WEVERIFY: WIDER AND ENHANCED VERIFICATION FOR YOU PROJECT OVERVIEW AND TOOLS

Zlatina Marinova[1], *Jochen Spangenberg*[2], *Denis Teyssou*[3], *Symeon Papadopoulos*[4], *Nikos Sarris*[5], *Alexandre Alaphilippe*[6], *Kalina Bontcheva*[7]

(1) Sirma AI, trading as Ontotext, (2) Deutsche Welle, (3) Agence France-Presse, (4) CERTH-ITI, (5) ATC, (6) EU Disinfo Lab, (7) University of Sheffield

ABSTRACT

This paper presents an overview of the WeVerify H2020 EU project, which develops intelligent human-in-the-loop content verification and disinformation analysis methods, tools and services. Social media and web content are analysed and contextualised within the broader online ecosystem, in order to expose fabricated content, through cross-modal content verification, social network analysis, micro-targeted debunking, and a blockchain-based public database of known fakes.

1. INTRODUCTION

The past few years have highlighted the influential role of social networks and other digital media in shaping public debate on current affairs and political issues and perception of information. The rising influence of disinformation and the often so-called 'alternative media ecosystem' on societal debates, polarisation, and participatory democracy are of particular concern. Even blatant lies may get thousands of posts and shares, while the respective debunkings often receive comparatively little attention [1].

The process of finding, verifying, and reporting on a breaking news event increasingly involves monitoring and analysing large volumes of social media and online news content. In 2016 alone, the Duke Reporters' Lab[1] has established a staggering 50% increase in global fact-checking by media, press, journalists, and independent fact-checking organisations. This is making the news reporting process even more time consuming and costly. In addition to practical verification skills, journalists and media organisations are increasingly in need of collaborative verification tools, assistance through intelligent algorithms for automatic content verification, and the ability to cross-check quickly with peers and others whether a given claim or media item has already been proven false by other fact-checking organisations.

The urgent need to address all these major challenges and develop a new generation of content verification tools has also been recognised by the pan-European High Level Expert Group (HLEG) on Fake News and Online Disinformation [2], which emphasises the need to:

> *...undertake source-checking, establish content provenance, and forensically analyse images and videos at scale and speed, to counter disinformation (including when published by news media) and to document and publicize who produces and promotes it.*

This paper introduces the WeVerify tools and open platform, as well as related project activities. Their novelty lies in:

- Improving the breadth of capabilities for content verification, in particular towards social network analysis and deep fake detection;
- Scaling up and speeding up the verification process;
- Developing and making available a blockchain database of "known fakes";
- Employing a holistic, cross-modal verification workflow, supported by an open-source browser plugin and a user-friendly collaborative verification workbench.

2. RELATED WORK

State-of-the-art content verification tools and methods have largely focused on identifying manipulated or fabricated content, but algorithmic support for discovering "deep fakes" (also known as synthetic media) is still in its infancy. There is also a need for cross-modal contextual analysis approaches which combine metadata, social interactions, visual cues, user profiles, and other information surrounding a textual or multimedia item posted online, to assist a user who is checking on such items with its verification. With respect to online tools, the InVID plugin [3] and the Amnesty International "YouTube DataViewer" (citizenevidence.amnestyusa.org/) are the two tools frequently used by professionals. The latter offers YouTube metadata listing and image-based similarity search using keyframes. The former offers a larger toolset, including coverage of other platforms (Facebook & Twitter

[1] http://reporterslab.org/global-fact-checking-up-50-percent/

978-1-7281-1485-9/20/$31.00 ©2020

videos), verification-related comment detection, text-based location identification, and an advanced up-to-the-minute Twitter search for any time interval.

At the same time, tools for identifying sources of disinformation are mostly limited to spam bot detection, e.g. the Botometer tool (botometer.iuni.iu.edu/), which is predominantly based on social behaviour features (e.g. tweet frequency, hashtag use).

Existing and completed projects and tools mostly focus on media forensics and verification (e.g. InVID [3], REVEAL (revealproject.eu/) [4]), crowdsourced verification, e.g. CheckDesk (meedan.com/en/check/), Veri.ly (veri.ly), fact checking claims made by politicians (e.g. Politifact (www.politifact.com), FactCheck.org (www.factcheck.org), FullFact (fullfact.org/) citizen journalism (e.g. Citizen Desk), repositories of checked facts/rumours/websites (e.g. Emergent [5], FactCheck, Decodex (www.lemonde.fr/verification), or pre-trained machine learning models and tools, which however cannot be adapted easily by journalists to new data (e.g. PHEME [6], REVEAL [7]. There are also related tools for visualising and analysing online rumours: TwitterTrails [8], Hoaxy [9], CrossCheck (crosscheck.firstdraftnews.org/france-en/), Meedan's Check and ClaimBuster [10].

In contrast, our focus is on tools for multimodal content verification. Moreover, the use of Social Network Analysis (SNA) in journalist verification practices is currently under-explored, and yet much needed. With the role of social media becoming so dominant in spreading false content, journalists increasingly need to identify quickly the key sources, influencers, and propagation networks. Current verification tools, however, fall short of supporting such complex analyses, which can then be used also for more effective debunking.

3. WEVERIFY: A HIGH LEVEL OVERVIEW

The WeVerify project (weverify.eu) is developing a platform and a suite of content verification tools and algorithms covering the complete content verification workflow:

1. **Verification of content and source**: verification of textual claims, images, and videos; content provenance and source trustworthiness.
2. **Analysis of disinformation flows**: propagation analysis and community detection; early disinformation discovery on fringe platforms (e.g. 4chan, 8chan).
3. **Debunking of disinformation**: alert and warn users sharing, replying or liking fabricated content by providing them with evidence and additional context.
4. **Cataloguing and publishing**: a decentralised database of already debunked claims and tampered media.

This paper presents a number of already developed WeVerify tools that address the following steps of the verification workflow:

- Step 1: Verification of Content and Source: the veracity and stance analysis tool (Section 4);
- Step 2: Analysis of Disinformation Flows: the disinformation network analysis tool (Section 5).

We also present two multi-function, professional-oriented tools that bring together the above WeVerify tools alongside pre-existing and widely used verification technology, such as reverse image search:

- Extension and further development of an open-source content verification browser plugin, which is being used by journalists, fact-checkers, and human rights defenders to verify particular multi-modal content (images, text, video). See Section 6;
- The Truly Media collaborative verification workbench, which enables teams of journalists/fact-checkers to work collaboratively on the verification of a collection of social media and news content, circulating around a particular event. See Section 7.

Here and elsewhere in the paper by journalists we mean media professionals who adhere to and follow certain professional standards in media production and reporting.

The following chapters will canvass in more detail some of the work that is being undertaken in WeVerify.

4. VERACITY AND STANCE ANALYSIS OF ONLINE CONVERSATIONS

Online conversations (currently on Twitter) can be analysed and marked up for their veracity, with the help of an automatic, state-of-the-art rumour veracity classification algorithm [11]. It is a recurrent network which classifies the post initiating the conversation into three categories: true, false or unverified/uncertain. To aid with determining the veracity of the source post (tweet in this case), we use an algorithm that determines automatically the stance of each reply post, i.e. whether the reply agrees, disagrees, questions, or comments on the original post.

In order to convey the algorithm results in an easy to understand manner, we have built a web-based Graphical User Interface (GUI) front-end that can be used standalone or be integrated easily within verification toolboxes such as the browser plugin (see Section 6 and Section 7).

The process starts by the journalist entering a tweet URL; the tool then fetches its content, replies, and user profile information, as well as processes them with the algorithms in the backend. The background of the post that is being verified is coloured according to the judgement by the veracity analysis algorithms. Different levels of colour intensity convey the algorithm certainty in its predictions.

The automatic judgement is simply a suggestion, which the journalist can easily override manually, after they have examined the presented evidence. The manual judgement will be instantly stored in the database, allowing the classifier to be updated regularly by leveraging the newly annotated data.

The journalist can currently make two types of annotation. Firstly, annotations on the veracity of the rumour itself: whether it is true, false, or unverified, and journalists are encouraged to provide evidence for making this judgement. Secondly, they can annotate the stance of the responses in the thread. The stance of the response is one of support, deny, question or comment. When creating a dataset for re-training with user-provided annotations, each tweet, for both rumour veracity and stance classification, uses the class with the majority vote. Each tweet must also have 50% or more votes in the majority category to be used.

5. VISUAL EXPLORATION OF DISINFORMATION NETWORKS

We have developed a methodology and tools to support the sourcing and tracking of disinformation flows, based on Social Network Analysis (SNA). The current experiments have been based on implicit Twitter networks (e.g. retweets, mentions, replies). Next, we plan to generalise them to other social platforms and implicit networks, and offer support for tracking multimedia content, using both an actor-based network approach and a content-based network approach.

We have developed a web-based interface for visualising disinformation communities and information flows. It graphs a network centred around a user-selected account. Different colours of nodes and edges in the graph are used to represent different communities that have been identified automatically, using the Louvain community detection method [12]. If the user clicks on a given community, they can see a word cloud characterising this community, derived from the users' profile texts. The accounts most closely connected to a chosen account are also listed. It is possible to restrict the disinformation network to a specific time period and thus observe its change and evolution over time.

To enable better tracking of content that is spreading, we plan to integrate a near-duplicate retrieval method for images and videos [13] in the disinformation flow analysis by generating links between accounts that share near-duplicate content. Besides supporting analysis of disinformation flows for multimedia content, these functionalities can be used to support multi-modal content verification, by allowing analysis to be performed on clusters of near-duplicate posts instead of isolated items.

6. OPEN-SOURCE CONTENT VERIFICATION BROWSER PLUGIN

The content verification browser plugin is a new redesigned version of the InVID verification plugin [3], which has so far been downloaded and used by over 20,000 users. The browser extension is conceived as a verification "Swiss army knife" (implying it contains a versatile set of features and functionalities) helping journalists, fact-checkers, and human rights defenders to analyse and debunk disinformation.

The plugin encompasses tools for video analysis, video key frames extraction, investigation of YouTube thumbnail images for videos, user-friendly advanced Twitter search, image magnifier, Exif metadata viewer, and image forensics.

Work in progress on the plugin is a new capability to analyse propagation on social networks as well as memes and online adverts in order to extract the text from them automatically. This can then be sent to Google translate for example, to help journalists understand what is being said if they do not speak the language in question. It is also possible to index the image with keywords or the full meme/ad for later retrieval or search. The rumour and social network analysis tools are also planned for integration, after user testing and refinement.

7. TRULY MEDIA: COLLABORATIVE CROSS-MODAL VERIFICATION WORKBENCH

Truly Media is a collaborative content verification platform, which allows users to find, organise, and collaboratively verify content coming from social media or other digital sources. It addresses the complete verification and debunking workflow [14, 15]. The workflow includes the following steps:

1. **Find content:** a) Set up 'streams' of content from various social media sources, such as Twitter, Facebook, YouTube, or VKontakte. b) Create and refine these streams through a wide range of filters such as location, time, source, and language. c) Quickly browse through items, examining additional information.

2. **Organise content:** a) Create collections of content for specific investigations defining the team of experts who will have access to these collections for collaborative activities. b) Add content to these collections picking content from streams, social media browsers (Twitter, Facebook, YouTube), local files, or URLs of social media items or web articles. c) Exchange views on a collection through real time chatting and messaging. d) Easily browse through content applying different filters and add tags and notes to collection items.

3. **Verify content:** a) Preview quick analytics for the item source. b) Extract and visualise useful information with a set of powerful tools such as Google Maps, Wolfram Alpha, TinEye, Mapillary, and many more. c) Annotate items (including image and video) in a structured way keeping a record of every change in annotations. d) Collaborate via real-time chatting and messaging with the team.

Truly Media connects to TruthNest (www.truthnest.com) allowing users to run a great number of analytics on Twitter content in order to gather additional insights for any Twitter account.

8. BLOCKCHAIN DATABASE OF KNOWN FAKES

Increasingly, online disinformation contains older images and videos or already debunked claims and false narratives [16]. To automate the retrieval of such "known" fakes, we are creating a database of already debunked content, being populated both by the WeVerify verification tools and by debunks published by IFCN fact-checking organisations, and possibly others (depending on e.g. accessibility and rights matters).

The WeVerify database holds detailed information about already debunked content, which is represented using a slightly extended version of ClaimReview metadata schema (schema.org/ClaimReview). Importantly, the database itself does not store the content, just metadata describing the content: its type (article, image, video, etc.), location, identification (based on a hash value), and finally the claim/narrative that is being debunked.

Moreover, the database holds what we call "Verification Actions". These are veracity assessments made by journalists or verification professionals on a piece of content. These include classifying the content as false, misleading, unverifiable, etc., as well as additional information, e.g. supporting evidence (sources and reasoning used) and relevant context.

In order to ensure data integrity, the database is enhanced with a blockchain ledger. Each time a debunk is stored in the database, a new record is also created in the ledger: <agent key, content key, verification action key>. Then, when someone retrieves the content from the database, they also retrieve those values from the blockchain ledger to confirm the content is unmodified.

9. CONCLUSION AND FUTURE WORK

In conclusion, when it comes to understanding online disinformation and its impact on society, there are still many outstanding questions. The WeVerify project aims to address some of them. Most notable is studying the dynamics of the interaction between disinformation sources, amplifiers, and fact checks over time. This would help us quantify better what kinds of messages result in misinformation spreading accounts gaining followers and re-tweets, how human-like was the behaviour of the successful ones, and also whether any of these accounts are connected to the alternative media ecosystem and how. Another focus is synthetic media (aka "deep fakes"), their use in online disinformation campaigns, and the development of machine learning methods for detecting them.

Acknowledgments

This research is partially supported by the European Union under grant agreement No. 825297 WeVerify.

10. REFERENCES

[1] Soroush Vosoughi et al., "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] HLEG, "A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation. European Commission," 2018.

[3] Denis Teyssou et al., "The InVID plug-in: Web video verification on the browser," in *Proceedings of the First International Workshop on Multimedia Verification*, New York, NY, USA, 2017, MuVer '17, pp. 23–30, ACM.

[4] Markos Zampoglou et al., "Web and social media image forensics for news professionals," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[5] William Ferreira and Andreas Vlachos, "Emergent: a novel data-set for stance classification," in *Proc. of the 2016 conf. of the NAACL HLT*, 2016, pp. 1163–1168.

[6] Michal Lukasik et al., "Gaussian processes for rumour stance classification in social media," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 20:1–20:24, Feb. 2019.

[7] Christina Boididou et al., "Verifying information with multimedia content on twitter," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 15545–15571, 2018.

[8] Panagiotis Metaxas et al., "Using twittertrails.com to investigate rumor propagation," in *Proceedings of the 18th ACM Conference Companion on CSCW*, New York, NY, USA, 2015, CSCW'15 Companion, pp. 69–72, ACM.

[9] Chengcheng Shao et al., "Hoaxy: A platform for tracking online misinformation," in *Proc. of the 25th Int. Conf. WWW Companion*, 2016, pp. 745–750.

[10] Naeemul Hassan et al., "Claimbuster: the first-ever end-to-end fact-checking system," *Proceedings of the VLDB Endowment*, vol. 10, pp. 1945–1948, 08 2017.

[11] Ahmet Aker, Alfred Sliwa, Fahim Dalvi, and Kalina Bontcheva, "Rumour verification through recurring information and an inner-attention mechanism," *Online Social Networks and Media*, vol. 13, pp. 100045, 2019.

[12] Vincent Blondel et al., "Fast unfolding of communities in large networks," *Journal of stat. mechanics: theory and experiment*, vol. 2008, no. 10, pp. P10008, 2008.

[13] Giorgos Kordopatis-Zilos et al., "Visil: Fine-grained spatio-temporal video similarity learning," in *International Conference on Computer Vision (ICCV)*, 2019.

[14] John Cook and Stephan Lewandowsky, "The debunking handbook," 2011.

[15] Craig Silverman, "Verification handbook," 2015.

[16] Vasileios Mezaris et al., *Video Verification in the Fake News Era*, Springer, 2019.