Challenges for Automatic Detection of Fake News Related to Migration

Invited paper

Werner Bailer, Georg Thallinger
DIGITAL JOANNEUM RESEARCH Forschungsgesellschaft mbH
Graz, Austria
{firstname.lastname}@joanneum.at

Gerhard Backfried, Dorothea Thomas-Aniola
HENSOLDT Analytics GmbH
Vienna, Austria
{firstname.lastname}@hensoldt-analytics.com

***Abstract*—Fake news and misinformation is a widespread phenomenon these days, affecting social media, alternative and traditional media. In a climate of increasing polarization and perceived societal injustice, the topic of migration is one domain that is frequently the target of fake news, addressing both migrants and citizens in host countries. The problem is inherently a multi-lingual and multi-modal one in that it involves information in an array of languages, material in textual, visual and auditory form and often involves communication in a language which may be unfamiliar to recipients or which these recipients only may have basic knowledge of. We argue that semi-automatic approaches, empowering users to gain a clearer picture and base their decisions on sound information, are needed to counter the problem of misinformation. In order to deal with the scale of the problem, such approaches involve a variety of technologies from the field of Artificial Intelligence (AI). In this paper we identify a number of challenges related to implementing approaches for the detection of fake news in the context of migration. These include collecting multi-lingual and multi-modal datasets related to the migration domain, providing explanations of AI tools used in verification to both media professionals and consumers. Further efforts in truly collaborative AI will be needed.**

Keywords—fake news, migration, content verification, data sets, multilinguality

I. INTRODUCTION

Disinformation and misinformation, frequently referred to as fake news, are widespread phenomena these days, affecting social media, alternative and traditional media. Issues with high societal, cultural or economic impacts typically provide the backdrop for such contents. It is evident that topics creating strong emotions and anxieties among audiences are particularly affected by fake news. In many regions of the world, the topic of migration features prominently among those domains.

Fake news related to migration may target migrants in their countries of origin, on their journey or in host countries. For example, InfoMigrants[1] documents and verifies a number of stories in multiple languages in order to counter misinformation targeting migrants at different points of their journey. Other examples are reports about fears among migrants caused by fake news about COVID-19, both in refugee camps[2] and in Europe[3]. In addition, fake news may target the majority population in host countries, for example, claiming that scenes of migrants drowning are merely staged (by using video footages from a completely different context[4]). Especially in settings of increasing polarization and perceived societal injustice, they find fertile grounds for reception. The widespread use of social media acts as a further amplifier and rapid distribution channel in this process.

The problem of fake news is inherently a multilingual and multi-modal one in that it involves information in an array of languages, material in textual, visual and auditory form, and often involves communication in a language which may be unfamiliar to recipients or which these recipients may only have basic knowledge of. Due to the scale of the problem, we argue that semi-automatic approaches, empowering users to gain a clearer picture and base their decisions on sound information, are needed. These tools involve a variety of technologies from the field of Artificial Intelligence (AI).

In this paper we identify a number of challenges related to this problem. We first review related work on verification tools and data in Section II, and discuss relevant technologies and their maturity for this application area in Section III. In Section IV, we formulate the challenges ahead. Section V concludes the paper.

II. RELATED WORK

In this section we review related work on verification tools and datasets, while the relevant technical building blocks are discussed in Section III.

A. Verification tools

Existing verification tools target either media consumers or professionals such as journalists. While end users require fully automated solutions, professionals require tools that provide automation support for mundane tasks, but leave the final assessment to the professional user. For consumers, a number

[1] https://www.infomigrants.net/en/tag/fake%20news/
[2] https://www.bbc.com/news/av/world-africa-53695376
[3] https://theconversation.com/coronavirus-misinformation-is-leading-to-fake-news-anxieties-in-dutch-refugee-communities-141830
[4] https://www.mimikama.at/aktuelles/filmteam-inszenierte-keine-ertrinkenden-fluechtlinge-in-griechenland/

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

of verification tools and sites are available. For example, the European InVid project developed a browser plugin[5] to enable consumers to verify content they find on the web. FactStream[6] is an app providing verification for consumers, tapping into different verification sources. TruthNest[7] provides credibility analytics of Twitter accounts. Buster.AI[8] is another verification service aimed at consumers, available as a browser plugin, web portal or via API to be integrated into third parties' products.

The situation for the professional sector is slightly different. The authors of [1] state that about half of journalists worldwide use social media sources, while this applies to the majority of journalists in Europe and the US (e.g., 96% in the UK). As found by a recent study commissioned by the European Parliament [2], AI-based verification tools are still only in experimental use in media organizations. Most verification strategies rely on traditional workflows, e.g. talking to trusted sources and eyewitnesses, while the use of tools is only applied in some cases. The LiT.RL News Verification Browser [3] is a tool that provides an assessment of the credibility of content based on linguistic analysis. Further tools target more collaborative verification rather than only consuming verification information. Truly Media[9] is a collaborative platform for content verification aimed at professionals such as journalists. WeVerify[10] offers a similar concept, aiming to bring together professionals and citizen journalists. The aspect of collaborative use of AI to support cooperation between humans and automated components still needs further research. This aspect is envisaged to provide tremendous potential and is crucial to deploy efficient workflows for professional users.

Other types of tools address the analysis of videos and video verification, focusing on forensic techniques to detect alterations. Examples are DeepFakeNet[11], Quantum Integrity[12] and Sensity[13]. A more complete analysis of the subject, addressing forensic techniques, content provenance and contextual semantic aspects is provided in [4].

When verification sites provide explanations for their decisions, those are often created by humans. The currently emerging topic of explainable AI (XAI) could be beneficial for verification tools in order to provide the basis for users to make their decisions based on the reasoning provided by the automatic algorithm. Some early works in this field have been published. In [5], a fake news detection system is complemented by an attention module which highlights the sentences in the sources that contributed to the assessment. The authors of [6] assess news items based on 10 features derived from the content, source and context. Shapely values[14] are used to describe the influence of each feature on the decision. In a similar way, [7] visualize the influence of features and deploy a tree visualization of supporting data samples. An approach for assessing the factfulness of tweets using their retweet data is described in [8]. Explanations are generated based on per-word attention values.

B. Datasets

Data-driven approaches, like machine learning methods (ML) commonly used today, depend critically on the availability of (annotated) datasets for training and validation. The growing interest in AI-based fake news detection has led to the creation of datasets, typically consisting of verified news items. While there seems to be a solid basis for developing fake news detectors, it turns out that many of the datasets are created with data from a rather limited number of fact checking sites. The robustness of the resulting models and their ability to generalize to further domains is thus very limited. Datasets like the ISOT Fake News Dataset[15], FNID[16] and LIAR[17] are based on PolitiFact. FakeNewsNet[18] also uses PolitiFact[19] in combination with GossipCop[20]. FNC-1[21] is another US focused dataset. NELA-GT[22] and FakeNewsCorpus[23] are datasets mined from a number of media websites, focusing mainly on US politics.

[9] propose an image dataset called Fauxtography, mined primarily from the US site snopes.com. FakeEddit[24] is a multimodal dataset collected from 22 different subreddits on Reddit, again mostly focusing on US politics. CREDBANK [10] is a dataset mined from Twitter's streaming API, and assessing the credibility of claims contained in tweets. "Some like it hoax" [11] is another social media dataset, mined from facebook pages related to conspiracy theories. Other datasets focus specifically on fakes[25] generated by (deep) neural networks. GermanFakeNC[26] is a dataset including a number of news related to crimes allegedly committed by migrants. FAKES [12] is an English language dataset containing (fake) news items related to the Syrian war.

To conclude, only a few datasets exist that seem relevant in the context of migration-related fake news. While some datasets cover at least a few relevant aspects targeting the majority population in host countries, datasets targeting fake

[5] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/
[6] https://www.factstream.co/
[7] https://www.truthnest.com/
[8] https://buster.ai/
[9] https://www.truly.media/
[10] https://weverify.eu/
[11] https://www.fakenetai.com/
[12] https://quantumintegrity.ch/
[13] https://www.sensity.ai
[14] https://en.wikipedia.org/wiki/Shapley_value
[15] https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php
[16] https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset
[17] https://github.com/thiagorainmaker77/liar_dataset
[18] https://github.com/KaiDMML/FakeNewsNet
[19] https://www.politifact.com/
[20] https://www.gossipcop.com/
[21] https://github.com/FakeNewsChallenge/fnc-1
[22] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ULHLCB
[23] https://github.com/several27/FakeNewsCorpus
[24] https://fakeddit.netlify.app/
[25] https://github.com/agermanidis/OpenGPT-2
[26] https://zenodo.org/record/3375714#.YBCHZNYxkng

news aimed at migrants, and in the relevant languages, are scarce if not absent.

III. Relevant Technologies

A. Metadata and sources

Algorithmic approaches to detect fake news typically combine factors concerning the credibility or reliability of sources and the veracity of content. These two groups of factors pose different challenges requiring different sets of technologies. Sources may be accommodated within a model of trust, relating sources and authors with each other. Such models allow relating information across multiple sources (potentially in multiple languages) and thus provide a richer context for verification.

The following aspects may serve as indicators regarding the level of credibility of sources:

- Trust in source/publisher (track record)
- Reputation, classification of source
- Trust in referenced sources
- Coverage of topic in other sources
- Publication by established publisher
- Verification of publisher's credentials
- Named, verifiable author
- Association with political actors
- Verifiable Internet footprint of publisher
- Number of advertisements
- Spamminess of advertisements
- Placement of advertisements
- Detection of social calls (share, like, ...)

Meta-data associated with publishers and documents may likewise serve as indicators of reliability. In particular elements regarding location and time as well as elements which can be used for cross-checks against the contents and other documents (such as the use of stock- or standard-images as profile-images, unusual ratio of followers: followees, timezones and locations) might provide valuable context and insights.

Social Media (SM) platforms provide a variety of meta-data associated with both actors as well as documents. In particular with regard to the reach and impact of a SM post, these meta-data may be key factors. Indicators pertaining to the account itself (verified user, image, age of account, ...) as well as related to user activity (bot-like behavior, following and follower relationships, likes, mentions, references, ...) may provide information concerning the credibility and reliability of SM accounts and posts.

An approach for verifying the claimed GPS location of an image based on a reference image set using neural networks is described in [13].

B. Technologies for text- content

In order to analyze textual content, a set of technologies from the area of Natural Language Processing (NLP) has to be employed. Not only do these technologies need to be robust enough to cope with non-standard language, different and mixed scripts, multiple languages, language-varieties and dialects, they also need to be able to deal with incomplete and incorrect inputs, in particular when processing textual contents from social media. A series of processing steps, from ingesting the input text, to finally producing an enriched sequence of elements needs to be carried out. [14] provide an overview of the associated challenges and the technologies required to cope with those challenges. In case of audio inputs, the content may be transcribed using Automatic Speech Recognition technology to produce content in textual format. ASR itself is associated with a series of challenges and the produced textual output requires special attention for further processing. [15] identify different categories of text-processing technologies which can be employed to determine a measure of trustworthiness of a text. This measure may subsequently be combined with further factors, such as trust in sources and authors, to arrive at an overall measure of trustworthiness. The following elements may serve as indicators in this mix:

- Elements identifying the source or author
- References to unnamed or unknown sources
- References to external sources
- Quotes and citations
- Stylistic measures, text layout, typos
- Text complexity measures
- Coherence between headline and body
- Number, diversity and proximity of mentioned entities
- Geographical and temporal references
- Emotional tone, sentiment (polarity as well as intensity), stance, exaggerations
- Clickbait in title and lead-in paragraphs
- Detection of previously published material / originality of content
- Detection of elements of known conspiracy theories

Automatic translation may allow comparing facts and indicators across sources in different languages (and different scripts).

In addition, visual elements included in the text (and related to the text) may yield further insights:

- Identification of duplicate/re-used images
- Fraudulent imagery
- Out-of-content images or videos

Furthermore, technologies to detect logical fallacies, inconsistent inferences or contradictions may be used. However, the latter are only available at a rudimentary stage requiring further work.

All of the above factors apply to the detection of fake news in general as well as to misinformation and disinformation in the domain of migration in particular. Domain-specific information about content (terminology, formulations, slang and code) as well as information about the involved actors (sources, SM accounts and news outlets) may be used to focus and adjust the general mechanisms to the domain of migration. Furthermore, geographic, cultural and linguistic adaptation and specialization can be carried out for effective application of technologies and methods. However, only little work has been carried out to this end leaving ample room for future research.

C. Technologies for audiovisual content

Provenance analysis for audiovisual content aims to track back the original source of a media item. While this does not answer the question of authenticity, it provides insight whether the media item has been modified, the claimed authorship is correct or it has been used in a different context. The main challenges are handling the scale of visual media items on the web and the robustness against transformations, such as re-encoding or cropping.

A scalable approach for image provenance detection is provided in [16], using clustering of similar images in the provenance graph. The paper also proposes a dataset mined from Reddit. [17] address the scalability issue by inferring possible provenance relations from metadata and thus significantly reduce the number of content matches that need to be performed. For video, this problem (also known as near-duplicate video retrieval) is even more computationally demanding [18]. In order to facilitate scalable matching of video content, approaches for compact descriptors (based on both hand-crafted and learned features) have been proposed. One example are the descriptors standardized in the MPEG CDVA standard [19][20].

For audiovisual content, the mismatch between the visual and audio modality can be exploited. [21] propose a system that determines the likelihood of match between both modalities, using both low-level and affective features.

D. Multimedia forensics

Multimedia forensic methods are used to identify whether a media item has likely been modified. The application of these approaches is often limited by the fact that no authentic copy may be available, but only other sources from (social) media that have gone through various processing chains during upload and distribution.

Although no authentic version may be available, there are still methods to detect whether an image is composed of different parts, known as splice detection. A recent method [21] predicts whether different parts of the same image could have been produced by a single imaging pipeline, and whether these settings match the EXIF data of the image. Splice detection is in particular relevant in connection with provenance analysis (see above), where it may be possible to identify the inputs that were used to generate the manipulated image. In [22], the authors propose to exploit relations between regions of images with disjoint groups of source images to detect image forgery.

Face verification is an approach to check whether a depicted person is the one claimed, and a range of methods exist [23]. Recently, fueled by the attention around deep fakes, the detection of whether face regions in images have been modified [24][25], have gained much attention. However, the majority of fakes and misinformation is still rather based on putting images out of context, rather than using sophisticated tools. Thus more general face recognition and verification tools can provide valuable information concerning the match between depicted persons and names given in contextual information.

E. Multimodal technologies

Effective fake detection requires the fusion of all cues from textual, audiovisual and contextual (e.g., metadata, source) information. This includes mismatch between visual and textual content, file metadata or the thread in which content appears. This is a very active research topic, and recent works have shown advances in this area (e.g., [26]), using also emerging approaches such as graph neural networks [27].

The automatic generation of image and video captions is another topic that has been actively researched recently. Checking the validity of generated captions is a similar problem as verifying provided captions. In a recent work proposing a verification framework for discriminative models [28] its application to verifying caption generation is mentioned. Such approaches could also be used for multimodal content verification in future.

IV. CHALLENGES

We have identified a number of challenges for providing automatic verification of migration-related news content at scale. These challenges involve collecting multi-lingual and multi-modal datasets related to the migration domain, which would help to develop and validate AI tools that can support the detection of fake news. Another aspect is providing explanations of AI tools used in verification to both media professionals and consumers. In order to realize a workflow with a human user in the loop, further efforts in truly collaborative AI will be needed. The challenges are discussed in more detail in this section.

A. Datasets

Currently most datasets related to fake news detection are in English and focus on domestic topics in Western countries, predominantly the United States. Although some of the fake news in these datasets include migration-related topics, targeting majority populations in Western countries, the domain is not well covered. In addition, datasets covering fake news targeting migrants, and datasets in other languages than English are scarce.

In order to foster research, it is highly relevant that the community creates relevant datasets, and organizes benchmarks/challenges around migration-related fake news. These should cover multiple languages and have a geographic scope reflecting the phenomenon of migration.

B. Explainability

There is only preliminary research on explainability for content verification tools. Providing explanations of AI decisions is always highly dependent on the target group. For example, the explanation of an AI-supported diagnostics tool will be very different whether it is aimed at the patient or the doctor. Similarly, explanations for fake news verification aim at a very wide range of people, with strong differences in technical and media literacy, and possibly limited knowledge of the language in which the fake content is provided.

For professional users, the need is to enable better human-AI collaboration, i.e. workflows that interlink automatic and manual workflow steps, and AI methods that are able to learn "on the job" from their human collaborators. Current

verification tools often consist of an automated information extraction process, followed by review and validation by a human user. Future AI-based tools need to be able to directly learn from human user interactions to update their models and reassess content based on newly learned facts or to provide functionalities to interactively drill down on facts uncovered during the analysis.

C. *Tool availability and deployment*

While finally citizens are the consumers of verification results, media professionals play a pivotal role in fact checking, and preparing information. For the majority population this process can rely on an existing ecosystem, although the eroded trust in media companies requires new approaches. In order to address migrant communities, such tools and processes also need to be available to (small) media organizations serving these communities, and in a variety of languages. In order to enable cost-effective scalability, trusted translation tools (supporting also less common languages) play an important role.

V. CONCLUSION

In this paper we have reviewed existing verification tools and datasets in order to assess their suitability for handling fake news related to migration, which may address both migrants and the majority population in target countries. In particular datasets and tools for identifying misinformation aimed at migrants (and thus in various relevant languages) are clearly underrepresented in current research. We also surveyed technologies that may be relevant to automate the process of fact checking. We identified and described open challenges related to data sets, explainability for the respective user groups and the deployment/availability of these tools. Addressing these challenges in future research is crucial in order to counter misinformation related to migration. In particular, truly collaborative human-AI approaches will be required in order to combine the strengths of scalable automatic approaches and human supervision.

ACKNOWLEDGMENT

This work was partly supported by the EU Horizon 2020 project AI4Media, under contract no. 951911.

REFERENCES

[1] Brandtzaeg, Petter Bae, et al. "Emerging journalistic verification practices concerning social media." Journalism Practice 10.3 (2016): 323-342.

[2] G. Rehm, Research for CULT Committee - The use of Artificial Intelligence in the Audiovisual Sector, PE 629.221, May 2020.

[3] Rubin, Victoria, et al. "A news verification browser for the detection of clickbait, satire, and falsified news." The Journal of Open Source Software 4.35 (2019): 1.

[4] Nixon, Lyndon, Symeon Papadopoulos, and Denis Teyssou. Video Verification in the Fake News Era. Ed. Vasileios Mezaris. Springer, 2019.

[5] Cui, Limeng, et al. "defend: A system for explainable fake news detection." Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.

[6] Reis, Julio CS, et al. "Explainable machine learning for fake news detection." Proceedings of the 10th ACM conference on web science. 2019.

[7] Yang, Fan, et al. "XFake: explainable fake news detector with visualizations." The World Wide Web Conference. 2019.

[8] Lu, Yi-Ju, and Cheng-Te Li. "GCAN: Graph-aware co-attention networks for explainable fake news detection on social media." arXiv preprint arXiv:2004.11648 (2020).

[9] Zlatkova, Dimitrina, Preslav Nakov, and Ivan Koychev. "Fact-Checking Meets Fauxtography: Verifying Claims About Images." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[10] Mitra, Tanushree, and Eric Gilbert. "Credbank: A large-scale social media corpus with associated credibility annotations." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 9. No. 1. 2015.

[11] Tacchini, Eugenio, et al. "Some like it Hoax: Automated fake news detection in social networks." 2nd Workshop on Data Science for Social Good, SoGood 2017. CEUR-WS, 2017.

[12] Salem, Fatima K. Abu, et al. "Fa-kes: A fake news dataset around the syrian war." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 13. 2019.

[13] Cheng, Jiaxin, et al. "Image-to-gps verification through a bottom-up pattern matching network." Asian Conference on Computer Vision. Springer, Cham, 2018.

[14] Oshikawa, Ray, Jing Qian, and William Yang Wang. "A survey on natural language processing for fake news detection." arXiv preprint arXiv:1811.00770 (2018).

[15] Bara, George & Backfried, Gerhard & Thomas-Aniola, Dorothea. (2019). Fake or Fact? Theoretical and Practical Aspects of Fake News. 10.1007/978-3-030-03643-0_9

[16] Moreira, Daniel, et al. "Image provenance analysis at scale." IEEE Transactions on Image Processing 27.12 (2018): 6109-6123.

[17] Bharati, Aparna, et al. "Beyond pixels: Image provenance analysis leveraging metadata." 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019.

[18] Kordopatis-Zilos, Giorgos, et al. "Finding Near-Duplicate Videos in Large-Scale Collections." Video Verification in the Fake News Era. Springer, Cham, 2019. 91-126.

[19] Bailer, Werner, Stefanie Wechtitsch, and Marcus Thaler. "Compressing visual descriptors of image sequences." International Conference on Multimedia Modeling. Springer, Cham, 2017.

[20] Lou, Yihang, et al. "Compact deep invariant descriptors for video retrieval." 2017 Data Compression Conference (DCC). IEEE, 2017.

[21] Mittal, Trisha, et al. "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues." Proceedings of the 28th ACM International Conference on Multimedia. 2020.Huh, Minyoung, et al. "Fighting fake news: Image splice detection via learned self-consistency." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[22] Mayer, Owen, and Matthew C. Stamm. "Exposing fake images with forensic similarity graphs." IEEE Journal of Selected Topics in Signal Processing 14.5 (2020): 1049-1064.

[23] Amato, Giuseppe, et al. "Face verification and recognition for digital forensics and information security." 2019 7th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2019.

[24] Cakir, Duygu, Merve Aritürk, and Özge Yücel. "An Overview of Fake Face Detection Approaches." International Online Conference on Intelligent Decision Science. Springer, Cham, 2020.

[25] Nirkin, Yuval, et al. "DeepFake detection based on the discrepancy between the face and its context." arXiv preprint arXiv:2008.12262 (2020).

[26] Giachanou, Anastasia, Guobiao Zhang, and Paolo Rosso. "Multimodal Fake News Detection with Textual, Visual and Semantic Information." International Conference on Text, Speech, and Dialogue. Springer, Cham, 2020.

[27] Wang, Youze, et al. "Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks." Proceedings of the 2020 International Conference on Multimedia Retrieval. 2020.

[28] Che, Tong, et al. "Deep verifier networks: Verification of deep discriminative models with deep generative models." arXiv preprint arXiv:1911.07421 (2019).