WeVerify: Wider and Enhanced Verification for You Project Overview and Tools

Zlatina Marinova
Ontotext AD
Zlatina.Marinova@ontotext.com

Denis Teyssou
Agence France Presse
Denis.Teyssou@afp.com

Nikos Sarris
ATC
n.sarris@atc.gr

Jochen Spangerberg
Deutsche Welle
jochen.spangenberg@dw.com

Symeon Papadopoulos
CERTH
papadop@iti.gr

Alexandre Alaphilippe
EU DisInfo Lab
aa@disinfo.eu

Kalina Bontcheva
University of Sheffield
K.Bontcheva@dcs.shef.ac.uk

Abstract

This paper presents an overview of the WeVerify H2020 EU project, which is developing intelligent human-in-the-loop content verification and disinformation analysis methods. 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.

In particular, we introduce the following already developed tools: cross-modal content verification; blockchain-based database of known fakes; an open source content verification browser plugin; a collaborative verification workbench.

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. The rising influence of disinformation 1 and the often so-called alternative media ecosystem on societal debates, polarisation, and participatory democracy are of a particular concern. Even blatant lies get thousands of posts and shares, while the respective debunking often receives comparatively little attention (Vosoughi et al., 2018).

The process of finding, verifying, and reporting on a breaking news event in modern news production increasingly involves monitoring and analysing large volumes of social media and online news content (often of uncertain origin and trustworthiness). So-called alternative facts are continuously repeated online, even when proven untrue through fact-checking by mainstream media or independent experts. 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. All this is making the news reporting process even more time consuming and costly. In addition to practical verification skills and know-how, 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 (e.g. an image or video) has already been proven as false by other fact-checking or media 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. In particular, their recently published report (HLEG, 2018) 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. Their novelty lies in:

- Improving the breadth and quality of content verification, in particular towards social network analysis;
- Scaling up and speeding up the verification process;

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

- Developing a blockchain database of "known fakes";

- Employing a holistic, cross-modal verification workflow, supported by an open-source verification 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" is in its infancy. There is also a need for cross-modal contextual analysis approaches, which combine metadata, social interactions, visual cues, the user profiles, and all other information surrounding a textual or multimedia item posted online, to assist a user with its verification. With respect to online tools, the InVID plugin (Teyssou et al., 2017) and the Amnesty International "Youtube DataViewer"[2] are the two typically used by professionals. The latter offers YouTube metadata listing and image-based similarity search using keyframes. The former offers a much fuller toolset, including coverage of other platforms (Facebook & Twitter videos), verification-related comment detection, weather analysis, text-based location identification, and Twitter search for identifying reposts of a video.

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

In more detail, existing projects and tools are mostly focused on images/video forensics and verification (e.g. InVID (Teyssou et al., 2017), REVEAL[3]), crowdsourced verification (e.g. CheckDesk[4], Veri.ly[5]), fact-checking claims made by politicians (e.g. Politifact[6], FactCheck.org[7], FullFact[8]), citizen journalism (e.g. Citizen Desk), repositories of checked facts/rumours/websites (e.g. Emergent (Ferreira and Vlachos, 2016), FactCheck [7] , Decodex[9]), or pre-trained machine learning models and tools, which however cannot be adapted easily by journalists to new data (e.g. PHEME (Lukasik et al., 2019), REVEAL[3]).

There are also existing tools for visualising and analysing online rumours which are related to the user interface of our system:

- RumorLens (Resnick et al., 2014) is a prototype aimed at citizens and journalists, to help detect rumours early, then classify posts as spreading or correcting the given rumour, and also visualising its spread. A human-in-the-loop learning showed good results on the tweet retrieval task. This motivated us to propose extending this approach to other rumour and misinformation analysis tasks.

- TwitterTrails (Metaxas et al., 2015) is an interactive, web-based tool that visualises the origin and propagation characteristics of a rumour and its refutation, on Twitter. Another visualisation-based framework for studying rumour propagation is RumourFlow (Dang et al., 2016).

- Hoaxy (Shao et al., 2016) is a recent open-source tool focused on visualising and searching over claims and fact checks. Such sophisticated visualisations are out of scope of our system, but relevant open-source visualisation tools, e.g. from Hoaxy, could be integrated in the future.

- CrossCheck[10] was a collaborative rumour checking project led by First Draft and Google News Lab, during the French elections. Its output was a useful dataset of false or unverified rumours.

- Meedan's Check[4] is an open-source breaking news verification platform, which however does not support continuously updated machine learning methods.

- ClaimBuster (Hassan et al., 2017) is a tool which gathered volunteer and expert-based claim annotations to train machine learning methods for claim classification (factual vs non-factual). In contrast, our focus is on tools for multimodal content verification.

[2] https://citizenevidence.amnestyusa.org/
[3] https://revealproject.eu/
[4] https://meedan.com/en/check/
[5] https://veri.ly
[6] https://www.politifact.com/
[7] https://www.factcheck.org/
[8] https://fullfact.org/
[9] https://www.lemonde.fr/verification/
[10] https://crosscheck.firstdraftnews.org/france-en/

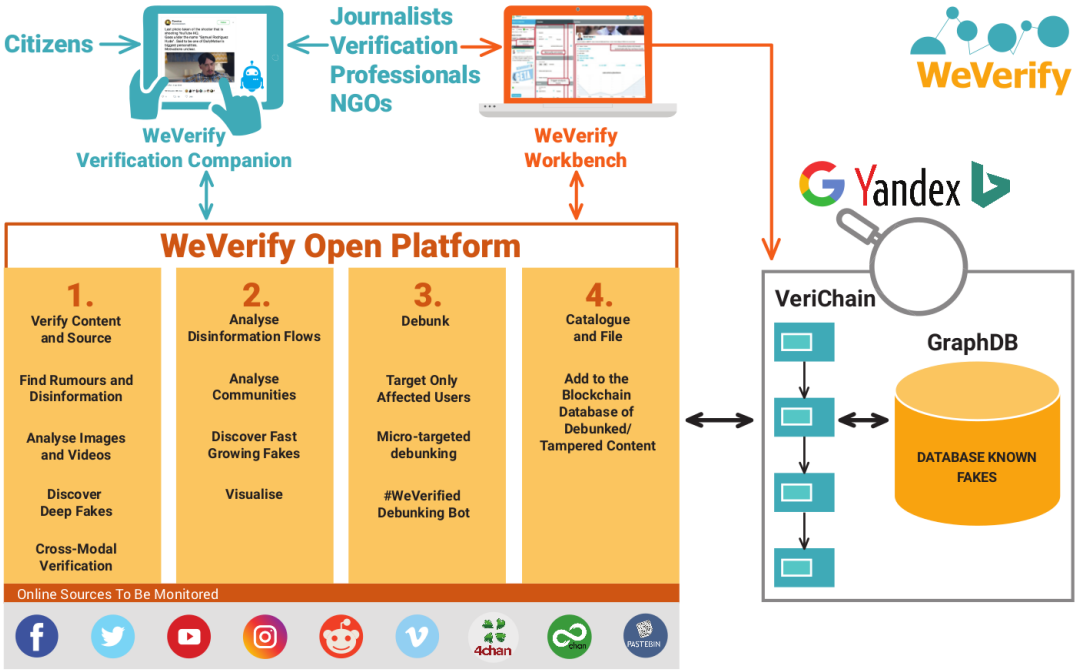

Figure 1: The WeVerify Verification Workflow

Moreover, the use of Social Network Analysis (SNA) in journalist verification practices is currently underexplored, and yet much needed (newsrooms mostly have access to top trends only). 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 is developing a platform and a suite of content verification tools and algorithms covering the complete content verification workflow (see Figure 1):

1. **Verification of content and source**: verification of textual claims, images, and videos (incl. AI-generated fakes); cross-modal content verification; 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**: creation of a decentralised database of already debunked claims and tampered images and videos, accessible both programmatically (e.g. by search engines or social platforms) and via a user-friendly web interface.

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:

- An open-source content verification browser plugin, which is being used by individual journalists to verify particular multi-modal content (images, text, video). See Section 6;

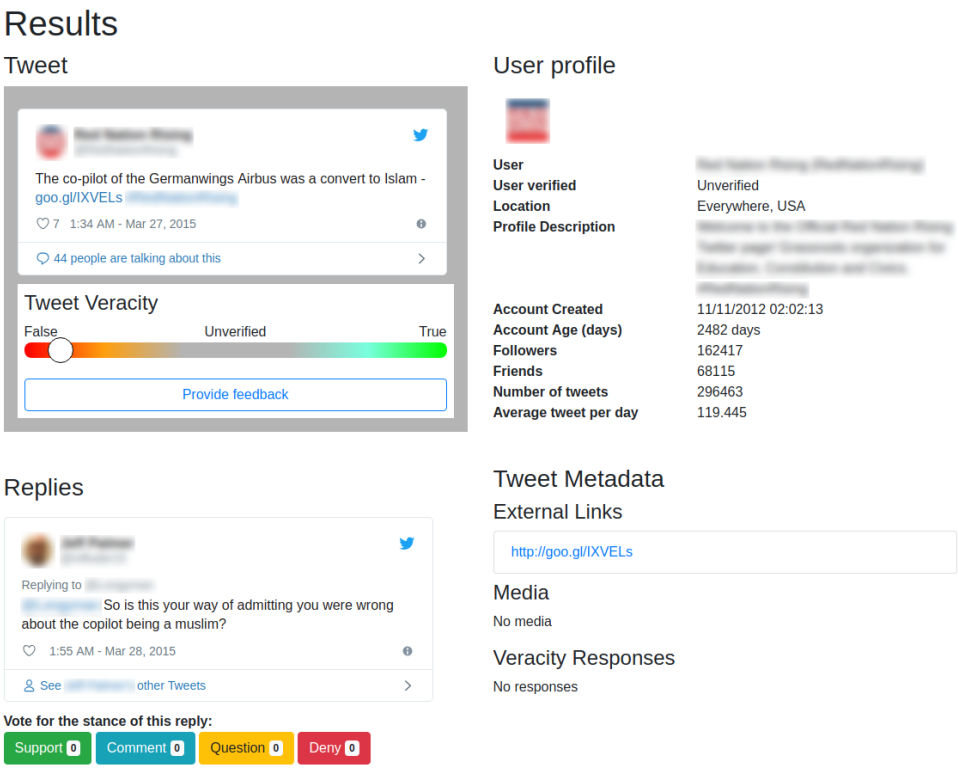

Figure 2: A screenshot of the web-based UI for veracity and stance analysis. The source post (tweet in this case) is shown on the top left. The automatic veracity classification is displayed below the post on a single axis scale that ranges between False (red), Unverified/Uncertain (grey) and True (green). Metadata about the post is shown on the right. Replies with their stance are shown on the bottom left.

- The TrulyMedia collaborative verification workbench, which enables teams of journalists/fact-checkers to work collaborative on the verification of a collection of social media and news content, circulating around a particular event. See Section 7.

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 Aker et al. (2019). It is a recurrent network which classifies the post originating 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 (Fig. 2) that can be used standalone or be integrated easily within verification toolboxes such as the browser plugin (see Section 6 and TrulyMedia 7. The process starts by the journalist entering a tweet URL and the tool then fetches its content, replies, and user profile information, as well as processing them with the algorithms in the backend.

As can be seen in Figure 2, the results are then displayed to the user. The background of the originating post that's being verified is coloured according to the judgement made by the veracity analysis algorithms. The different levels of colour intensity are used to convey how certain is the algorithm in its predictions.

The automatic judgement is simply a suggestion, which the journalist can easily override through manual input, after they have examined the presented evidence. The manual judgement will be instantly stored into the database, allowing the classifier to be updated regularly by leveraging the newly annotated data, when a sufficient amount of it becomes available.

The journalist can currently make two types of

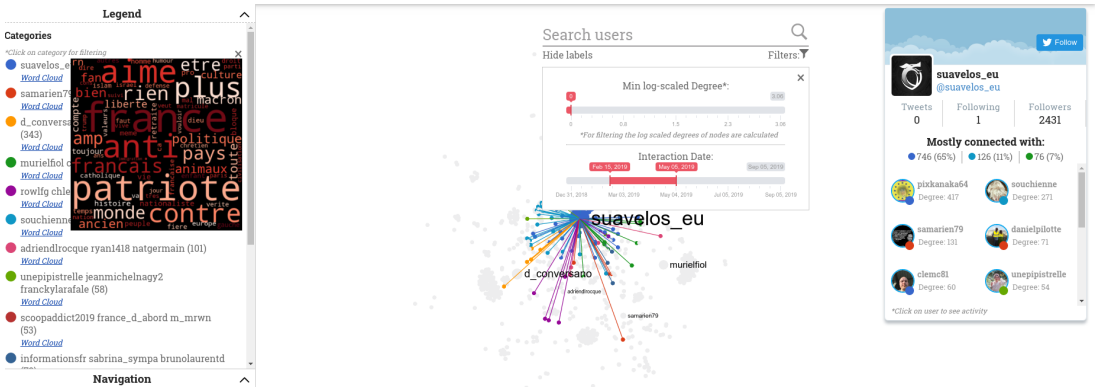

Figure 3: UI for Analysing Disinformation Networks

annotations. Firstly, annotations on the veracity of the rumour itself. Whether it is true, false, or unverified, and 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 misinformation flows, based on Social Network Analysis (SNA). The current experiments are based on implicit Twitter networks (e.g. retweets, mentions, replies), but in next steps we plan to generalise them to other social platforms and implicit networks, as well as enable support for tracking multimedia content, bringing both an actor-based network approach and a content-based network approach.

Unlike other tools for social media analysis, we do not simply crawl through user accounts according to keywords or geographical location, but also harness implicit networks, as well as additional information such as topics and sentiment.

We have developed a web-based interface for visualising disinformation communities and information flows (Figure 5, which we plan to integrate within the content verification browser plugin and the TrulyMedia collaborative verification workbench. In more detail, Figure 5 shows the specific implicit, centred around a user-selected account (suavelos_eu). The different colours of nodes and edges reflect the different communities that were identified automatically, using social network analysis. 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 shown on the right. It is also possible to restrict the disinformation network to a specific time period and thus observe its change and evolution over time.

In order to enable better tracking of content spread, we have developed a near-duplicate retrieval method for images and tracking them for the purpose of disinformation flow analysis. Besides supporting analysis of disinformation flows for multimedia content, these functionalities can be used to support multimodal 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 (Figure 4 is a new redesigned version of the InVID verification plugin (Teyssou et al., 2017), which has so far been downloaded and used by over 13,600 users. The browser extension is conceived as a verification "Swiss army knife" helping journalists, fact-checkers and human rights defenders to debunk disinformation.

As can be seen from Figure 4, it encompasses tools for video analysis (includig Twitter and Facebook videos), key frame extraction from video, investigation of the YouTube published thumbnail images for videos, user-friendly advanced Twitter search, image magnifier, EXIF

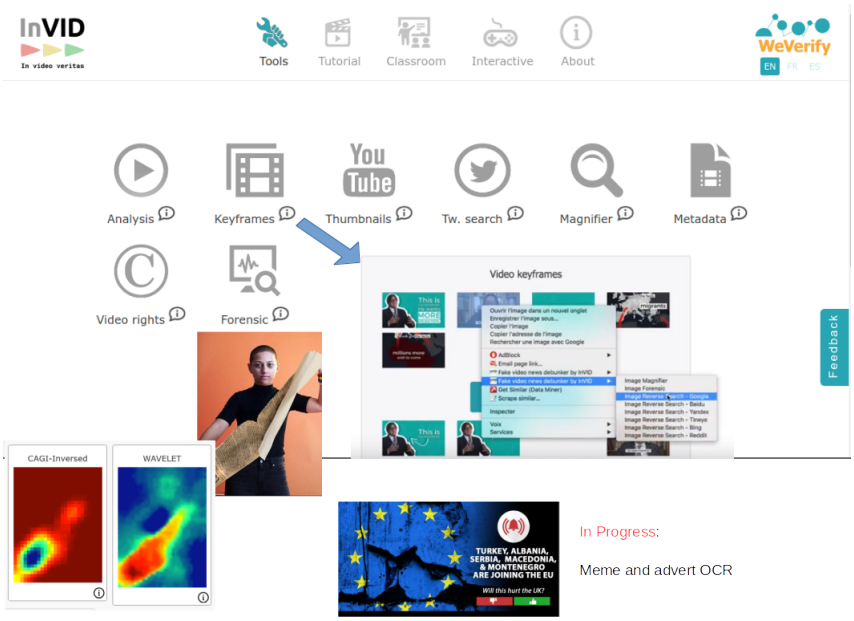

Figure 4: Screenshots of the Browser Plugin

metadata viewer, and AI-based image forensics. In particular, Figure 4 shows two example analyses. Firstly, it shows how a propaganda video of Guy Verhofstadt is analysed by first extracting automatically the key frames, followed by a reverse image search (which can be carried out against Google, TinEye, Yandex, and several other engines. Figure 4 also shows how a tampered image of Emma Gonzalez (allegedly showing her tearing up the US constitution) is analysed with the automatic image forensics tools which show the suspect areas of the image in red.

Work in progress on the plugin is a new capability to analyse memes and online adverts and extract the text from them automatically. This can then be sent to Google translate for example, to help journalists understand what's being said if they do not speak the language. It is also possible to index the image with keywords or the full of the meme/ad for later retrieval or search.

The rumour and social network analysis tools are also planned for integration, after currently undergoing user testing and refinement.

7 TrulyMedia: Collaborative Cross-Modal Verification Workbench

Truly Media (www.truly.media) is a collaborative content verification platform (Figure 5) which allows users to find, organise and collaboratively verify content coming from social media or web sources. It addresses the complete verification and debunking workflow (Cook and Lewandowsky, 2011; Silverman, 2015) and is in the process of integrating the machine learning methods for cross-modal content verification (Section 4) and the social network analysis methods for sourcing and tracking disinformation flows (Section 5).

More specifically, Truly Media allows users to:

1. **Find Content**
   (a) Set up 'streams' of content coming from various Social Media sources, such as Twitter, Facebook, YouTube, or VKontakte.
   (b) Create and refine streams through a wide range of filters such as location, time, source, and language.
   (c) Quickly browse through items, examining additional information, as user details, or translated text in the full conversation thread.
2. **Organise Content** - see Figure 5
   (a) Create collections of content for specific investigations defining the team of experts who will have access and will collaborate to gather relevant content and verify it.
   (b) Add content to the collection through: 'drag and drop' content from streams; pick content from ordinary Social Media

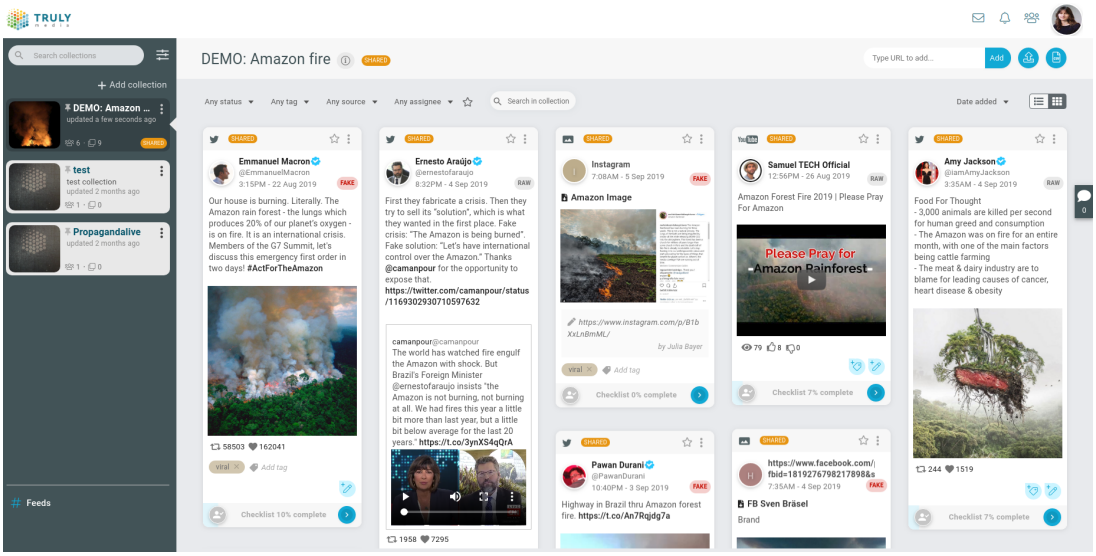

Figure 5: The TrulyMedia Collaborative Verification Workbench: Screenshot of the top level dashboard

browsers (Twitter, Facebook, YouTube); upload files from local folders; provide the link of any social media item or web article.

(c) Exchange views on a collection through real time chatting and messaging.

(d) Easily browse through content applying different filters

3. **Verify Content** - following a checklist methodology (see Figure 6:

(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, Pipl and many more.

(c) Annotate item in a structured way keeping a record of every change in annotations.

(d) Collaborate via real-time chatting and messaging with the team.

Adding to the above, 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 a specific account. More specifically, the user can:

1. Browse through deep analytics on the activity, network, or influence of any source on Twitter. The analytics are based on AI and produce a set of alerts or flags which highlight suspicious (bot-like) behaviour.

2. Gain insights about a Twitter account by accessing a wide variety of metrics, most of them not visible by mere checking of a users Twitter account.

3. Assess whether the Twitter account exhibits bot-like behaviour.

While currently TrulyMedia is focused primarily on verifying Twitter, Facebook, and YouTube content, support for Reddit, and 4chan is currently being added.

8 Blockchain Database of Known Fakes: Work in Progress

Increasingly, online disinformation contains older images and videos or already debunked claims/ false narratives. In order to automate the retrieval of such already "known" fakes, WeVerify is creating a database of already debunked content. This database is being populated not only by the verification tools described here, but is also being expanded with debunks published by some IFCN member fact-checking organisations.

The blockchain database (see the right-hand side of Figure 1 has two complementary aspects: the database of debunked content and the blockchain, which we refer to as the VeriChain.

The WeVerify database holds detailed information about already debunked content, which is being represented using a slightly extended ClaimReview[11] metadata schema. In order to avoid a

[11] https://schema.org/ClaimReview

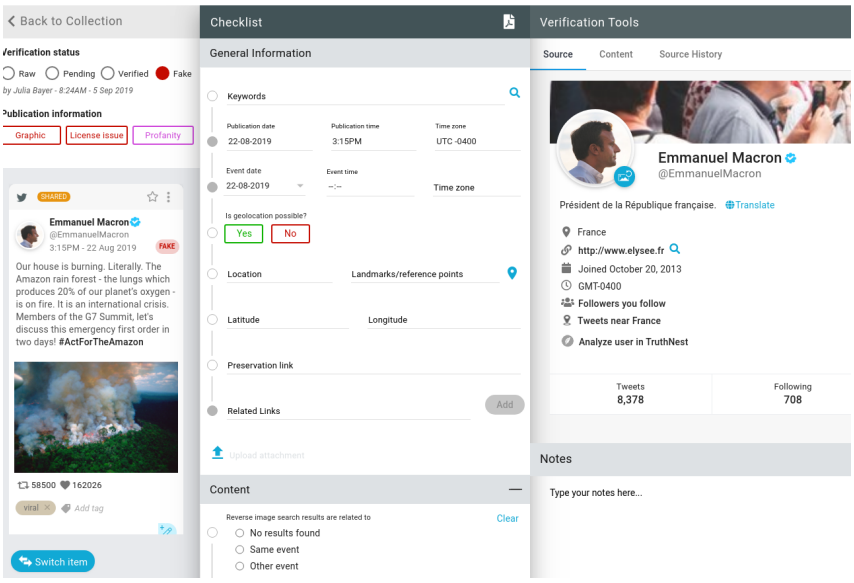

Figure 6: The TrulyMedia Collaborative Verification Workbench: Verification Checklist

single point of failure, the database is kept in parallel at several locations. Importantly, the database itself does not store the content, but only stores metadata about the content, encoded using the extended ClaimReview-based schema.

Firstly, there is metadata describing the content – what type it is (article, image, video, etc.), where it can be found, how it can be identified (based on a hash value), and finally the claim/narrative that is being debunked. It is possible to extend the schema in future.

Secondly, the database holds Verification Actions. One Verification Action is a judgement of veracity made by a journalist or a verification professional on a piece of content. The verification includes classifying the content as false/misleading/unverifiable/etc., but it also includes additional information, e.g. supporting evidence (sources and reasoning used) and relevant context (e.g. the claim is not true now but was true in the past or might be true in the future). All this data is being made available programmatically via a SPARQL-based query interface.

The blockchain, while not the appropriate tool for holding large quantities of data, is the ideal tool for verifying the consistency of other data sources, which is how it is used in this case. Each time a professional verifies a piece of content (creating a new Verification Action in the central database), a new record is also created in the VeriChain, writing an agent key-content key-verification action key triple. Then when someone retrieves them from the database, they also retrieve those values from the blockchain and confirm the contents are unmodified.

9 Conclusion and Future Work

In conclusion, when it comes to understanding online misinformation and its impact on society, there are still many outstanding questions. The WeVerify project aims to address some of them in the remaining two years of the project. Most notable is studying the dynamics of the interaction between disinformation sources, amplifiers, and fact checks over time. This would help us quantify better (amongst other things) 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 were any of these accounts connected to the alternative media ecosystem and how.

Another key focus is on studying synthetic media (ako “deep fakes”), their use in online disinformation campaigns, and development of machine learning methods for analysing and recognising synthetic media.

Acknowledgments

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

References

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

John Cook and Stephan Lewandowsky. 2011. The debunking handbook.

Anh Dang, Abidalrahman Moh'd, Evangelos Milios, and Rosane Minghim. 2016. What is in a rumour: Combined visual analysis of rumour flow and user activity. In *Proceedings of the 33rd Computer Graphics International*, CGI '16, pages 17–20, New York, NY, USA. ACM.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Naeemul Hassan, Anil Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10:1945–1948.

HLEG. 2018. A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation. European Commission. https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.

Michal Lukasik, Kalina Bontcheva, Trevor Cohn, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2019. Gaussian processes for rumour stance classification in social media. *ACM Trans. Inf. Syst.*, 37(2):20:1–20:24.

Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. 2015. Using twittertrails.com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, CSCW'15 Companion, pages 69–72, New York, NY, USA. ACM.

Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*, pages 10121–0701.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 745–750, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Craig Silverman. 2015. Verification handbook.

Denis Teyssou, Jean-Michel Leung, Evlampios Apostolidis, Konstantinos Apostolidis, Symeon Papadopoulos, Markos Zampoglou, Olga Papadopoulou, and Vasileios Mezaris. 2017. The invid plug-in: Web video verification on the browser. In *Proceedings of the First International Workshop on Multimedia Verification*, MuVer '17, pages 23–30, New York, NY, USA. ACM.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.