapidly flag content that had already been verified or debunked in the past, thereby enhancing efficiency. In tandem with this, the AI systems deployed NLP and machine learning models to fact-check textual claims in real time. This wasn't merely a word-by-word analysis. The models evaluated the text in its entirety, interpreting the context in which claims were made. The AI was capable of understanding nuances, including sarcasm, cultural references, and localized jargon, leading to a nuanced and comprehensive assessment of the veracity of the information. The fact-checking process was not limited to textual content. In the context of the Hong Kong protests, where the authenticity of numerous images and videos shared online was under intense scrutiny, AI took on a pivotal role. Using advanced reverse image search techniques, the AI could trace the original source of an image, its prior use cases, and any alterations made. Similarly, video analysis tools were employed to check the authenticity of videos, identifying elements like manipulated frames or deepfakes. When the AI identified a claim as false or misleading, automated systems were initiated to create detailed reports debunking the claim. These systems did not just flag false information; they also generated explanations and evidence to support the debunking. This transparency was paramount in ensuring user understanding. It helped the public see why a particular claim was false, facilitating a better grasp of the underlying facts and fostering an atmosphere of trust around the AI's fact-checking abilities.

Crowdsourcing efforts further enhance real-time fact-checking, with collaborative platforms allowing a community of users to collectively verify information. Users can submit claims for verification and provide evidence to either support or debunk claims, as seen on the Europe-wide fact-checking platform "FactCheckEU" in 2023. Automated alerts and notifications play a critical role in real-time verification systems. These systems can alert users when a potentially false or misleading claim is detected, thus helping prevent the rapid spread of misinformation. While automation is vital in real-time fact-checking, human fact-checkers still provide essential expertise and judgement, especially for complex or context-dependent claims. AI systems assist human fact-checkers by prioritizing tasks, summarizing information, and offering evidence from knowledge graphs. For example, during the COVID-19 pandemic in Brazil, a joint human-AI effort was instrumental in combating the misinformation surrounding the virus and vaccines.[15] To stay abreast of the ever-evolving landscape of misinformation and the ongoing development of AI technologies, it's essential to continuously update and refine these methods and techniques. Ensuring the reliability and accuracy of automated fact-checking systems remains paramount, with human oversight and validation playing a critical role.

4. Cross-lingual Fact-Checking:

Cross-lingual fact-checking, the practice of verifying claims and information across multiple languages, presents a unique set of challenges due to the linguistic and cultural differences between languages. Nevertheless, advanced methods and techniques have been developed to tackle these challenges, leading to the creation of a robust framework for cross-lingual fact-checking.

The use of multilingual pre-trained language models, such as the transformer-based language models like multilingual BERT (mBERT) and XLM-R, stands as a pillar of this framework. Trained on vast multilingual corpora, these models are

[15] Abonizio, Hugo Queiroz, Ana Paula Ayub da Costa Barbon, Renne Rodrigues, Mayara Santos, Vicente Martínez-Vizcaíno, Arthur Eumann Mesas, and Sylvio Barbon Junior. "How people interact with a chatbot against disinformation and fake news in COVID-19 in Brazil: the CoronaAI case." International Journal of Medical Informatics (2023): 105134.

proficient in handling multiple languages, becoming instrumental in various NLP tasks, including fact-checking.[16] For instance, in India, mBERT has been used to fact-check information in a range of languages including Hindi, Bengali, and Tamil, serving as a crucial bulwark against automated information manipulation efforts that take local dialects into account.[17] This was bolstered by the application of multilingual knowledge graphs. These are repositories with cross-lingual information, that can align data from various languages, thereby allowing fact-checkers to leverage existing data for verification. A prime example is the utilization of Wikidata, a multilingual knowledge graph, for fact-checking during the 2022 European Union elections.[18] Parallel fact-checking datasets also play an integral role in cross-lingual fact-checking. A parallel fact-checking dataset might contain one dataset of fact-checked claims about a specific event (like an election or a natural disaster) in English, and another dataset of fact-checked claims about the same event in another language, like Spanish or Mandarin. These parallel datasets can be extremely valuable in training machine learning algorithms for multilingual fact-checking and by learning from these parallel datasets, an algorithm can potentially understand how misinformation or facts may be presented differently in different languages or cultural contexts, and thereby improve its ability to fact-check claims across languages or contexts.

By creating parallel datasets containing fact-checked claims in multiple languages, these datasets enable supervised learning approaches for claim verification in different languages. A successful implementation of this was seen in Canada, a country with two official languages - English and French. During the Canadian Federal Election of 2021, researchers created parallel fact-checking datasets that included claims from various sources in both English and French.[19] These datasets captured not only the linguistic differences but also the cultural and regional nuances of misinformation in both languages. They were then used to train an AI model, which was successful in identifying and debunking false claims made in either language during the election campaign. This use of parallel fact-checking datasets proved instrumental in ensuring the integrity of the electoral process, highlighting their potential in multilingual and multicultural contexts.

In scenarios where the language of the claim does not match the language of the available fact-checking datasets, machine translation becomes a vital tool. Claims are translated from one language to another, allowing the use of existing models to verify these translated claims. Cross-lingual transfer learning also builds on this by fine-tuning models pretrained on one language using data from another language, allowing for specific task adaption. Crowdsourcing platforms, involving speakers of multiple languages, also facilitate cross-lingual fact-checking. By

[16] Xu, Haoran, Benjamin Van Durme, and Kenton Murray. “Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation.” arXiv preprint arXiv:2109.04588 (2021).

[17] Kar, Debanjana, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. “No rumours please! a multi-indic-lingual approach for covid fake-tweet detection.” In 2021 Grace Hopper Celebration India (GHCI), pp. 1-5. IEEE, 2021.

[18] Rudnik, Charlotte, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphaël Troncy, and Xavier Tannier. “Searching news articles using an event knowledge graph leveraged by wikidata.” In Companion proceedings of the 2019 world wide web conference, pp. 1232-1239. 2019.

[19] Al-Rawi, Ahmed, and Abdelrahman Fakida. “The methodological challenges of studying “fake news”.” Journalism Practice 17, no. 6 (2023): 1178-1197.

leveraging language expertise from individuals worldwide, claims can be verified in a multitude of languages. This method was notably used during the COVID-19 pandemic where a number of crowdsourced efforts and global volunteers fact-checked misinformation in numerous languages.

In Italy, the Bruno Kessler Foundation launched the COVID-19 Infodemic Observatory.[20] This effort employed machine learning to monitor the surfeit of information—or the "infodemic"—associated with COVID-19 on social media platforms. The platform not only identified trending claims but also posted them on their website, inviting the global community to participate in the fact-checking process and engage in informed discussions. Meanwhile, Factcheck.org, a respected non-partisan fact-checking platform, dedicated a special section of its website to fact-checking information about the pandemic. This section served as a hub for users worldwide, allowing them to submit claims they encountered for verification by the platform's team of professional fact-checkers. Simultaneously, the International Fact-Checking Network (IFCN) at the Poynter Institute coordinated the CoronaVirusFacts/DatosCoronaVirus Alliance, an unprecedented global fact-checking operation.[21] This alliance connected over 100 fact-checking organizations from over 70 countries to jointly debunk false claims about the pandemic. In the realm of private messaging, WhatsApp launched a novel feature in certain countries, empowering users to forward messages to fact-checking organizations.[22] Given the rapid spread of misinformation through such apps, this step was a crucial measure to contain the proliferation of false information. Furthermore, Snopes, one of the pioneers of online fact-checking, provided a comprehensive guide to discern fake news sites known for peddling pandemic-related misinformation. This guide proved invaluable in aiding internet users in assessing the credibility of their information sources. Teyit, a leading fact-checking organization based in Turkiye, offered an interactive platform where users could submit claims about the COVID-19 pandemic for verification. By opening up this process to the public, Teyit enabled a more widespread and democratic process of fact-checking.

Techniques that measure semantic similarity between sentences are adapted for cross-lingual fact-checking. The semantic similarity measure can be used to find equivalent claims in different languages and validate information.[23] Additionally, the integration of cross-lingual knowledge bases provides valuable context and evidence for verification. Fact-checking becomes even more potent when other modalities like images, videos, or audio are incorporated, a technique termed multimodal cross-lingual fact-checking. For example, during the 2022 Olympics in Tokyo, multimedia content was cross-checked across languages to ensure the veracity of the shared information[24]. Research in learning language-

[20] https://covid19obs.fbk.eu/

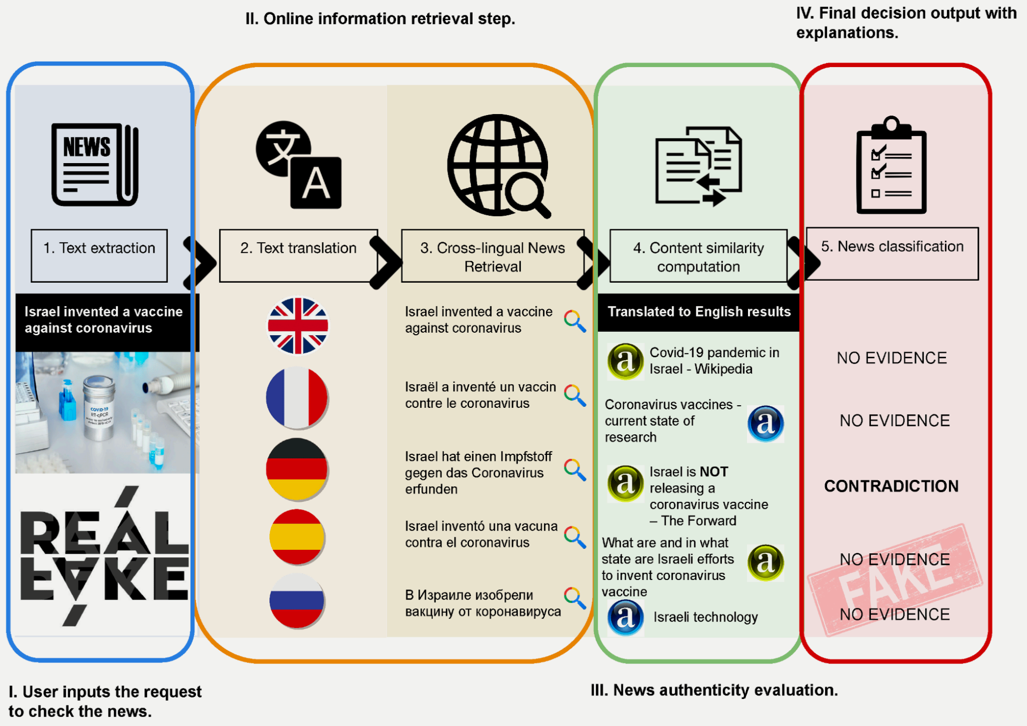
[21] https://www.poynter.org/coronavirusfactsalliance/

[22] Saurwein, Florian, and Charlotte Spencer-Smith. "Combating disinformation on social media: Multilevel governance and distributed accountability in Europe." Digital journalism 8, no. 6 (2020): 820-841.

[23] Sravanthi, Pantulkar, and B. Srinivasu. "Semantic similarity between sentences." International Research Journal of Engineering and Technology (IRJET) 4, no. 1 (2017): 156-161.

[24] Japanese fact-checking body set to counter online misinformation. The Yomiuri Shimbun. 29 September 2022. https://japannews.yomiuri.co.jp/society/general-news/20220929-61235/

agnostic representations can enable fact-checking models to handle claims with low-resource languages, where labeled data may be scarce. This technique played a critical role in fact-checking efforts during the Sudanese political crisis in 2022, where low-resource languages were prevalent.



II. Online information retrieval step.
IV. Final decision output with explanations.
NEWS
1. Text extraction
2. Text translation
3. Cross-lingual News Retrieval
4. Content similarity computation
5. News classification
Israel invented a vaccine against coronavirus
REAL FAKE
Israel invented a vaccine against coronavirus
Israël a inventé un vaccin contre le coronavirus
Israel hat einen Impfstoff gegen das Coronavirus erfunden
Israel inventó una vacuna contra el coronavirus
В Израиле изобрели вакцину от коронавируса
Translated to English results
Covid-19 pandemic in Israel - Wikipedia
Coronavirus vaccines - current state of research
Israel is NOT releasing a coronavirus vaccine – The Forward
What are and in what state are Israeli efforts to invent coronavirus vaccine
Israeli technology
NO EVIDENCE
NO EVIDENCE
CONTRADICTION
NO EVIDENCE
FAKE
NO EVIDENCE
I. User inputs the request to check the news.
III. News authenticity evaluation.

A sample multi-lingual automated fact-checking flowchart. Source: Dementieva, D.; Kuimov, M.; Panchenko, A. Multiverse: Multilingual Evidence for Fake News Detection. J. Imaging 2023, 9, 77. https://doi.org/10.3390/jimaging9040077

Cross-lingual fact-checking is still a challenging and evolving area of research. Language differences, variations in writing styles, and the availability of resources for different languages add layers of complexity to the task. Additionally, models performing well for high-resource languages may not generalize well to low-resource languages due to the lack of data and linguistic diversity. Therefore, a combination of automated techniques, human expertise, and collaborative efforts are often required to ensure accuracy and coverage across languages, ensuring accurate verification.

5. Deepfake and Media Verification:

Deepfake and media verification are paramount in the domain of automated fact-checking, crucial for counteracting the proliferation of manipulated content and misinformation. Advanced methodologies and techniques have been developed, underpinning the evolution of this field and boosting its efficacy. Among the most transformative advances is the application of deepfake detection algorithms. By leveraging computer vision and deep learning techniques, these algorithms are capable of identifying manipulated images and videos.

The use of advanced machine learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in recent US elections underlines their potential in combating manipulated media content. CNNs, renowned for their image classification capabilities, were used to analyze image patterns pixel by pixel and spot any irregularities that could point to manipulation. In contrast, RNNs, known for their sequence prediction properties, were employed to inspect the temporality of video frames and uncover any unnatural transitions or alterations. These AI techniques not only scaled up the fact-checking process but also improved its accuracy by detecting sophisticated manipulations that might elude human observers.

In the realm of personal authentication, there were significant advances in face and voice recognition methods. During the 2023 G7 Hiroshima summit, these technologies played a vital role in safeguarding against deepfake impersonations. The summit, with its global ramifications, was a prime target for misinformation campaigns. To counter this, the authorities implemented state-of-the-art facial recognition technology that could identify subtle facial features and expressions, and voice biometric systems that could recognize individual vocal characteristics, even through alterations. As a result, they were able to validate the authenticity of every image and video clip attributed to the world leaders attending the summit. Further, during the course of the summit, real-time deepfake detection algorithms were also put into action. For example, an AI system was employed to continuously monitor the digital content generated from the summit. It analyzed the facial movements and voice patterns of world leaders in videos to look for anomalies indicative of deepfake technology.

A significant instance was during the Australian bushfires in 2022. Misinformation during such calamities can exacerbate the situation, and to combat this, fact-checkers used reverse image search techniques extensively. For example, when an image depicting a group of koalas supposedly saved from the fires went viral, fact-checkers employed reverse image searches and found the image was actually taken years before the fires. By debunking such misleading content, fact-checkers could focus public attention on the real and pressing issues related to the bushfires. Another case study that highlights the importance of these techniques was during the Syrian Civil War. Misinformation was rife, and several images and videos of incidents of violence and destruction were shared widely. Fact-checkers and OSINT initiatives such as Bellingcat used reverse video search techniques to trace back the source of these videos.[25] In many instances, they found the videos were either doctored or taken from entirely different conflict zones, thereby debunking the false claims and bringing attention to the actual atrocities occurring in the war-torn region.

[25] Ikonen, Pasi, Jere Hokkanen, Turo Uskali, Ville JE Manninen, and Heidi Kuusniemi. “The Networked Utilisation of Satellite Images and Geospatial Technology in Journalism.” In Futures of Journalism: Technology-stimulated Evolution in the Audience-News Media Relationship, pp. 245-260. Cham: Springer International Publishing, 2022.

Analyzing the metadata of images and videos has emerged as another effective verification technique. Metadata provides valuable information about the origin of a media piece and possible alterations. It can disclose details about the device used, location, and editing history, which has been particularly useful in debunking altered images shared on social media platforms like Instagram. The role of forensic experts has not diminished in this AI-dominated era. Instead, their expertise in deploying advanced tools to scrutinize media content for signs of tampering, such as artifacts and inconsistencies, is invaluable. Their involvement was particularly important during the conflict in Eastern Ukraine, where they helped verify images and videos for authenticity.[26]

Another method employed in media verification is content-based hashing. These techniques create unique 'fingerprints' for media content, facilitating fast and efficient comparison for similarity analysis. Multimedia forensics, another prominent approach, is dedicated to developing algorithms and methods for detecting image and video manipulations, such as splicing, retouching, and superimposition. The advent of Generative Adversarial Networks (GANs) led to a surge in deepfakes, pushing researchers to develop GAN detection methods that focus on identifying the traces left by GANs in manipulated content. In a similar vein, multimodal analysis, the confluence of text, image, and video analysis, offers a comprehensive understanding of the context and potential manipulation of media content. Furthermore, organizations host challenges and competitions to promote the development of effective deepfake detection methods, fostering innovation in synthetic media detection. The robustness of deepfake detection models is further enhanced by data augmentation and adversarial training techniques. These simulate different manipulation scenarios and adversarial attacks, enhancing model resilience. For example, Facebook's Deepfake Detection Challenge[27] used these techniques to train models and improve their effectiveness.

Explainable AI plays a crucial role in enhancing transparency and user trust in how fact-checkers adopt new technologies. The development of fast and efficient algorithms for real-time deepfake detection, particularly for immediate media verification during live events, stands as an essential milestone. However, as deepfake and media verification methods evolve, so do the techniques deployed by malicious actors aiming to evade detection. Hence, automated fact-checking systems must stay updated with the latest developments and continuously improve their methodologies to tackle these challenges. The value of human oversight and collaboration with field experts should not be underestimated, as they ensure the accuracy and reliability of media verification processes.

[26] Chen, Emily, and Emilio Ferrara. "Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 17, pp. 1006-1013. 2023.

[27] https://tech.facebook.com/ideas/2020/9/dfdc/

CROWDSOURCING AND COLLABORATIVE CROSS-PLATFORM MONITORING FOR FACT-CHECKING

Fact-checking organizations and social media platforms have been forging increasingly robust alliances worldwide in the shared mission to battle the spread of misinformation and disinformation. One of the most significant examples is Facebook's Third-Party Fact-Checking Program, initiated in 2016.[28] This program involved Facebook joining forces with over 80 fact-checking organizations around the globe that have certification from the non-partisan International Fact-Checking Network (IFCN). Through this initiative, Facebook significantly curtails the distribution of stories that have been flagged as false and supplements them with related articles from fact-checkers to provide additional context.

Twitter also undertook a similar initiative with the launch of its Birdwatch program in 2021. While this pilot program was initially exclusive to the U.S. and mainly relied on users to identify and annotate misleading information, Twitter has expressed intentions of integrating established fact-checking organizations into the Birdwatch program, thereby enhancing its credibility and effectiveness. In parallel, YouTube has made strides to integrate fact-checking panels into its platform. This has been particularly emphasized for breaking news events and search queries related to topics that are frequently subjected to misinformation.

In India, for example, YouTube has established partnerships with several organizations, including BOOM FactCheck and Fact Crescendo, which aid in the verification process and provide users with critical information. Sharing information and insights can lead to more comprehensive detection of disinformation campaigns and coordinated manipulation efforts. Collaborative Cross-Platform Monitoring is a powerful approach that brings together fact-checking organizations, social media platforms, tech companies, and researchers to collectively combat misinformation and verify information across various online platforms. By focusing on the technicalities of new technologies and techniques associated with Collaborative Cross-Platform Monitoring, this extended analysis explores its impact on fact-checking and information verification.

1. Data Sharing and Interoperability:

Data sharing and interoperability stand as the pillars of crowdsourcing and collaborative fact-checking. These crucial aspects facilitate the flow of information, foster collaboration among fact-checking organizations and individuals, and enhance the efficiency of the fact-checking process. Many fact-checking organizations such as FullFact, Factmata and MediaWise offer open APIs and

[28] https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking

have adopted common data standards. This approach allows other platforms and tools to access their fact-checking data in a standardized format, ensuring interoperability and smooth integration between various fact-checking systems. For instance, the global news agency Reuters employs such an open API to share its fact-checking data, promoting the harmonization of fact-checking processes worldwide.

In line with this, the use of Linked Data principles and the Resource Description Framework (RDF) has become increasingly common to publish and interconnect fact-checking data in a decentralized and standardized way.[29] This approach enables data from diverse sources to be linked, simplifying the verification and cross-referencing of claims. The BBC, for example, employs these methods to streamline its fact-checking efforts.[30] Furthermore, online platforms acting as data clearinghouses have been crucial in fostering data sharing and collaboration between fact-checking organizations. These platforms host datasets, APIs, and tools for fact-checkers to access and contribute data, an approach exemplified by the European Union's digital platform for fact-checking.

Fact-checking data can also be integrated into a knowledge graph, allowing it to be linked and used alongside other related information. Knowledge graphs facilitate advanced querying and reasoning, amplifying the effectiveness of collaborative fact-checking efforts. Google's Knowledge Graph is a prime example of this, offering an efficient way to interconnect disparate pieces of information.[31] With the advent of federated learning, multiple parties can collaboratively train machine learning models without sharing raw data. This method, applied in the realm of fact-checking, improves model accuracy without compromising data privacy, an approach effectively used by privacy-oriented organizations such as DuckDuckGo. Blockchain technology, renowned for its immutable and transparent record-keeping capabilities, has found its application in fact-checking as well. It ensures data provenance and builds trust among participating organizations by tracking data sharing and contributions. For instance, the New York Times' News Provenance Project utilizes blockchain for this very purpose.[32] Considering the sensitivity of data, privacy-preserving techniques like differential privacy have been employed to ensure that confidential information remains protected while still allowing data sharing.

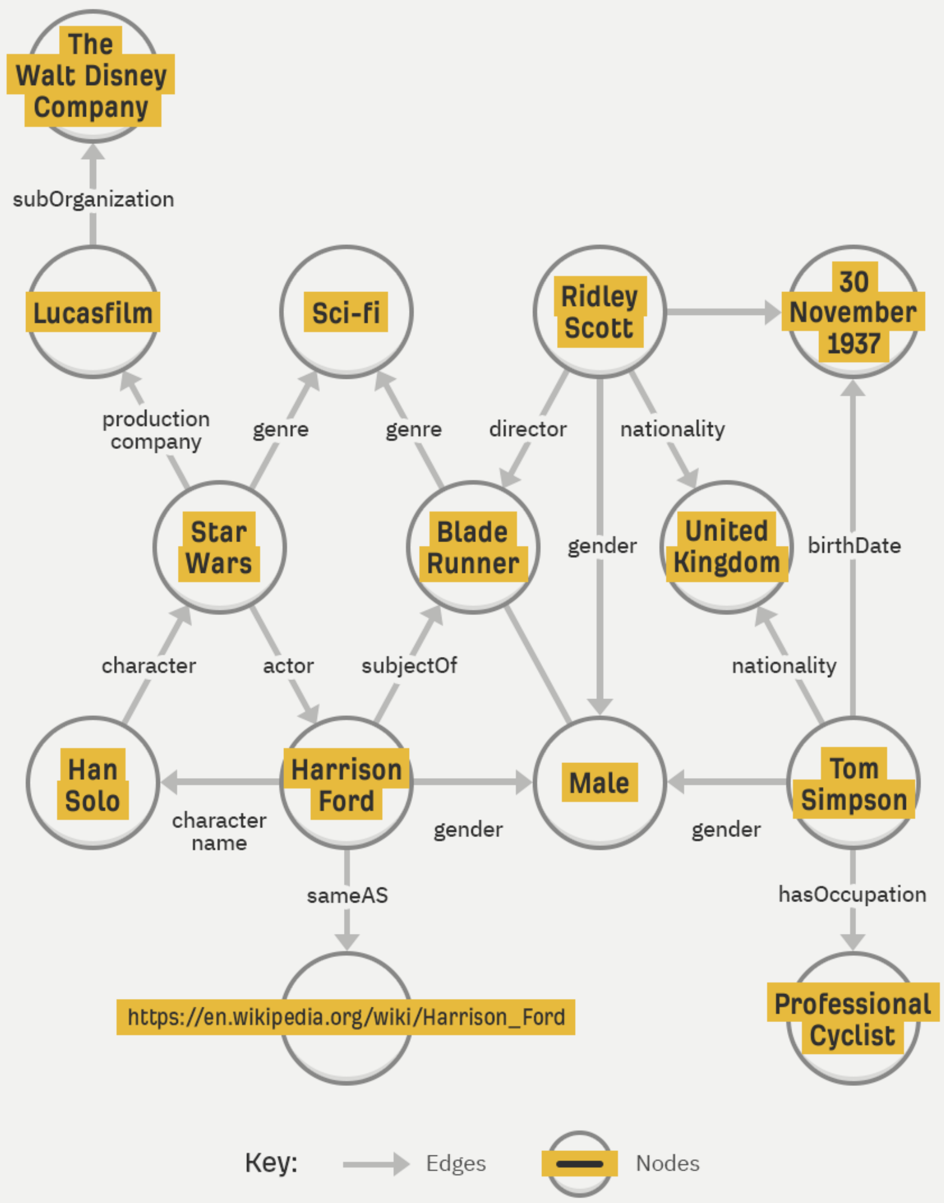
[29] https://www.w3.org/wiki/LinkedData

[30] BBC Ontologies. https://www.bbc.com/ontologies

[31] Tchechmedjiev, Andon, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. "ClaimsKG: A knowledge graph of fact-checked claims." In The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18, pp. 309-324. Springer International Publishing, 2019.

[32] https://www.newsprovenanceproject.com/

What Google's Knowledge Graph Looks Like



A sample blueprint of how Google Knowledge Graphs operate. Source: Pecanek, Michal. 'Google's Knowledge Graph Explained: How It Influences SEO'. September 2020. https://ahrefs.com/blog/google-knowledge-graph/

Collaborative fact-checking also relies heavily on dedicated online platforms that promote collective efforts among volunteers, journalists, and experts. These platforms, like Wikipedia, offer tools for data sharing, claim verification, and collaborative editing, thus facilitating large-scale fact-checking projects. In the fast-paced world of news and information, real-time data sharing mechanisms have become essential.They enable fact-checkers to access the most recent information and contribute to live fact-checking efforts, an approach that has proved vital during events like presidential debates or during a breaking news situation. To ensure transparency, collaborative fact-checking platforms often implement version control and revision tracking mechanisms. This helps keep track of changes to shared data, a method employed by software platforms like GitHub.

Additionally, curation mechanisms to validate and verify shared data help maintain data quality and reliability, a strategy often used in scientific data repositories. In conclusion, effective data sharing and interoperability in crowdsourcing and collaborative fact-checking hinge on trust, transparency, and cooperation among all participants. By employing advanced data sharing techniques, more comprehensive and accurate fact-checking results can be achieved, enabling multiple organizations and individuals to jointly combat misinformation and promote credible information sharing.

2. API Integration and Real-Time Data Collection:

API integration and real-time data collection serve as key facilitators in crowdsourcing and collaborative fact-checking initiatives. Their crucial function lies in ensuring smooth communication between fact-checking platforms, data sources, and contributors, paving the way for an efficient and updated information exchange. To keep up with the rapidly changing landscape of information, sophisticated web scraping and crawling techniques are used to collect data from various online sources, such as news websites, social media platforms, and official statements. For example, firms like Brandwatch employ real-time monitoring to make sure that fact-checkers have access to the freshest information as it surfaces.

Fact-checking platforms can open their APIs to external applications and contributors, granting access to their data, tools, and fact-checking functionalities. An instance of this would be the Google Fact Check Tools API which encourages cooperation between different fact-checking initiatives and fosters the integration of real-time data. Real-time communication between servers is made possible by webhooks and WebSockets, enabling instant updates and notifications when new data or contributions become available. For example, Slack, a popular communication platform, uses webhooks to send real-time notifications. Fact-checking platforms can also federate their APIs, creating a unified and interoperable ecosystem for real-time data exchange among various stakeholders.

Data streaming techniques such as Apache Kafka or RabbitMQ can be harnessed to manage real-time data flows, ensuring efficient data delivery to relevant parties. In the realm of social media, integration with their APIs allows fact-checkers to monitor and collect data from platforms like Twitter, Facebook, and YouTube in real-time. This proves particularly useful in identifying and debunking viral misinformation campaigns, as seen in Facebook's partnership with third-party fact-checking organizations. Geolocation-based APIs and those that analyze user-generated content, such as images and videos, also bring immense value to the table. They allow fact-checkers to filter and collect information based on specific locations or regions, and to detect deepfakes and manipulated media. For instance, Google Cloud Vision API allows for the analysis of user-generated content.

Knowledge graph APIs further expand the horizons for fact-checkers, granting access to structured information and facilitating claim verification against existing facts. On top of that, APIs that perform data validation and quality control checks, as well as machine learning APIs, play significant roles in ensuring the accuracy, reliability, and relevance of real-time data collected from multiple sources. Blockchain technology, due to its nature, allows for tracking the origin and history of data collected in real-time, thereby ensuring transparency and trust. For instance, IBM's Blockchain Transparent Supply solution offers a clear provenance of data and its history.

To cap it all, rate limiting and throttling mechanisms in APIs help control data flow and prevent overload on fact-checking servers, allowing for efficient and uninterrupted operations. Leveraging these tools and technologies allows crowdsourcing and collaborative fact-checking initiatives to efficiently access and analyze information, making their efforts more effective in combating misinformation and promoting accurate reporting. However, alongside these advancements, ensuring data quality, privacy, and security remains of paramount importance in these collaborative efforts. It is through the deployment of these advanced techniques that such objectives can be achieved.

3. Real-Time Alerts and Early Warning Systems:

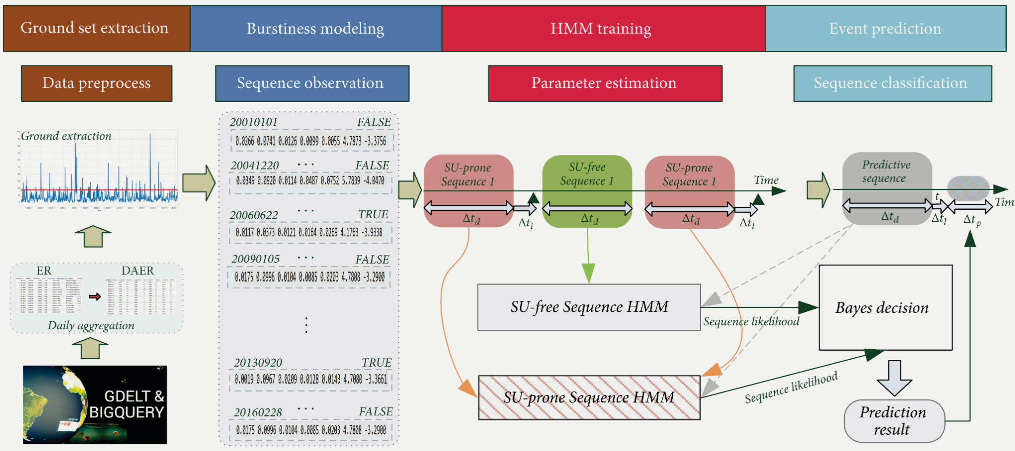
Real-time alerts and early warning systems play a pivotal role in crowdsourcing and collaborative fact-checking initiatives. The Early Model Based Event Recognition using Surrogates, or EMBERS, is an innovative tool developed by the Discovery Analytics Center at Virginia Tech.[33] This platform harnesses open-source data, from tweets to restaurant reservations or even currency exchange rates, to anticipate significant societal events like protests or riots before they occur. In fact, EMBERS successfully forecasted the protests in Brazil back in 2013, demonstrating its predictive potential. On a global scale, we have the Global Database of Events, Language, and Tone, or GDELT. This system diligently monitors media from sources around the world in a multitude of languages. Utilizing natural language processing, GDELT generates data on events, pinpoints their location, and identifies the involved actors. This comprehensive analysis can effectively detect the simmering build-up of social unrest. Moreover, we have the Integrated

[33] Saraf, Parang, and Naren Ramakrishnan. "EMBERS autogsr: Automated coding of civil unrest events." In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 599-608. 2016.

Crisis Early Warning System, known as ICEWS. This platform, initiated by the Defense Advanced Research Projects Agency (DARPA), leverages a diverse range of data, including international news, to predict political crises like protests or riots.[34] Finally, a different approach is seen in the application of PredPol, an algorithm-based software used in law enforcement. PredPol predicts areas of potential criminal activity, including civil unrest, by analyzing historical data.

Their main aim is to quickly detect and flag potentially misleading or false information right at its inception. This way, fact-checkers can respond promptly, stalling the rapid spread of misinformation before it takes root. Among the effective methods employed in this context is real-time data collection and monitoring. Advanced web scraping and data crawling techniques continuously scan online sources, ranging from news websites to social media platforms and official statements, in search of new information and claims. This real-time data collection guarantees fact-checkers have immediate access to the freshest information, such as the approach employed by web data integration platforms like Import.io.

Sentiment analysis and topic modeling techniques have proven invaluable in this endeavor. They assist in identifying trending topics and gauging the sentiment surrounding certain issues. For instance, IBM's Watson Tone Analyzer can detect sudden spikes in negative sentiment or discussions around specific topics, which may indicate the birth of potential misinformation. Analyzing the virality and engagement of posts on platforms like Twitter or Facebook helps identify rapidly spreading information. A high level of engagement with specific content often necessitates urgent fact-checking. Similarly, social media analytics tools monitor the reach and impact of information across different platforms, pinpointing influential sources and amplifiers of possible misinformation.



A sample workflow integrating event datasets to produce forecasts through parameter estimation and sequence classification. Source: Qiao, Fengcai, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang. "Predicting social unrest events with hidden Markov models using GDELT." Discrete Dynamics in Nature and Society 2017 (2017).

[34] Wang, Wei, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. "Growing pains for global monitoring of societal events." Science 353, no. 6307 (2016): 1502-1503.

In the technology infrastructure supporting these efforts, the implementation of event-driven architectures permits real-time processing and an immediate response to incoming data and alerts. This architecture, as used in systems like Apache Kafka, ensures that actions are taken promptly based on the emerging information. Machine learning also comes to aid in these endeavors, especially for anomaly detection. Algorithms can be trained to detect anomalies in data patterns, thus identifying sudden spikes in the dissemination of potentially false information. A good example of this application is Twitter's real-time anomaly detection system. Natural Language Processing (NLP), with techniques such as keyword extraction and named entity recognition, can be applied to detect factual claims and potential misinformation in real-time data streams.

These techniques facilitate the identification of check-worthy claims. Similarly, setting up keyword-based alerts helps fact-checkers receive notifications when specific terms or phrases associated with misinformation are detected. Fact-checkers can further expand their watch by monitoring multiple platforms simultaneously. This can reveal coordinated efforts to spread misinformation across different channels, enhancing the effectiveness of response measures. Real-time verification APIs can be integrated into the fact-checking workflow, enabling quick assessments of the accuracy of claims and timely responses. For example, the Google Fact Check Tools API provides real-time verification capabilities. Collaborative alerting systems foster a collective approach in early warning and debunking.

Platforms that allow users and fact-checkers to submit potential misinformation and receive real-time alerts help build a collaborative wall against the onslaught of misinformation. Advanced algorithms for the early detection of deepfakes and manipulated media, like those used by Deeptrace Labs, provide an additional line of defense, aiding in early identification and debunking. The effectiveness of real-time alerts and early warning systems relies heavily on the quality and speed of data collection, the precision of detection algorithms, and the timely collaboration among fact-checking organizations and individuals. By leveraging advanced technologies and collaborative efforts, these systems play a significant role in mitigating the impact of misinformation, thereby enhancing the credibility of information shared in real-time.

4. Privacy and Data Protection:

The importance of privacy and data protection is elevated when it comes to crowdsourcing and collaborative fact-checking efforts, given that several parties and contributors are involved. It becomes crucial to ensure that sensitive information remains secure and confidential. Advanced methods and techniques have been developed and implemented to ensure this security. Data minimization plays a key role in protecting privacy. By adopting data minimization practices, organizations make sure to collect and retain only the necessary and relevant data, mitigating the risk of exposing sensitive information. For instance, Apple's data minimization approach with its user data has been lauded as a benchmark in the tech industry.

To further ensure the protection of identities of contributors and users, data is often anonymized or pseudonymized. This makes it extremely difficult to trace the data back to an individual. Differential privacy techniques go a step further by adding noise to the data, which provides individual privacy protection but still enables an accurate aggregate analysis. Keeping data safe during transmission and storage is another vital aspect of data protection. To achieve this, secure protocols, such as encryption and secure sockets layer (SSL), are implemented.[35] This protects the data from falling into the wrong hands. Similarly, access control and role-based permissions serve to limit access to sensitive data, ensuring that only authorized personnel can view and handle such information.

Another critical process is conducting Data Protection Impact Assessments (DPIAs).[36] DPIAs can identify and address potential privacy risks in the crowdsourcing and fact-checking process, allowing for proactive measures to mitigate any identified risks. Privacy-preserving machine learning techniques, such as federated learning and secure multi-party computation, come in handy in protecting privacy while enabling model training. This approach allows for model training without the need for sharing raw data, a practice famously used by Google's Gboard keyboard app for personalized text predictions. Trustworthy data sharing agreements among participating organizations serve to ensure compliance with privacy regulations and protect the rights of data contributors. They set clear rules and boundaries about how the data can be used. Transparency and consent are two pillars in privacy protection. Obtaining explicit user consent for data collection and processing is essential. Providing clear and transparent information about data usage allows users to make informed decisions about their data.

Every fact-checking platform should have a comprehensive privacy policy that outlines how user data is handled and provides mechanisms for users to exercise their privacy rights. These policies should be clear and readily available to the users. Regular auditing and compliance monitoring ensure that privacy measures are continuously adhered to and that potential breaches are detected and promptly addressed. This process helps maintain user trust and confidence in the system. Data deletion and retention policies dictate that data should not be stored longer than necessary and should be properly disposed of when it is no longer needed. These policies ensure that data does not linger, and risk being compromised.

In instances dealing with sensitive data, such as personally identifiable information

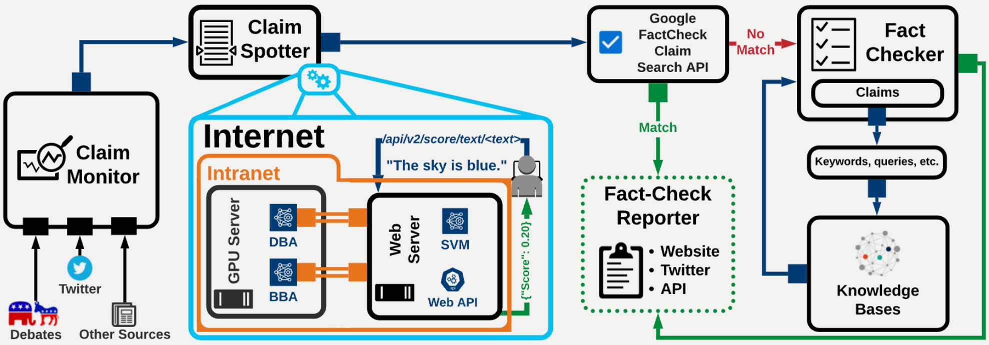
[35] Wang, Wei, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. "Growing pains for global monitoring of societal events." Science 353, no. 6307 (2016): 1502-1503.

[36] Demetzou, Katerina. "Data Protection Impact Assessment: A tool for accountability and the unclarified concept of 'high risk'in the General Data Protection Regulation." Computer Law & Security Review 35, no. 6 (2019): 105342.

(PII), explicit and informed consent from users is a non-negotiable requirement.[37] Furthermore, using secure communication channels, such as end-to-end encrypted messaging, ensures the protection of sensitive information during data sharing and communication. Finally, if third-party vendors are involved, it is crucial to ensure they adhere to stringent security and privacy standards to safeguard the data. This diligence extends the protective measures beyond the immediate organization. By incorporating these advanced privacy and data protection methods into the crowdsourcing and collaborative fact-checking process, organizations can cultivate trust with contributors and users while maintaining the integrity and confidentiality of sensitive information. Implementation of these measures is not only good practice but also aligns with privacy regulations, promoting responsible data handling practices within the fact-checking ecosystem.

5. The Role of Blockchain in Data Integrity:

Blockchain technology, known for its decentralization and immutability, is proving to be a game-changer in the field of crowdsourcing and collaborative fact-checking. Its unique attributes bolster data integrity, enhance transparency, and foster trust among participants. Several innovative applications of blockchain have emerged in this arena. Blockchain technology provides data provenance and integrity by creating an immutable record of data contributions and fact-checking results. This is akin to how blockchain is utilized in supply chain management, where each stage of a product's life cycle can be tracked and verified.[38]



ClaimBuster's blueprint for multimodal fact-checking. Sources: https://idir.uta.edu/claimbuster/

In this context, every addition or update to the blockchain is cryptographically linked, ensuring the traceability and integrity of the data. Blockchain's decentralized nature allows for the creation of a decentralized fact-checking platform. Here, numerous participants can contribute and verify information

[37] Schwartz, Paul M., and Daniel J. Solove. "The PII problem: Privacy and a new concept of personally identifiable information." NYUL rev. 86 (2011): 1814.

[38] Edwardsson, Malin Picha, and Walid Al-Saqaf. "Drivers and barriers for using blockchain technology to create a global fact-checking database." Online Journal of Communication and Media Technologies 12, no. 4 (2022): e202228.

without being reliant on a central authority, creating a more democratic and reliable system. It also offers reliable timestamping for data and claims, indicating when information was added to the blockchain. Timestamp verification becomes crucial in validating the accuracy and recency of data during the fact-checking process, much like how Bitcoin transactions are timestamped and verified on the blockchain.[39] Smart contracts on the blockchain can automate the collaboration process in crowdsourcing and collaborative fact-checking.

These contracts can be used to define rules, rewards, and penalties, thereby ensuring transparency and fairness among participants. This bears similarities to how smart contracts are used in Ethereum to enforce agreements without the need for intermediaries. Blockchain-based platforms can use tokens or cryptocurrencies to incentivize contributors for their fact-checking efforts. Such a mechanism can encourage more participation and enhance the accuracy of results, similar to how blockchain-based content platforms reward creators. Blockchain's consensus mechanisms, such as Proof of Work (PoW) or Proof of Stake (PoS), can be employed for data verification and agreement among participants.[40] These mechanisms ensure that the information is agreed upon and validated by a majority, adding to the trustworthiness of the data. Blockchain can also serve as a tool for privacy and data protection. Techniques like zero-knowledge proofs, which allow users to prove they possess certain information without revealing it, help users retain control over their data while still contributing to the fact-checking process.[41] The technology can also facilitate the creation of collaborative knowledge graphs.

Verified facts and data are stored in a decentralized and tamper-resistant manner, allowing easy access and reference to previously fact-checked information. Blockchain can foster reputation and trust systems for participants, enhancing the credibility of contributors. This is somewhat similar to decentralized identities on the blockchain, where the reputation of an entity can be tracked and verified. Cross-platform data exchange is another area where blockchain excels. It enables a secure and standardized way of exchanging data between different fact-checking platforms and databases, promoting interoperability and facilitating the sharing of information among fact-checking organizations. Moreover, the distributed ledger of the blockchain ensures data consistency across the network, thereby preventing the duplication or alteration of information.

Finally, all fact-checked claims can be stored on the blockchain as immutable records, creating a comprehensive archive of verified information. While leveraging blockchain technology can enhance the reliability, transparency, and efficiency of crowdsourcing and collaborative fact-checking initiatives, it's vital to bear in mind the scalability and energy consumption concerns of blockchain systems when deploying them for large-scale fact-checking projects. To conclude, Collaborative

[39] Dwivedi, Ashutosh Dhar, Rajani Singh, Sakshi Dhall, Gautam Srivastava, and Saibal K. Pal. "Tracing the source of fake news using a scalable blockchain distributed network." In 2020 IEEE 17th international conference on mobile ad hoc and sensor systems (MASS), pp. 38-43. IEEE, 2020.

[40] Akbar, Nur Arifin, Amgad Muneer, Narmine ElHakim, and Suliman Mohamed Fati. "Distributed hybrid double-spending attack prevention mechanism for proof-of-work and proof-of-stake blockchain consensuses." Future Internet 13, no. 11 (2021): 285.

[41] Sun, Xiaoqiang, F. Richard Yu, Peng Zhang, Zhiwei Sun, Weixin Xie, and Xiang Peng. "A survey on zero-knowledge proof in blockchain." IEEE network 35, no. 4 (2021): 198-205.

Cross-Platform Monitoring is heralding a transformative era in fact-checking and information verification. The utilization of novel technologies and techniques enables fact-checkers to comb through vast data sets, spot misinformation trends, and unearth coordinated disinformation campaigns across various online platforms. This cross-platform collaboration, powered by technological advancements, presents a unified front in combatting misinformation, fostering a more informed and resilient information ecosystem. While challenges do persist, the potential benefits offered by Collaborative Cross-Platform Monitoring in countering misinformation make it an attractive strategy for fact-checkers and information verification initiatives.

OPEN-SOURCE INVESTIGATION/INTELLIGENCE (OSINT)

Fact-checkers can leverage open-source intelligence and publicly available data to verify claims and debunk false information. Utilizing crowd-sourcing and collaborative platforms can enable fact-checkers to access a wide range of expertise and resources. Open-Source Investigation (OSINT) has emerged as a powerful tool in fact-checking and information verification, leveraging publicly available information from various sources to corroborate claims, identify misinformation, and uncover hidden truths. By focusing on the technicalities of new technologies and techniques associated with OSINT, this extended analysis explores its impact on fact-checking and information verification.

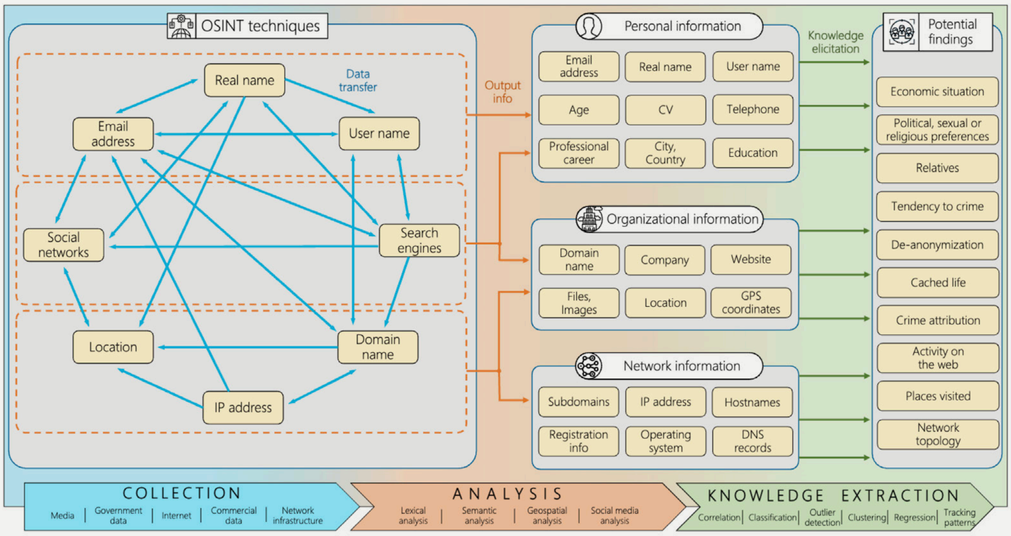
1. Digital Forensics and Image Verification:

Open-source intelligence (OSINT) is a critical tool that relies heavily on digital forensics techniques for the verification of images and videos that are shared online. To authenticate visual content and identify manipulated or misleading images, fact-checkers utilize a number of tools and techniques including image reverse search, metadata analysis, and the investigation of contextual clues. In the current digital era, these tools and techniques play a significant role in ensuring the integrity of information and shielding the public from deceptive practices. In particular, they are pivotal in the fight against foreign information manipulation and interference.

Leading these advancements is the realm of deep learning, which has been adopted to discern tampered or manipulated media content, notably in the form of deepfakes. These artificial intelligence models, including Generative Adversarial Networks (GANs), are used not only to create deepfakes but also to detect them.[42]

[42] Parveen, Azra, Zishan Husain Khan, and Syed Naseem Ahmad. "Classification and evaluation of digital forensic tools." TELKOMNIKA (Telecommunication Computing Electronics and Control) 18, no. 6 (2020): 3096-3106.

GANs are essentially a contest between two AI systems - one creates the deepfake, and the other tries to detect it, thereby improving the detection capability. A key part of media forensics is image forensics, which involves identifying alterations in images or uncovering the original source of an image. Techniques such as Error Level Analysis (ELA) and JPEG Ghost Detection are often utilized in this regard.[43] ELA identifies areas of an image at different compression levels, highlighting potential areas of modification. JPEG Ghost Detection, on the other hand, helps detect if a portion of an image has been manipulated by comparing it with the rest of the image based on JPEG compression inconsistencies. On the video side of things, tampering detection techniques are often employed, which rely on examining the inconsistencies in video elements, such as lighting, shadows, or even subtle physiological signals like heart rate or breathing rate that can be detected through pixel-level analysis of video frames. The field of audio forensics, meanwhile, is replete with techniques such as spectrogram analysis for uncovering manipulations in audio files, voice biometrics to verify the identity of the speaker, and even advanced AI models to identify artificially synthesized speech.



Source: Pastor-Galindo, Javier, Pantaleone Nespoli, Félix Gómez Mármol, and Gregorio Martínez Pérez. "The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends." IEEE Access 8 (2020): 10282-10304.

Media forensics - such as techniques used by Forensic Architecture - are essential tools, primarily used in the analysis of images and videos. By detecting signs of manipulation like splicing, retouching, and superimposition, fact-checkers can identify inconsistencies in visual content and thereby spot potential foreign interference. The analysis is performed across various platforms and social media networks, helping reveal any coordinated foreign campaigns aimed at spreading misinformation. Further enriching the fact-checking process are linguistic analysis and language identification techniques. These tools help identify the language

[43] Harish Kumar, J., and T. Kirthiga Devi. "Fingerprinting of Image Files Based on Metadata and Statistical Analysis." In Proceedings of International Conference on Deep Learning, Computing and Intelligence: ICDCI 2021, pp. 105-118. Singapore: Springer Nature Singapore, 2022.

and linguistic patterns used in misinformation campaigns conducted by foreign actors.[44] Coupled with social network analysis tools, fact-checkers can uncover connections and communication patterns between foreign entities and local influencers. Contextual analysis, as well as the comparison of current claims with past misinformation campaigns, sheds light on the credibility of the information at hand and reveals recurring patterns of misinformation.

2. Geolocation and Mapping:

Open-source intelligence (OSINT) techniques such as geolocation and mapping tools are crucial to the verification of the location and context of events. These methods allow fact-checkers to corroborate claims and debunk false narratives by using satellite imagery, geotagged data, and location-based information. Their application is especially relevant in combatting foreign information manipulation and interference, where they help verify the origin of information and pinpoint potential misinformation campaigns conducted by foreign actors.

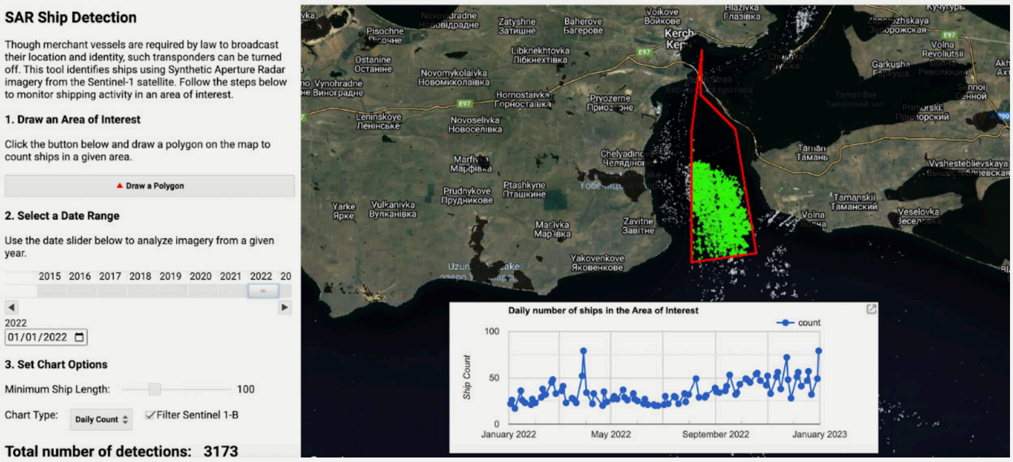
For instance, the geolocation of social media posts can serve as a powerful tool to verify the authenticity of information and identify potential misinformation campaigns originating from foreign locations. Advanced algorithms can analyze the content and metadata of these posts to determine their location of origin.[45] This technique has been crucial in identifying the true location of various posts and images, including those shared during the conflict in Ukraine. Moreover, geolocation can also be combined with sentiment analysis, a process that discerns the sentiments and reactions to specific information in different geographic regions. This approach helps identify variations in the impact of misinformation based on location, as has been seen in the response to various political events worldwide.

Analyzing the digital footprint left by foreign actors can help identify their geographic location and potential involvement in misinformation campaigns. This analysis, coupled with advanced IP address tracking techniques, can identify the source of online information and even detect the use of VPNs to conceal the origin of data. These tactics played a crucial role during the investigations into Russian interference in the 2016 U.S. elections. The power of visualization should not be underestimated either. Mapping and visualization tools can effectively display the geographic distribution of information, helping identify patterns and anomalies across regions. Satellite imagery can provide a real-world view of specific locations, serving as a powerful means to verify events and information, as seen in the verification of human rights abuses in Xinjiang, China.

[44] Yadav, Ashok, Atul Kumar, and Vrijendra Singh. "Open-source intelligence: a comprehensive review of the current state, applications and future perspectives in cyber security." Artificial Intelligence Review (2023): 1-32.

[45] Evangelista, João Rafael Gonçalves, Renato José Sassi, Márcio Romero, and Domingos Napolitano. "Systematic literature review to investigate the application of open source intelligence (OSINT) with artificial intelligence." Journal of Applied Security Research 16, no. 3 (2021): 345-369.

Fact-checking in local languages aids in combating misinformation campaigns targeting specific regions. Analyzing social media connections and interactions between foreign actors and local influencers can reveal potential manipulation efforts. Fact-checkers can also use geolocation to verify the authenticity of sources and confirm their proximity to specific events or incidents. Cross-referencing information with official government data or publicly available sources is another invaluable technique in the verification of claims. For instance, during the COVID-19 pandemic, fact-checkers have often cross-referenced claims about case numbers and mortality rates with data from health departments and the World Health Organization.



SAR Ship Detection

Though merchant vessels are required by law to broadcast their location and identity, such transponders can be turned off. This tool identifies ships using Synthetic Aperture Radar imagery from the Sentinel-1 satellite. Follow the steps below to monitor shipping activity in an area of interest.

1. Draw an Area of Interest

Click the button below and draw a polygon on the map to count ships in a given area.

2. Select a Date Range

Use the date slider below to analyze imagery from a given year.

2022
01/01/2022

3. Set Chart Options

Minimum Ship Length: 100

Chart Type: Daily Count

☑ Filter Sentinel 1-B

Total number of detections: 3173

Bellingcat's work on using open source Synthetic Aperture Radar to identify ships breaking the Russian arms embargo offshore Ukraine. Source: https://www.bellingcat.com/news/2023/05/11/grain-trail-tracking-russias-ghost-ships-with-satellite-imagery/

Fact-checkers can also enhance the accuracy of geolocation data and provide valuable context by collaborating with local experts and organizations. Techniques to verify the authenticity of geo-tagged images and videos ensure they are not manipulated or misleading. These have been useful, for example, in verifying images related to natural disasters or violent incidents. Real-time location verification through geolocation methods empowers fact-checkers to respond quickly to emerging misinformation campaigns. This timely response was crucial during high-profile events such as elections and referendums, where misinformation can have immediate real-world impact. The fusion of geolocation and mapping techniques with other fact-checking methods amplifies the effectiveness of efforts to combat foreign information manipulation and interference. Leveraging advanced technologies and collaborating with local experts enables fact-checkers to better identify the sources and origins of misinformation, thus ensuring accurate and trustworthy information reaches the public.

3. Dark Web Monitoring:

Open-source intelligence (OSINT) in some cases necessitates the monitoring of the dark web to identify disinformation campaigns and malicious actors. While navigation through the dark web poses several challenges, the information that can be uncovered from these hidden corners of the internet proves invaluable for uncovering clandestine efforts to spread misinformation.[46] The importance of dark web monitoring in fact-checking and combatting foreign information manipulation and interference cannot be understated. The dark web, often linked to illegal activities, is a component of the deep web that requires specific software for access and has become a thriving marketplace for the dissemination of misinformation, disinformation, and other forms of harmful content. To navigate the complex web of the darknet, advanced methods and techniques are employed, such as deep and dark web crawling.[47] By using specialized tools and services, fact-checkers can access hidden websites and forums, enabling the collection of a wider range of information and exposing potential sources of misinformation. An example of such exploration was seen in the fight against drug trafficking, where law enforcement officials used similar techniques to identify and track illegal drug marketplaces.

Equally important in this endeavor are NLP techniques, used to analyze textual content from dark web forums, chat logs, and discussions. Sentiment analysis and topic modeling through NLP can help pinpoint coordinated disinformation campaigns and identify trending misinformation topics. One such application was during the 2020 US Presidential Election, where NLP was used to uncover disinformation campaigns aimed at manipulating public opinion.[48] Visual content analysis is another technique fact-checkers employ. Advanced image and video analysis can detect and analyze potentially manipulated or misleading visual content shared on the dark web, with deep learning algorithms serving as tools to detect deepfakes and identify inconsistencies in images and videos. One of the distinct features of the dark web is the prevalence of cryptocurrency.

Monitoring cryptocurrency transactions helps trace the funding sources of misinformation campaigns and identify financial links between actors. Tracking these payments can reveal patterns of funding related to foreign interference efforts, similar to the investigations into ransomware attacks. In the same vein, the dark web also hosts marketplaces that sell and distribute fake documents, hacked data, and other types of disinformation. By monitoring these platforms, fact-checkers can uncover the availability and demand for misinformation-related products, just as security researchers do when tracking cybercrime trends. Artificial intelligence plays an integral role in dark web monitoring as well.

[46] He, Siyu, Yongzhong He, and Mingzhe Li. "Classification of illegal activities on the dark web." In Proceedings of the 2nd International Conference on Information Science and Systems, pp. 73-78. 2019.

[47] Rawat, Romil, Vinod Mahor, Sachin Chirgaiya, and Bhagwati Garg. "Artificial cyber espionage based protection of technological enabled automated cities infrastructure by dark web cyber offender." Intelligence of Things: AI-IoT Based Critical-Applications and Innovations (2021): 167-188.

[48] Ibrishimova, Marina Danchovsky, and Kin Fun Li. "A machine learning approach to fake news detection using knowledge verification and natural language processing." In Advances in Intelligent Networking and Collaborative Systems: The 11th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2019), pp. 223-234. Springer International Publishing, 2020.

Machine learning models can be trained to recognize specific keywords or phrases associated with misinformation campaigns, providing crucial aid in identifying patterns and anomalies that may indicate foreign information manipulation and interference. Fact-checking organizations often collaborate with law enforcement and intelligence agencies to access and analyze dark web data and identify potential threats. These cooperative relationships, similar to those formed to combat terrorism and cybercrime, enhance the effectiveness of dark web monitoring efforts. Data privacy and security measures are crucial when dealing with dark web data. Encryption and secure data storage protect sensitive information and safeguard the identities of researchers and fact-checkers. These practices are not dissimilar to those used by journalists and activists working in hostile environments or dealing with sensitive information. As misinformation knows no linguistic boundaries, monitoring dark web content in various languages is essential. Advanced language processing capabilities help analyze content in different languages, much like how security researchers monitor international cyber threats.

Dark web intelligence platforms offer specialized tools and data feeds for monitoring and analyzing dark web activities. These resources provide significant support for fact-checking organizations in detecting and combating misinformation, similar to how cybersecurity firms use them to track threats. Real-time monitoring of dark web activities allows for timely responses to emerging threats and misinformation campaigns. This immediate detection allows for quick fact-checking and debunking, analogous to cybersecurity incident response in dealing with real-time cyber threats. Engaging in dark web monitoring demands a strong foundation in cybersecurity, data analysis, and understanding legal considerations. Ethical considerations are paramount, as is collaboration with relevant authorities and experts to ensure responsible and effective monitoring. By employing these advanced methods and techniques, fact-checking organizations can proactively identify and counter foreign information manipulation and interference, thereby safeguarding public discourse and promoting accurate and reliable information.

4. Ethical Considerations and Source Verification:

As fact-checkers navigate the vast and complex web of information made available through open-source intelligence (OSINT), it is vital that they adhere to stringent ethical guidelines. Their responsibility extends beyond simply verifying sources to encompass respecting privacy and handling sensitive information responsibly in their investigations. One of the foundational ethical considerations in fact-checking is transparency and disclosure.

Fact-checking organizations are expected to maintain clarity about their methodologies, sources of funding, and any potential conflicts of interest that may arise. By openly disclosing these aspects, they can build trust with their audience and safeguard the integrity of their fact-checking results. For example, Full Fact, a prominent UK-based fact-checking organization, openly shares its funding sources to maintain public trust. Similarly, impartiality and non-bias form another critical element of ethical fact-checking. Regardless of the political or ideological context, fact-checking must be carried out impartially, using objective criteria to evaluate claims. This commitment to neutrality ensures that the assessments produced are fair and accurate. However, the pursuit of truth in fact-checking should also be balanced with the potential impact on individuals and communities.

Fact-checkers must consider the potential repercussions of their findings and should be cautious not to amplify harmful content while debunking misinformation. This can be seen in sensitive situations such as election cycles or times of social unrest where amplified misinformation can lead to real-world harm. Maintaining privacy protection is a crucial part of handling user data and personal information, and fact-checkers must comply with data privacy regulations. Techniques such as anonymizing data and obtaining explicit consent from users are paramount in this regard. In their mission to debunk misinformation, fact-checkers must also be cautious not to inadvertently spread the misinformation further. This calls for a nuanced approach in which they debunk false claims without giving them undue visibility. This principle was exemplified during the COVID-19 pandemic when health organizations and fact-checkers had to carefully balance providing correct information without amplifying potentially harmful misinformation. With the increasing incorporation of artificial intelligence and automation into fact-checking, responsible use of these technologies becomes more crucial.

Despite the benefits these technologies offer, fact-checkers must ensure their implementation is guided by ethical considerations to avoid potential biases and errors. Education also plays a crucial role in fact-checking efforts. By focusing on educating the public about critical thinking and media literacy, fact-checkers can empower the audience to identify misinformation independently. This form of outreach has been used effectively by media organizations around the world, for instance, in the United Nations' campaign to improve global media literacy. Lastly, fact-checkers need to be aware of potential retaliation from malicious actors. Just as journalists sometimes face threats for their work, fact-checkers too must have measures in place to protect their security. Implementing such measures ensures that fact-checkers can continue their vital work in the face of potential retaliation, supporting the ongoing fight against misinformation.

In conclusion, Open-Source Investigation has significantly impacted fact-checking and information verification. By leveraging digital forensics, social media analysis, data mining, and other OSINT techniques, fact-checkers can verify claims, detect misinformation, and provide accurate information to the public. New technologies and automation enhance the efficiency of OSINT efforts, enabling fact-checkers to sift through vast amounts of data and identify potential misinformation rapidly. Collaboration, multilingual capabilities, and ethical considerations are essential in maximizing the benefits of OSINT and ensuring the credibility and reliability of fact-checking and information verification initiatives. As technology continues to advance, OSINT will remain a crucial tool in combating misinformation and fostering a more informed and accountable information ecosystem.

MACHINE LEARNING FOR CONTENT CLASSIFICATION

Machine learning models can be trained to classify content as potentially misleading or factual, assisting fact-checkers in prioritizing their efforts and focusing on the most critical cases of information manipulation. Machine Learning for Content Classification techniques has revolutionized fact-checking and information verification by automating the process of identifying misinformation, categorizing content, and prioritizing fact-checking efforts. By focusing on the technicalities of new technologies and techniques associated with Machine Learning for Content Classification, this extended analysis explores its impact on fact-checking and information verification.

1. Active Learning and Human-in-the-Loop (HITL) Approach:

The combination of Active Learning and the Human-in-the-Loop (HITL) Approach provides a powerful toolkit for significantly enhancing fact-checking methods and combating foreign information manipulation and interference[49]. By uniting the power of machine learning algorithms with human expertise, these techniques optimize the fact-checking process, improve accuracy, and efficiently target misinformation campaigns.

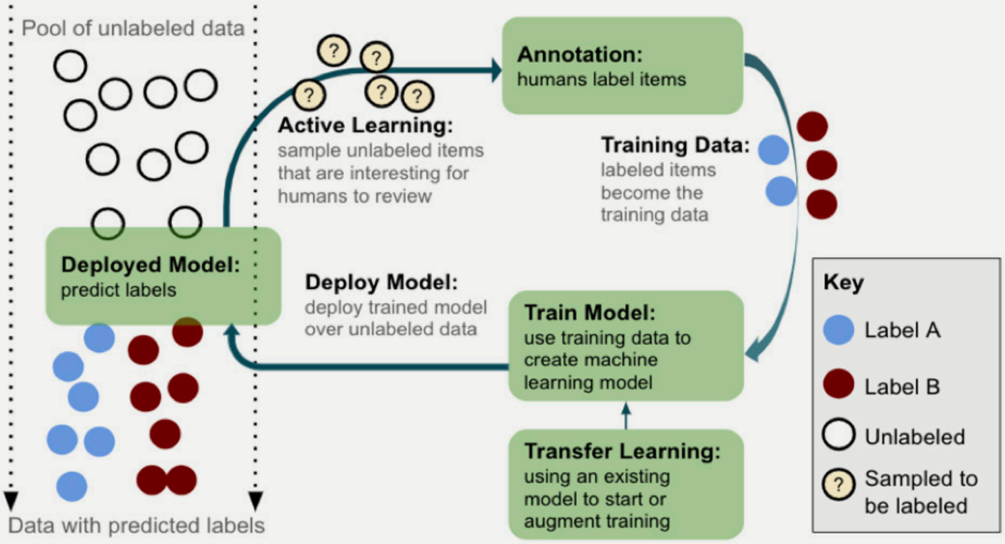
Active learning, a machine learning approach, is instrumental in enhancing the efficiency of fact-checking.[50] It works by selecting the most informative data points for human annotation, helping prioritize which claims or pieces of information require the most urgent human verification. In practice, the machine learning model is initially trained on a small labeled dataset. It then identifies data points that present uncertainty or challenge, which are subsequently annotated by fact-checkers. The model is retrained on this updated dataset, thereby iteratively improving its performance over time. This process reduces the number of manual annotations needed, thereby achieving high accuracy in a more efficient manner.

[49] Liu, Zimo, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. "Deep reinforcement active learning for human-in-the-loop person re-identification." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6122-6131. 2019.

[50]

Complementing active learning, the HITL approach merges human expertise with machine learning algorithms to perform complex tasks. Within the realm of fact-checking, this approach empowers fact-checkers to validate and guide the decisions of AI systems, integrating human judgment and reasoning into the process. Machine learning models propose suggestions or predictions, which are then reviewed and validated by fact-checkers before the results are published. This cooperative process ensures that fact-checking efforts are accurate and reliable, while also taking advantage of the speed and scalability of machine learning. Within these frameworks, specific strategies such as uncertainty sampling and confidence-based labeling play crucial roles. In uncertainty sampling, the model selects data points it is unsure about for human review. This enables fact-checkers to focus on claims the model is least confident about, rectifying potential errors, and enhancing the model's performance.This cooperative process ensures that fact-checking efforts are accurate and reliable, while also taking advantage of the speed and scalability of machine learning. Within these frameworks, specific strategies such as uncertainty sampling and confidence-based labeling play crucial roles. In uncertainty sampling, the model selects data points it is unsure about for human review. This enables fact-checkers to focus on claims the model is least confident about, rectifying potential errors, and enhancing the model's performance.

Complementing active learning, the HITL approach merges human expertise with machine learning algorithms to perform complex tasks. Within the realm of fact-checking, this approach empowers fact-checkers to validate and guide the decisions of AI systems, integrating human judgment and reasoning into the process. Machine learning models propose suggestions or predictions, which are then reviewed and validated by fact-checkers before the results are published. This cooperative process ensures that fact-checking efforts are accurate and reliable, while also taking advantage of the speed and scalability of machine learning. Within these frameworks, specific strategies such as uncertainty sampling and confidence-based labeling play crucial roles. In uncertainty sampling, the model selects data points it is unsure about for human review. This enables fact-checkers to focus on claims the model is least confident about, rectifying potential errors, and enhancing the model's performance.



Pool of unlabeled data
Annotation:
humans label items
Active Learning:
sample unlabeled items that are interesting for humans to review
Training Data:
labeled items become the training data
Deployed Model:
predict labels
Deploy Model:
deploy trained model over unlabeled data
Train Model:
use training data to create machine learning model
Transfer Learning:
using an existing model to start or augment training
Key
Label A
Label B
Unlabeled
Sampled to be labeled
Data with predicted labels

A comparison between Active Learning and Transfer Learning approaches.
Source: https://livebook.manning.com/book/human-in-the-loop-machine-learning/chapter-1/v-11/16

Confidence-based labeling operates similarly, where the model assigns a confidence score to each prediction. Claims with low confidence scores are referred to fact-checkers for manual verification, which reduces the risk of propagating false information. Fact-checkers also utilize data diversity sampling to ensure the training dataset encompasses a diverse range of claims and information, which represents various topics and perspectives. This broad representation aids in comprehensive coverage of fact-checking efforts and counteracts misinformation across various domains. Additionally, feedback and updates from fact-checkers are continuously incorporated to improve the machine learning model. This system adapts and learns from new data and human judgment, leading to a more effective process. The HITL approach also helps mitigate potential biases in AI models by integrating diverse human perspectives and judgment. Fact-checkers can identify and correct these biases, upholding a fair and objective fact-checking process. This collaborative approach also enhances the scalability and efficiency of the fact-checking process. By focusing on the most relevant data points, active learning reduces the burden of manual annotation.

Moreover, the HITL approach enables real-time fact-checking, facilitating quick responses to emerging misinformation campaigns and foreign information manipulation. In conclusion, by adopting advanced methods such as active learning and the HITL approach, fact-checking organizations can significantly improve their efforts to combat foreign information manipulation and interference. This cooperative process between AI models and human fact-checkers ensures a robust and ethical fact-checking ecosystem that upholds the principles of truthfulness and accurate reporting.

2. Transfer Learning for Few-Shot Fact-Checking:

In an age where misinformation spreads at an unprecedented rate, few-shot fact-checking powered by transfer learning stands as a powerful, innovative approach that efficiently combats this issue. This technique uses pre-trained models alongside a limited set of labeled data to perform accurate fact-checking. This is particularly advantageous in scenarios where labeled data is scarce. Such an approach holds substantial potential for combating foreign information manipulation and interference, allowing fact-checkers to rapidly respond to emerging misinformation campaigns and effectively verify claims with limited resources.

In few-shot fact-checking, pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and RoBERTa (A Robustly Optimized BERT Pretraining Approach) are instrumental.[51] These models, trained on vast amounts of textual data, capture rich representations of language, understanding complex linguistic patterns, context, and semantics. Acting as the base, they provide a robust foundation for comprehending and analyzing text in the realm of fact-checking. Building on this foundation, the pre-trained model undergoes a process known as fine-tuning. Here, the model is trained on a smaller dataset of fact-checking examples, typically consisting of labeled claims and their associated veracity (true or false).

This specialized training allows the model to adapt its knowledge to the context of fact-checking, aligning its understanding with the specific task at hand. Complementing transfer learning, the few-shot learning paradigm is employed. This approach trains models to make accurate predictions based on a minimal number of examples per class (in this case, fact-checking labels). Consequently, the model effectively generalizes its knowledge from a small number of labeled samples. This notion of generalization is further enhanced through the use of prototypical networks, a few-shot learning method that encapsulates the training data into prototypes for each class. Within the context of fact-checking, these prototypes represent true and false claims, allowing the model to predict the veracity of new claims with limited labeled data. This process is further enhanced through meta-learning and metric learning. These approaches train models to quickly adapt to new tasks based on a minimal number of examples and learn a distance metric that measures the similarity between data points.

These methods play a vital role in few-shot fact-checking, where the model's ability to quickly adapt to new claims with limited labeled data is paramount. The flexibility of transfer learning extends into cross-domain contexts. Here, the model transfers knowledge from one domain, like news articles, to another, such as

[51] Lee, Nayeon, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. "Towards few-shot fact-checking via perplexity." arXiv preprint arXiv:2103.09535 (2021).

social media posts. This flexibility proves valuable in fact-checking across various platforms, where information manipulation can occur. The active learning and Human-in-the-Loop (HITL) strategies, previously discussed, also augment few-shot learning. By focusing on the most informative data points for human annotation, active learning works hand in hand with HITL, which integrates human reviewers into the fact-checking process, thus validating and guiding the model's predictions to ensure accuracy. The amalgamation of few-shot learning and transfer learning paves the way for real-time fact-checking and rapid response capabilities. It allows for the model to swiftly adapt to new claims, which are pivotal for curbing emerging misinformation campaigns and foreign information manipulation.

Furthermore, the reduced need for extensive labeled data as a result of using transfer learning with few-shot learning enhances the scalability and efficiency of fact-checking efforts. In summary, the adoption of advanced methods in transfer learning for few-shot fact-checking enables fact-checking organizations to effectively counteract foreign information manipulation and interference, even in scenarios with limited labeled data. By amalgamating pre-trained language models, few-shot learning, and human validation within the loop, a robust and accurate fact-checking ecosystem is fostered, promoting the swift dissemination of verified information to the public.

3. Ensemble Methods for Improved Accuracy:

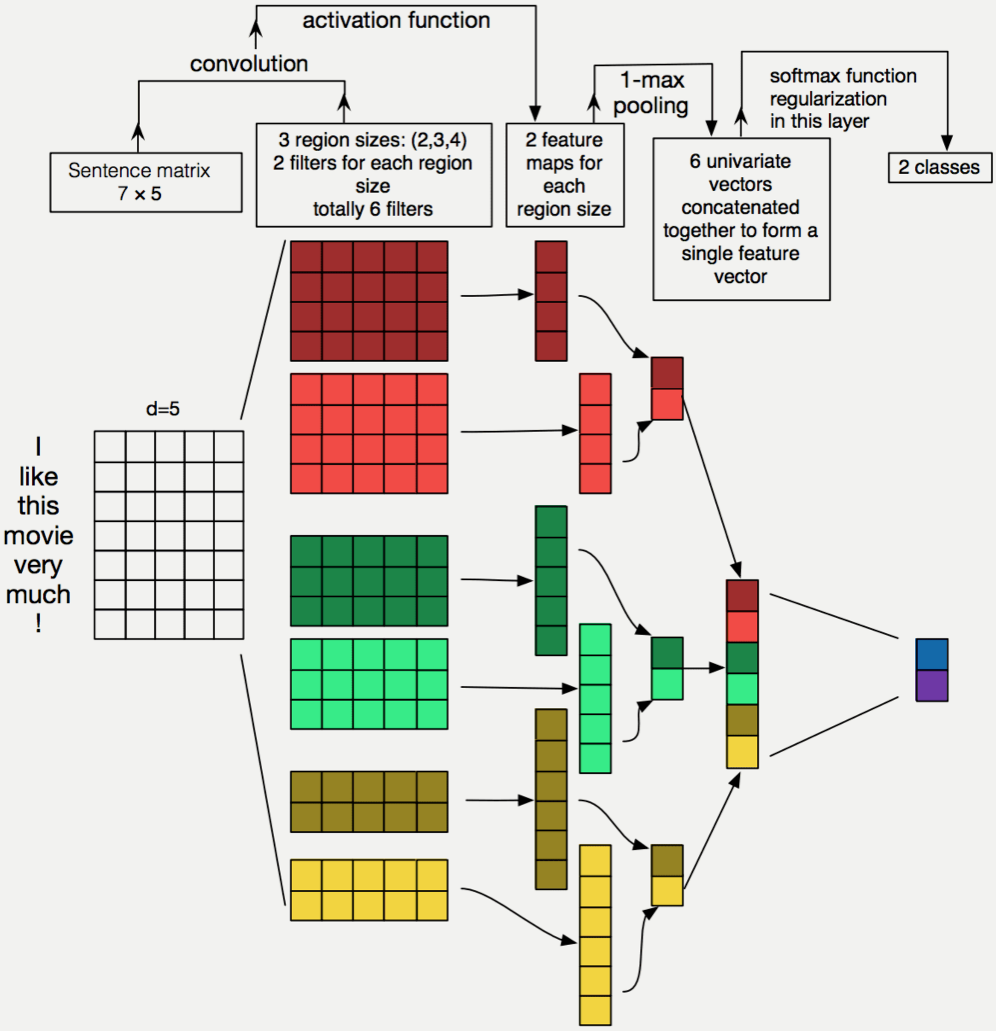
Ensemble methods, which are sophisticated techniques used in fact-checking, serve as a robust countermeasure to foreign information manipulation and interference. These techniques uniquely blend the outputs of diverse individual models or algorithms to make collective predictions, enhancing the accuracy and reliability of the fact-checking process.[52] Through this approach, ensemble methods effectively decrease bias, improve generalization, and deliver accurate and robust fact-checking results. A significant strength of ensemble methods lies in their diversity of model architecture. By employing a variety of architectures such as pre-trained language models, graph-based models, and deep neural networks among others, they introduce diversity into the ensemble. Each model, trained on different aspects of the data, captures and encodes unique patterns and relationships. For instance, pre-trained language models might excel at understanding the linguistic nuances of a claim, while graph-based models might be more effective at analyzing the relationships between entities involved in the claim.

Key techniques in ensemble methods include Bagging (Bootstrap Aggregating) and Boosting.[53] Bagging involves training multiple instances of the same model on

[52] Ahmad, Iftikhar, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. "Fake news detection using machine learning ensemble methods." Complexity 2020 (2020): 1-11.

[53] Gupta, Surbhi, and Munish Kumar. "Forensic document examination system using boosting and bagging methodologies." Soft Computing 24 (2020): 5409-5426.

different subsets of the training data, known as bootstrap samples. The predictions of individual models are then combined through methods such as voting or averaging, reducing the impact of outlier predictions. Boosting, on the other hand, aims to enhance the accuracy of weak learners - models that display modest performance. This is achieved by assigning more weight to instances that are misclassified. Through iterative boosting, which updates model weights based on the performance of previous iterations, the overall accuracy is improved. Ensemble methods also make use of sophisticated techniques such as Stacking and Random Forest.[54] In Stacking, predictions from several models are combined through a meta-model or a higher-level model. Here, the predictions of base models are used as features for the meta-model, allowing the ensemble to leverage multiple sources of information. The Random Forest method, which is an ensemble method based on bagging, utilizes multiple decision trees to make predictions, which reduces overfitting and enhances accuracy by averaging the predictions from individual trees.


activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Convolutional Neural Network Architecture for Sentence Classification. Source: Zhang, Ye, and Byron Wallace. "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification." ArXiv, (2015). Accessed August 3, 2023. /abs/1510.03820.

[54] AG, Priya Varshini, Anitha Kumari K, and Vijayakumar Varadarajan. "Estimating software development efforts using a random forest-based stacked ensemble approach." Electronics 10, no. 10 (2021): 1195.

Gradient Boosting Machines (GBM) and Adaboost represent further advanced techniques in ensemble methods[55]. GBM is a boosting technique where new models are trained to correct the errors made by previous models, thus iteratively improving accuracy by combining the predictions of multiple weak learners. Adaboost, a popular boosting algorithm, assigns weights to training instances and trains models iteratively to emphasize misclassified instances, hence enhancing accuracy by focusing on difficult-to-classify examples. Another implementation of gradient boosting is XGBoost (Extreme Gradient Boosting), which is optimized for speed and performance, and includes regularization and parallel processing, making it highly effective within ensemble methods. Notably, ensemble methods also allow diversity in data representation, combining predictions from models trained on different types of data such as text, images, and metadata, to improve accuracy and robustness. They also can estimate the uncertainty of predictions, providing confidence scores for fact-checking results. This approach is particularly valuable in the field of fact-checking, where presenting confidence in predictions can help to build trust with audiences.

Online ensemble learning further extends the capabilities of these methods by facilitating continuous updates and improvements to the ensemble as new data becomes available. This real-time adjustment and learning allow for real-time fact-checking and a rapid response to emerging misinformation. In summary, ensemble methods offer a powerful approach to enhance the accuracy of fact-checking. By leveraging these advanced methods and techniques, fact-checking organizations can achieve reliable and comprehensive results in their fight against foreign information manipulation and interference. By harnessing the collective intelligence of multiple models, ensemble methods enhance the resilience of fact-checking efforts and promote well-informed public discourse. The effectiveness of these methods can be further amplified through collaborative efforts among different fact-checking organizations, diversifying data sources, and model perspectives.

4. Neural Language Models for Contextual Understanding:

At the forefront of advancements in fact-checking and counteracting foreign information manipulation and interference, we find neural language models. These models bring together deep learning techniques and large-scale pre-training on copious amounts of text data, building a profound understanding of language context and semantics. With this enhanced contextual understanding, neural language models are equipped to accurately comprehend complex claims, identify nuanced misinformation, and effectively verify the veracity of information. Among the most advanced neural language models, Transformer-Based Models, such as BERT and GPT, are making substantial strides in neural language

[55] Verma, Pawan Kumar, Prateek Agrawal, Vishu Madaan, and Radu Prodan. "MCred: multi-modal message credibility for fake news detection using BERT and CNN." Journal of Ambient Intelligence and Humanized Computing (2022): 1-13.

understanding. By employing self-attention mechanisms, these models have the ability to capture relationships between words in a sentence, facilitating their understanding of long-range dependencies. The process of developing these models often involves two critical steps: pre-training and fine-tuning. Initially, these models are pre-trained on large-scale corpora of text data. Through this, they learn rich language representations and contextual embeddings. Following this, they undergo fine-tuning, wherein they are trained on specific fact-checking data, thereby adapting their language understanding abilities for the task of verifying claims.

Neural language models achieve a superior understanding of context through the generation of contextual word embeddings. These embeddings capture the meaning of words based on their surrounding context. This enables models to understand polysemous words and phrases, which results in better fact-checking performance. Some advanced neural models are capable of integrating multi-modal inputs, such as text, images, and metadata. This multi-modal contextual understanding provides a comprehensive interpretation of claims, combining various types of data for more accurate and holistic fact-checking. Attention mechanisms, employed by these models, enable them to focus on the most relevant parts of a sentence or document when processing information. This helps identify key elements in claims that require fact-checking and ensures better contextual understanding. Furthermore, these models are equipped with language generation techniques and adversarial training. Language generation can simulate and detect potential misinformation campaigns or generate counterarguments to false claims.

Adversarial training, on the other hand, can make the model robust against adversarial attacks and misinformation attempts.[56] Leveraging transfer learning, neural language models can apply knowledge from one domain, such as news articles, to another domain, such as social media posts. This cross-domain understanding is vital for fact-checking diverse sources of information to combat foreign information manipulation. Other capabilities include entity recognition for veracity assessment, where neural models identify key entities in claims and verify the information related to those entities. Zero-shot learning also plays a pivotal role as it enables models to make predictions for unseen or novel claims, even if they were not explicitly trained on those exact claims. This quick response to emerging misinformation is invaluable to fact-checkers. Expanding beyond a single language, these models can be fine-tuned for multiple languages, facilitating cross-lingual fact-checking.

Lastly, they can also undergo continuous learning, updating their knowledge with new data and fact-checking results to improve performance over time. This

[56] Tariq, Abdullah, Abid Mehmood, Mourad Elhadef, and Muhammad Usman Ghani Khan. "Adversarial Training for Fake News Classification." IEEE Access 10 (2022): 82706-82715.

supports real-time fact-checking and response to emerging misinformation campaigns. By tapping into the contextual understanding capabilities of neural language models, fact-checking organizations can significantly enhance their efforts in combating foreign information manipulation and interference. The models' ability to process context, semantics, and nuances in language improves the accuracy and efficiency of fact-checking, enabling the timely dissemination of reliable information to the public. Ensuring these models are fine-tuned and optimized for robust and ethical fact-checking practices requires collaborative efforts among researchers, language experts, and fact-checkers.

In conclusion, Machine Learning for Content Classification has significantly impacted fact-checking and information verification by automating the identification of misinformation, categorizing content, and prioritizing fact-checking efforts. NLP, supervised learning, active learning, and transfer learning are among the key technical aspects driving the success of these techniques. By leveraging advanced Machine Learning models, fact-checkers can process vast amounts of content rapidly and accurately, enhancing the efficiency and effectiveness of fact-checking initiatives. The continuous development of these technologies will continue to play a pivotal role in combatting misinformation and fostering a more informed and resilient information ecosystem.

DEEPFAKE DETECTION TECHNOLOGIES

Given the rise of deepfakes in FIMI campaigns, fact-checkers can adopt advanced deepfake detection technologies to identify manipulated audio and video content accurately. Deepfake Detection Technologies have become crucial in the fight against misinformation, particularly as the sophistication of deepfake technology continues to advance. By focusing on the technicalities of new technologies and techniques associated with Deepfake Detection, this extended analysis explores their impact on fact-checking and information verification.

1. Facial Recognition and Landmark Detection:

Facial recognition and landmark detection have emerged as critical elements in the battle against deepfake content and foreign information manipulation and interference. These techniques leverage the power of computer vision and deep learning to accurately identify and analyze facial features, making it easier to spot manipulated faces within deepfake content. Within the realm of deep learning, Convolutional Neural Networks (CNNs) have sparked a revolution in facial recognition tasks. These sophisticated models have the capacity to learn distinctive facial features and embeddings, which facilitates accurate face matching and identification.

Going a step further, ^{3}D facial recognition techniques utilize depth information to generate robust facial representations. This approach significantly improves performance under challenging lighting conditions and poses, making it a valuable tool in deepfake detection. To enhance accuracy and robustness in the identification of manipulated faces, multiple facial recognition models can be combined through ensemble methods. These ensemble models can detect manipulated faces by comparing them with multiple reference databases. Furthermore, pre-trained facial recognition models can be fine-tuned on deepfake detection datasets, specializing in the identification of manipulated faces. This practice, known as transfer learning, allows for efficient training, even with limited labeled deepfake data.

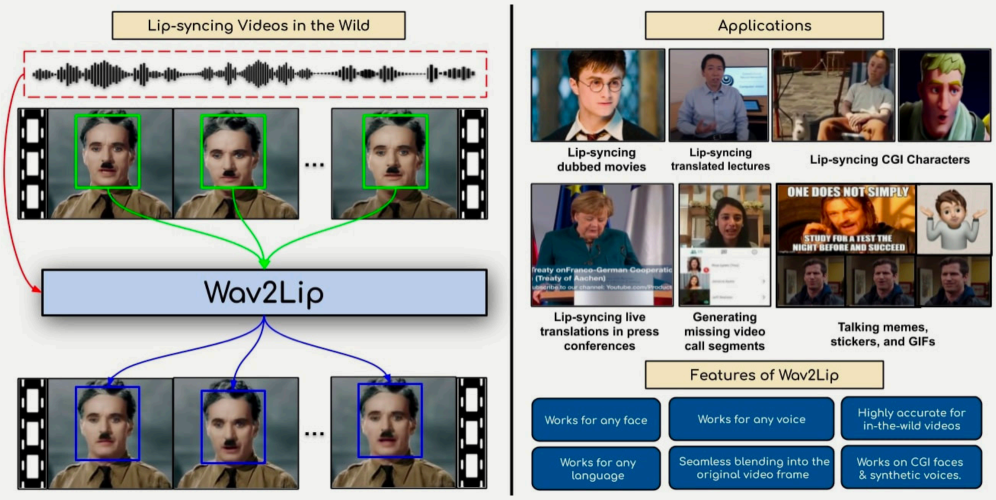
Attention mechanisms are also a powerful tool in this context. They can be used to highlight facial landmarks such as the eyes, nose, and mouth, which aids in accurate face analysis and landmark detection. For precise landmark localization, cascaded regression models are particularly effective. They iteratively refine the positions of facial landmarks, increasing the accuracy of detection. Moreover, real-time facial landmark detection in videos enables continuous tracking and analysis of facial expressions and movements. Similarly, ^{3}D facial landmark detection techniques, which analyze facial depth data, can accurately locate landmarks and offer robustness against ^{2}D image manipulations. For a direct comparison between faces, facial landmark alignment techniques can be used. They bring facial features into a standardized position, which significantly aids in the detection of deepfakes.

To improve robustness against adversarial attacks designed to evade detection, facial recognition models can be equipped with adversarial training. This form of training greatly enhances the adversarial robustness of facial recognition systems. For swift response to emerging deepfake content in real-time, it is crucial to optimize models for real-time inference, which enables rapid facial recognition and landmark detection. Furthermore, cross-domain techniques allow these models to generalize across different image and video sources, thereby ensuring reliable deepfake detection across diverse platforms. Finally, in a world increasingly concerned about privacy, privacy-preserving techniques can be incorporated to protect the identities of individuals during the facial recognition and landmark detection processes.

By harnessing these advanced facial recognition and landmark detection methods and techniques, fact-checking organizations can greatly enhance their deepfake detection capabilities. The ability to accurately identify and analyze manipulated faces allows fact-checkers to rapidly identify deepfake content and prevent its widespread dissemination. To adapt to emerging challenges in deepfake detection and manipulation, it is essential to foster collaboration among researchers, computer vision experts, and fact-checkers.

2. Lip-Sync and Speech Analysis:

Lip-sync and speech analysis are key components in the field of deepfake detection and in the fight against foreign information manipulation and interference. This is particularly the case when manipulated audiovisual content is being used to disseminate misinformation. Deep learning, natural language processing, and audiovisual analysis are leveraged in advanced methods and techniques to pinpoint discrepancies between lip movements and audio. This helps to expose and detect deepfake videos and audios.


Lip-syncing Videos in the Wild
Wav2Lip
Applications
Lip-syncing dubbed movies
Lip-syncing translated lectures
Lip-syncing CGI Characters
ONE DOES NOT SIMPLY
STUDY FOR A TEST THE NIGHT BEFORE AND SUCCEED
Treaty on Franco-German Cooperation (Treaty of Aachen)
Lip-syncing live translations in press conferences
Generating missing video call segments
Talking memes, stickers, and GIFs
Features of Wav2Lip
Works for any face
Works for any voice
Highly accurate for in-the-wild videos
Works for any language
Seamless blending into the original video frame
Works on CGI faces & synthetic voices.

A recent study has increased GAN audio with a 99% success rate. Source: K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

Lip-sync detection, which involves an analysis of the synchronization between the audio track and the lip movements of the speaker in a video, employs advanced deep learning models, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). These models are designed to capture temporal patterns and identify potential lip-sync errors, thereby providing a means of flagging manipulated content. Enhancing this process further, audio and visual features are combined in lip-sync analysis to provide a more comprehensive dataset for detection. Techniques like early fusion or late fusion enable the model to jointly process audio and visual data. This not only strengthens the accuracy of lip-sync detection but also provides a more robust defense against deepfake content.

On the speech analysis front, speech recognition models are used to convert audio to text, while Text-to-Speech (TTS) models transform text into synthesized speech. By comparing the recognized speech with the synthesized speech, discrepancies and potential deepfake manipulations can be detected more efficiently. Speaker verification and identification techniques further support the process of deepfake detection. Verification validates the identity of the speaker in the audio, while identification methods pinpoint known speakers. Both techniques are instrumental in assessing the authenticity of the audio in deepfake detection scenarios. Complementing these methods, voiceprint analysis creates unique voiceprints for individuals, enabling the identification of potential impersonations or voice manipulations. Additionally, an analysis of prosody and intonation patterns in the speech can reveal unnatural fluctuations or abnormalities that may indicate deepfake manipulation.

Moreover, cross-modal alignment allows for a direct comparison of lip movements and speech content by aligning the audio and visual data in time. This alignment further facilitates accurate lip-sync analysis. Perturbation techniques, applied to both the audio and visual components of deepfake content, assess robustness and authenticity. This can expose deepfake manipulations and provide additional layers of security against misinformation. The use of ensemble models in lip-sync and speech analysis enhances the reliability and effectiveness of deepfake detection by combining multiple models. This increases the robustness of the detection methods and contributes to an overall increase in the accuracy of the results. Importantly, these models are optimized for real-time inference, enabling the rapid analysis of lip-sync and speech patterns. This ensures quick responses to emerging deepfake content and helps prevent the spread of misinformation.

Advanced techniques also accommodate the analysis of content in multiple languages, enabling cross-lingual deepfake detection and helping to combat foreign information manipulation. Finally, adversarial robustness in speech analysis is bolstered by equipping speech analysis models with adversarial training techniques. This enhances robustness against adversarial audio attacks that might be used in deepfakes. These advanced methods and techniques in lip-sync and speech analysis provide fact-checking organizations with powerful tools to combat foreign information manipulation and interference.

The accurate identification of lip-sync discrepancies and speech abnormalities enables fact-checkers to quickly identify manipulated content and prevent the spread of false narratives. It is crucial that collaboration continues between researchers, NLP experts, audiovisual analysts, and fact-checkers to continuously refine and adapt these techniques to address the evolving challenges in deepfake detection and manipulation.

3. Behavior and Movement Analysis:

Deepfake Detection Technologies significantly utilize behavioral analysis as a means to identify unnatural movements and actions in videos. This form of detection is pivotal when combating foreign information manipulation and interference, particularly when manipulated videos aim to deceive by altering the behavior or movements of individuals. By leveraging computer vision, machine learning, and behavioral analysis, these advanced techniques compare the movements in a video with patterns expected in real-life scenarios. This approach aids in identifying anomalies, inconsistencies, and deepfake manipulations in videos, distinguishing between authentic and deepfake content.

A critical component of these techniques is pose estimation and tracking. Advanced pose estimation models can accurately detect and track key points on a person's body, such as joints and limbs, throughout a video. This tracking process enables the careful analysis of movement patterns, thereby identifying potential manipulations in body posture and behavior. Alongside this, gait analysis is utilized to examine an individual's walking pattern and rhythm. The advanced methods employed in this process can detect variations in gait that might be indicative of deepfake manipulation or impersonation. Further adding to the arsenal of deepfake detection tools is facial micro-expressions analysis. Facial micro-expressions, brief emotional expressions that can reveal subtle emotions and reactions, are carefully scrutinized. Advanced methods can identify changes in these micro-expressions that may indicate deepfake manipulation or inconsistencies in emotional responses.

The detection technologies also incorporate activity recognition models. These models can identify specific actions or behaviors in videos, such as talking, running, or gesturing. The analysis of these recognized activities plays a crucial role in detecting anomalies that might be present in deepfake content. This analysis is supplemented by emotion recognition techniques. By identifying the emotional state of individuals in videos and comparing these emotions with the context of the video, these techniques assist in assessing the authenticity of emotional expressions. Anomaly detection models are another vital tool in the fight against deepfakes. These models pinpoint rare or abnormal behavior and movements that deviate from expected patterns, aiding in identifying potential deepfake manipulations. Moreover, a contextual analysis of the video, such as the environment and interactions with other individuals, is employed to assess the realism and consistency of the scene. This, along with behavior profiling and biometrics, which create unique behavioral profiles for individuals and verify the identity of individuals based on behavioral traits, respectively, are crucial for identifying potential impersonations or manipulations.

Further augmenting the detection process are activity and behavior transfer detection techniques. These techniques can identify instances where the actions of one person have been transposed onto another in a manipulated video. The technologies also make use of ensemble models, which combine multiple behavior and movement analysis models. This combination enhances accuracy and robustness in deepfake detection, leading to a more reliable identification of manipulated content. To allow for quick responses to emerging deepfake content, these models are optimized for real-time inference. This enables the rapid analysis of behavior and movement patterns, preventing the spread of misinformation before it can cause damage. To ensure a comprehensive understanding of video content, multimodal behavior analysis is used. This involves the integration of multiple modalities, such as video, audio, and metadata, in behavior analysis. Lastly, to ensure adversarial robustness in behavior analysis, models can be equipped with adversarial training techniques. This enhances their robustness against adversarial attacks that may be used in deepfakes.

By leveraging these advanced methods and techniques in behavior and movement analysis, fact-checking organizations can significantly enhance their deepfake detection capabilities and combat foreign information manipulation and interference more effectively. The accurate identification of anomalies and inconsistencies in behavior and movement enables fact-checkers to quickly identify manipulated content, safeguarding against the spread of false narratives. To ensure these techniques continue to evolve with the challenges posed by deepfake detection and manipulation, it is essential for researchers, computer vision experts, behavioral analysts, and fact-checkers to continually collaborate and refine these techniques.

4. GAN Detection Techniques:

Deepfake detection, particularly the identification of Generative Adversarial Network (GAN) generated media, is an essential element in combating foreign information manipulation and interference. Since GANs are extensively employed in creating hyper-realistic deepfake content, there is a growing need for advanced methods and techniques capable of discerning the presence of such GAN-generated media. To meet this challenge, a combination of deep learning, statistical analysis, and forensic examination is being used to detect subtle GAN-specific artifacts and inconsistencies.

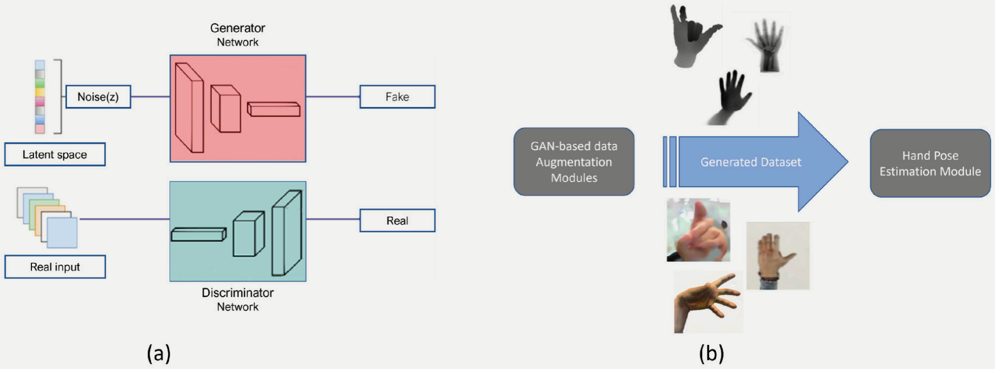
Adversarial feature learning is a cornerstone of these techniques. In this process, a separate classifier is trained to differentiate between GAN-generated and real media.[57] This classifier is capable of learning to recognize specific GAN artifacts and features, enabling it to accurately detect GAN-generated content. Complementing this process is the patch-based analysis technique, which dissects images or videos into smaller patches. Each patch is analyzed individually for GAN-specific patterns as GAN-generated patches often display unique artifacts, differing from those found in real media. To further refine the detection process, advanced statistical methods like Higher-Order Statistics (HOS) are used to identify irregularities and correlations in pixel patterns, indicative of GAN-generated content.[58] This is augmented by noise analysis which focuses on specific noise

[57] Wang, Can, Shangfei Wang, and Guang Liang. "Identity-and pose-robust facial expression recognition through adversarial feature learning." In Proceedings of the 27th ACM international conference on multimedia, pp. 238-246. 2019.

[58] Lee, Ji-Yeoun. "Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the saarbruecken voice database." Applied Sciences 11, no. 15 (2021): 7149.

patterns that may have been introduced during the synthesis process. Identifying inconsistencies in these noise patterns aids in distinguishing GAN-generated content from real media. Additional methods such as frequency domain analysis and color spaces analysis add another layer to this detection process.

GAN-generated images and videos might exhibit unusual frequency distributions in the Fourier domain, which can be identified through frequency domain analysis. Similarly, color spaces analysis detects specific color distribution patterns that might arise from the different color spaces utilized during the GAN generation process. Additionally, techniques have been developed to identify unique traces left by GANs, such as gradients and network activations that occur during the generation process. Analyzing these traces can yield valuable insights into the presence of GAN-generated content. Examination of metadata and compression artifacts also plays a crucial role in detecting GAN-generated content in video files.



A recent popular study that aims to automatically-detect GAN videos and images through generator and discriminator networks. Source: Farahanipad F, Rezaei M, Nasr MS, Kamangar F, Athitsos V. A Survey on GAN-Based Data Augmentation for Hand Pose Estimation Problem. Technologies. 2022; 10(2):43. https://doi.org/10.3390/technologies10020043

For robust and accurate detection, GAN classification ensembles are employed. These ensemble methods combine multiple GAN classifiers, each focusing on different aspects of GAN-specific artifacts. Moreover, techniques such as zero-shot GAN detection, which allows models trained on authentic data to identify unseen GAN-generated content without prior exposure to GAN examples, add to the sophistication of these detection methods. The demands of deepfake detection necessitate real-time responses, hence GAN detection models are optimized for real-time inference. This allows for the rapid analysis of media content, thereby enabling swift responses to emerging deepfake threats. To provide comprehensive detection capabilities, cross-modal GAN detection techniques have been developed that can analyze audiovisual content and metadata. Considering video-based deepfakes, advanced techniques focus on the temporal consistency across frames in videos to identify GAN-generated sequences.

This approach extends the scope of GAN detection beyond still images. By employing these advanced methods and techniques in GAN detection, fact-checking organizations significantly enhance their deepfake detection capabilities. The accurate identification of GAN-generated content enables fact-checkers to quickly identify manipulated media, thus preventing the spread of disinformation and safeguarding the integrity of the information ecosystem. To continue meeting the challenges posed by GAN-based manipulation and deception, collaboration among researchers, deep learning experts, and fact-checkers is essential.

In conclusion, Deepfake Detection Technologies are now a crucial part of fact-checking and information verification as deepfakes become increasingly sophisticated. They employ advanced image and video analysis, facial recognition, lip-sync analysis, and behavior analysis, all of which enable the accurate identification of synthetic media. The use of benchmark datasets and multi-modal approaches further enhance the effectiveness of deepfake detection, with explainability and real-time detection capabilities also being critical aspects. As deepfake technology continues to evolve, ongoing research and advancements in Deepfake Detection will be pivotal in maintaining the integrity of information and fostering a reliable and trustworthy digital ecosystem.

FACT-CHECKERS AND FACT-CHECKING IN TURKIYE

The largest and most comprehensive overview of the fact-checking ecosystem in Turkiye was previously published by EDAM.[59] This section will not go into the same level of detail with that extensive report, but will rather connect the discussions of this report to the broader fact-checking environment in Turkiye. Turkiye has a vibrant and dynamic media landscape, with a multitude of print, broadcast, and digital outlets catering to diverse audiences. However, this diversity also creates a breeding ground for misinformation and propaganda. Political polarization, censorship, and the government's control over media have further complicated the dissemination of accurate information. In this context, fact-checking has become increasingly important to separate truth from fiction and hold public figures and institutions accountable for their statements.

The demand for fact-checking in Turkey was initially triggered by a surge of medical hoaxes during the H1N1 virus outbreak in the late 2000s and then increased internet penetration. Later, due to political turmoil and a series of elections after 2014, another, second wave of fact-checkers emerged focusing on political verification as the country's political system grew more authoritarian. A third wave of pro-government fact-checkers then came into play, aimed at providing a counterbalance to what they perceived as an increasingly anti-government fact-checking ecosystem after 2018. Only two Turkish fact-checking platforms,

[59] Unver, Hamid Akin, Fact-Checkers and Fact-Checking In Turkey (2020). EDAM Research Reports, 2020, Available at SSRN: https://ssrn.com/abstract=3644285

Doğruluk Payı and Teyit.org, are members of the International Fact-Checking Network (IFCN), requiring them to adhere to transparency and editorial rules and undergo regular external audits.

Other groups like Evrim Ağacı and Yalan Savar also meet most IFCN criteria but are not members yet. Despite the rise of fact-checking platforms, a substantial portion of the public still do not use the internet for news or fact-check claims they encounter online. The ones that do often rely on traditional verification methods, such as consulting with friends and family or cross-checking with other news sources. However, this might change as the concept of fact-checking is still relatively new to Turkey. Though the relationship between fact-checking and disinformation remains inconclusive, some evidence suggests that fact-checking can counter high-profile disinformation cases.

Platforms like Doğruluk Payı, Evrim Ağacı, and Teyit.org shoulder most of the fact-checking burden, though they struggle with more complex and time-sensitive forms of disinformation. Nevertheless, the Turkish fact-checking ecosystem shows two promising trends. Firstly, fact-checkers have proven that they can survive, innovate, and serve the public even under adverse conditions, creating a demand for objective information. Secondly, despite political and legal pressures, they have been successful and politically sustainable, creating new verification protocols and engagement models that could serve as an inspiration to others, especially in countries experiencing democratic backsliding. However, the dependence on international funding for survival remains an issue, as it directly impacts a platform's ability to counter disinformation effectively. As such, the future course of the Turkish fact-checking ecosystem remains significant to international observers, and their progress, successes, and failures will continue to be relevant to fields such as comparative political communication, communication sociology, and technology-information.

Information manipulation in Turkiye often revolves around sensitive topics such as politics, religion, health, and national security. False narratives and rumors can spread rapidly on social media platforms, further exacerbating divisions within society and eroding public trust in the media. Additionally, the spread of misinformation during election periods can significantly impact voter perceptions and democratic processes. Fact-checking organizations in Turkiye play a critical role in addressing these challenges by providing evidence-based analysis and debunking false claims. They aim to equip citizens with accurate information and promote critical thinking, thus fostering a more informed and engaged society.
Teyit.org is one of Turkiye's pioneering fact-checking organizations, established in 2016. The word "teyit" means "verification" in Turkish, reflecting the organization's core mission. Teyit.org's team of journalists and researchers is committed to debunking false information, rumors, and misleading claims circulating in the media and on social platforms. Methodology: Teyit.org follows a systematic

approach to fact-checking, which includes claim selection based on public impact and credibility. They verify sources and collect relevant evidence to support or refute claims. The organization uses a rating system to categorize the accuracy of statements, ensuring clear and easily understandable summaries of their findings. Their fact-checking reports are transparent, explaining the methodologies and providing evidence for conclusions. Teyit.org addresses a wide range of topics, from political statements and viral social media posts to health-related claims and scientific misconceptions. They use open-source tools and data analysis techniques to identify patterns of misinformation and detect disinformation campaigns. One of Teyit.org's notable cases involved debunking false information related to COVID-19 during the pandemic. The organization fact-checked claims about potential cures, preventive measures, and the origins of the virus. By providing accurate information, Teyit.org contributed to public health awareness and the fight against misinformation during a critical period.

Doğruluk Payı: Founded in 2014, Doğruluk Payı, meaning "Share of Accuracy," is another prominent fact-checking organization in Turkiye. Its team is dedicated to verifying the statements of politicians, public figures, and media outlets. Methodology: Doğruluk Payı's fact-checking process involves rigorous source verification and evidence collection. They consult with subject matter experts for complex cases to ensure accuracy. The organization uses a rating system from "True" to "False" to categorize statements based on their accuracy. Doğruluk Payı has been actively involved in fact-checking political claims, especially during elections. They strive to maintain objectivity and independence despite pressures from various political actors. A specific case that Doğruluk Payı tackled involved fact-checking claims made by opposing political parties during a national election. The organization meticulously analyzed campaign promises and policy statements, providing voters with an unbiased assessment of the accuracy of these claims. By doing so, Doğruluk Payı empowered voters to make informed decisions based on reliable information.

Both Teyit.org and Doğruluk Payı are signatories of the International Fact-Checking Network (IFCN), a global alliance of fact-checking organizations. The IFCN, a project of the Poynter Institute, aims to promote excellence in fact-checking and strengthen the collaboration between fact-checkers worldwide. Through their affiliation with the IFCN, Teyit.org and Doğruluk Payı gain access to resources, best practices, and support from an international community of fact-checkers. This collaboration enhances their credibility and ensures that their fact-checking methodologies align with global standards. Additionally, IFCN's partnership with Facebook enables fact-checkers to review and debunk potentially false information on the social media platform, thereby reaching a broader audience. The IFCN fosters collaboration among its member organizations through regular virtual meetings, webinars, and workshops. Fact-checkers from different countries share knowledge and experiences, discuss emerging trends in misinformation, and explore innovative fact-checking tools and technologies.

Fact-Checking Europe is a collaborative project involving various fact-checking organizations from different European countries, including Turkiye. Teyit.org is an active participant in this network, contributing to cross-border fact-checking efforts and information sharing. Collaborative Initiatives: Fact-Checking Europe's initiatives include joint fact-checking projects during elections and major events, where fact-checkers from multiple countries work together to verify claims and counter false narratives. By collaborating with partners from other European countries, Teyit.org strengthens its fact-checking capacity and extends its reach to a wider audience. The collaboration also facilitates the exchange of data and research findings between fact-checking organizations. By working together, the member organizations can leverage each other's expertise and enhance the overall effectiveness of their fact-checking efforts.

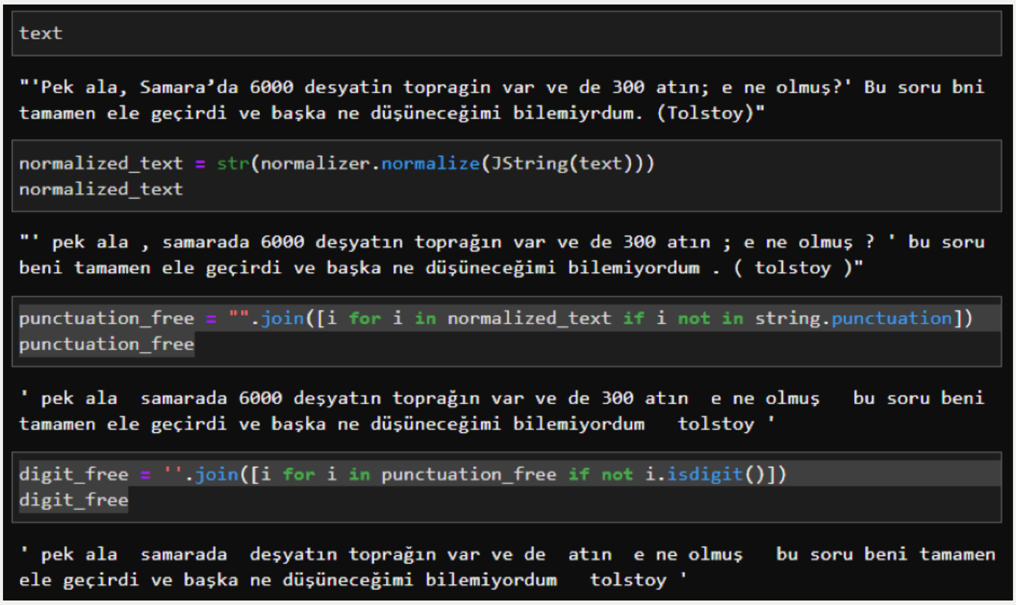
One of the most significant challenges facing fact-checking organizations in Turkiye is political pressure from government authorities. The government's tight control over the media has created an environment where fact-checkers may face intimidation or threats when debunking claims made by powerful figures or the ruling party. For example, Teyit.org faced numerous backlashes when it debunked statements made by high-ranking government officials during election periods. The organization's exposures of disinformation led to debunked officials' supporters lunching several smear campaigns against Teyit.org on social media, accusing the organization of bias and serving the interests of opposition parties.

NLP Tools Available to Turkish Fact-Checkers:

There are recent libraries and classifiers in Turkish that form the frontiers of NLP research in Turkish-language fact-checking.[60] These sentiment, mood and classifier libraries, including the Movie Sentiment Dataset,[61] ITU NLP Turkish Sentiment Analysis Dataset,[62] and Turkish Product Sentiment Analysis Dataset,[63] are just some of the most recent text-based tools available for the use of fact-checkers. These datasets can help machine learning models to better understand and interpret the sentiments expressed in Turkish text.

Additionally, some of the most recent advanceds feature valuable Named Entity Recognition datasets such as the BOUN Named Entity Recognition Dataset,[64] Turkish Wikipedia Dump[65] and 'Mukayese'[66] – a benchmarking platforms for Turkish NLP libraries – which can be useful for tasks like language modeling and text classification. Comprehensive and novel sets of stop words that can be

[60] A frequently updated list of Turkish NLP tools can be found in Gökçe Merdun's and Turkish NLP Suite GitHub page: https://github.com/agmmnn/turkish-nlp-resources and https://github.com/turkish-nlp-suite (all links from here onwards are accessed latest on 3 August 2023)
[61] http://humirapps.cs.hacettepe.edu.tr/tsad.aspx
[62] http://tools.nlp.itu.edu.tr/api_usage.jsp
[63] https://github.com/turkish-nlp-suite/Vitamins-Supplements-NER-dataset
[64] https://universaldependencies.org/treebanks/tr_boun/index.html
[65] https://www.kaggle.com/datasets/mustfkeskin/turkish-wikipedia-dump
[66] https://github.com/alisafaya/mukayese



```
text

"'Pek ala, Samara'da 6000 desyatin topragin var ve de 300 atın; e ne olmuş?' Bu soru bni
tamamen ele geçirdi ve başka ne düşüneceğimi bilemiyrdum. (Tolstoy)"

normalized_text = str(normalizer.normalize(JString(text)))
normalized_text

"' pek ala , samarada 6000 deşyatın toprağın var ve de 300 atın ; e ne olmuş ? ' bu soru
beni tamamen ele geçirdi ve başka ne düşüneceğimi bilemiyordum . ( tolstoy )"

punctuation_free = "".join([i for i in normalized_text if i not in string.punctuation])
punctuation_free

' pek ala  samarada 6000 deşyatın toprağın var ve de 300 atın  e ne olmuş   bu soru beni
tamamen ele geçirdi ve başka ne düşüneceğimi bilemiyordum   tolstoy '

digit_free = ''.join([i for i in punctuation_free if not i.isdigit()])
digit_free

' pek ala  samarada  deşyatın toprağın var ve de  atın  e ne olmuş   bu soru beni tamamen
ele geçirdi ve başka ne düşüneceğimi bilemiyordum   tolstoy '
```

Python code sample for Zemberek. Screen capture from: https://medium.com/technology-hits/turkish-text-preprocessing-with-zemberek-in-python-35930cc79afa Source: Akın, Ahmet Afsin, and Mehmet Dündar Akın. "Zemberek, an open source NLP framework for Turkic languages." Structure 10, no. 2007 (2007): 1-5.

beneficial for text preprocessing in NLP tasks are also in development. Stop words, often deemed as low value words, are usually filtered out in NLP processes. This repository provides two separate lists of Turkish stop words that could be beneficial for different forms of fact-checking and automated verification.

In terms of tools and libraries, there are an assortment of resources for performing NLP tasks in Turkish. 'Zemberek',[67] for example, is an open-source NLP library developed specifically for the Turkish language. It provides a multitude of utilities such as tokenization, morphology, spell checking, and more. 'Turkish Deasciifier',[68] a tool that can convert a Turkish text written using only ASCII characters into a correctly spelled version using the appropriate Turkish characters is also worth mention. For fact-checkers interested in Word Embeddings, 'Turkish Word2Vec'[69] offers pre-trained word vectors for Turkish. As NLP research continues to expand, so does the potential for such resources to contribute significantly to new discoveries and advancements in Turkish language processing.

CONCLUSION

The rapid evolution of the digital landscape has amplified the significance of accurate and trustworthy information. Indeed, the prevalence of misinformation and disinformation has become a challenge of epic proportions that reverberates across all sectors of society, impacting political, social, and economic stability on a global scale. Advancements in technology, particularly artificial intelligence and machine learning, have emerged as powerful tools in the battle against the spread of false information. They offer the capability to process vast amounts of data at speed, uncover patterns, and make predictions. In the context of fact-checking, these technologies provide the backbone for tools and systems designed to verify the authenticity and credibility of information.

However, the power of these technologies is a double-edged sword, as they are also used to create highly sophisticated and convincing false content such as deepfakes. Artificial neural networks, for instance, form the basis for numerous advanced fact-checking models. These models can sift through enormous amounts of textual data, identify claims, and evaluate their truthfulness. In the world of visual media, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been effectively used to spot manipulated content. Simultaneously, advanced face and voice authentication methods are being developed to verify the authenticity of individuals in images and videos, safeguarding against impersonations. Parallel fact-checking datasets and multilingual models have enabled the expansion of these verification techniques across different languages and cultural contexts, allowing the development of universal fact-checking tools. This has brought a new level of sophistication to multilingual fact-checking, as seen in Canada's handling of bilingual fact-checking during its elections. The advent of real-time alerts and early warning systems, powered by AI, has also enhanced the efficacy of fact-checking initiatives. By flagging potentially misleading or false information promptly, these systems enable fact-checkers to respond swiftly and stem the spread of misinformation. These techniques, paired with state-of-the-art media forensics methods, have further advanced the accuracy and speed of information verification, making it possible to quickly debunk doctored images or deepfake videos.

Crowdsourcing platforms and the engagement of the public in fact-checking have also played a pivotal role. Aided by technology, these platforms can engage a wide community of users, promoting collective intelligence in the verification process. The Europe-wide fact-checking platform "FactCheckEU" exemplifies this collaborative approach. Fact-checking organizations have recognized the power of collaboration and have started working closely with social media platforms. This collaboration has allowed the immediate flagging and debunking of false information, thereby reducing its spread and impact. Facebook's partnership with third-party fact-checking organizations and Twitter's Birdwatch are prime examples of this strategy.

The geopolitics of misinformation cannot be understated. As the digital space becomes the primary medium for information dissemination, the ability to spread misinformation has been weaponized, leading to the manipulation of public opinion, interference in democratic processes, and even inciting social unrest. Misinformation can ignite and fuel conflicts, destabilize nations, and shift geopolitical power dynamics. It can undermine trust in institutions and in democracy itself. The advancements in fact-checking and information verification technologies are key in this battlefield. By promoting the circulation of accurate information, these technologies contribute to the preservation of democracy, prevent the incitement of conflicts, and enhance societal resilience against information manipulation. However, the continued refinement of these technologies is paramount. As misinformation techniques advance, so must the technologies used to combat them. This necessitates continuous research, development, and investment in fact-checking and verification technologies.

Governments, organizations, and technology companies must collaborate closely, sharing resources and knowledge, to effectively combat the misuse of information. Furthermore, regulations must be put in place to ensure the ethical and responsible use of these technologies, guarding against potential misuse while also protecting freedom of speech. The development and deployment of fact-checking and information verification technologies must be transparent and accountable. In conclusion, the advancements in fact-checking and information verification technologies are having profound impacts on geopolitics, influencing the flow and trustworthiness of information on a global scale. The role of these technologies in preserving truth, promoting informed public discourse, and safeguarding democracy is of paramount importance.

Ekonomi ve Dış Politika
Araştırmalar Merkezi
edam

Address : Hare Sokak NO:16 AKATLAR 34335 İstanbul/Türkiye
Phone : +90 212 352 18 54
Fax : +90 212 351 54 65
Email : info@edam.org.tr