PROVENANCE: An Intermediary-Free Solution for Digital Content Verification

Bilal Yousuf[1,2], M. Atif Qureshi[1,2], Brendan Spillane[1], Gary Munnelly[1], Oisin Carroll[1], Matthew Runswick[1], Kirsty Park[3], Eileen Culloty[3], Owen Conlan[1] and Jane Suiter[3]

[1]*ADAPT Centre, Trinity College Dublin*

[2]*ADAPT Centre, Technological University Dublin*

[3]*Institute for Future Media, Democracy and Society, Dublin City University*

Abstract

The threat posed by misinformation and disinformation is one of the defining challenges of the 21st century. Provenance is designed to help combat this threat by warning users when the content they are looking at may be misinformation or disinformation. It is also designed to improve media literacy among its users and ultimately reduce susceptibility to the threat among vulnerable groups within society. The Provenance browser plugin checks the content that users see on the Internet and social media and provides warnings in their browser or social media feed. Unlike similar plugins, which require human experts to provide evaluations and can only provide simple binary warnings, Provenance's state of the art technology does not require human input and it analyses seven aspects of the content users see and provides warnings where necessary.

Keywords

Misinformation, Disinformation, Fake News, Social Media, Plugin, Browser Extension

1. Introduction

Provenance is an intermediary-free solution for digital content verification to combat misinformation and disinformation on the Internet and social media. As per [1], it is designed to aid users by providing them with warning notifications in their browser or social media feed when viewing content that may be dangerous or problematic. The detailed warning notifications inform users which of the seven criteria Provenance's state of the art technology has detected an issue with and why. It significantly improves upon all known similar solutions in two ways. Firstly, existing solutions do not analyse the content the user is viewing and are thus limited to providing users with warnings based on the news agencies historical publication record and behaviour. Secondly, existing browser

plugins only provide a single broad-spectrum warning about the content users are viewing whereas Provenance is capable of evaluating content under seven criteria and providing individual warnings for each. Provenance's warning notifications are also educational and designed to inspire users to be more cautious and critical of the information they consume. Thus, it will improve media literacy among users and make them less susceptible to the influence of misinformation and disinformation by making them more critical and reflective of the content they consume.

There are significant research challenges in the design and development of Provenance. The main challenges include the huge volume of news and other content published each day, the combination of multimedia formats in each article or story, the high churn-rate and short shelf-life of news, and the fact that news content is often republished from wire services or from other publishers. These are compounded by the fact that misinformation and disinformation are often designed to masquerade as real news. Many disinformation sources share characteristics with the Lernaean Hydra of Greek mythology and re-post problematic content through multiple easy to set up websites or social media groups and reappear under different guises when they are identified and shut down.

There are also a range of individual challenges within components of the Provenance platform. These include deriving a system to assign accurate writing quality scores for each piece of textual content, detecting when new facts introduced in a news article are indicative of disinformation or an evolution in an unfolding story, detecting image and video manipulations, or developing

a system that can differentiate between anger and fear in disinformation and anger and fear in opinion news articles. There is also some difficulty in differentiating between news articles from alternative and independent agencies and news articles from disinformation sources due to often lower quality writing, more emotive content, and the reuse of images and videos.

This paper provides an update on the ongoing progress of developing Provenance. The remainder of this paper is organised as follows. Section 2 *Motivation and Background* delves into the impetus for this project and situates it within other recent EU disinformation projects. Section 3 *Related Work* provides a detailed overview of similar browser plugins and describes how Provenance advances the state of the art. Section 4 *Architecture Overview* contains system architecture diagrams and descriptions of each component in the Provenance platform. Section 5 *Provenance in Action* provides a detailed explanation of how the Provenance browser plugin provides warnings to the user. Section 6 *Use Cases* presents two use cases for the Provenance plugin to show in what scenarios we envision it being used. Section 7 *Evaluation* briefly describes plans to evaluate the tool. Finally, section 8 *Conclusions* completes the paper with closing remarks.

2. Motivation and Background

The proliferation of misinformation and disinformation on social media has been described as a strategic threat to democracy and society in the European Union (EU) [2, 3]. A recent EU study on the issue found that the common narratives of society *"are being splintered by filter bubbles, and further ruined by micro-targeting."* [4]. The report points out that like a virus, misinformation and disinformation spread throughout society through social media and other platforms in open and closed groups to the detriment of democratic systems. This occurs when *"Susceptible users become weaponized as instruments for disseminating disinformation and propaganda"* [4].

The Presidents of the European Council, Commission and Parliament have all made increasingly public calls for concerted efforts to do more to combat the scourge of fake news to protect democracy. The President of the European Parliament has been the most forthright in this with a recent announcement that: *"We must nurture our democracy & defend our institutions against the corrosive power of hate speech, disinformation, fake news & incitement to violence."* [5]. As a result, the EU have funded a range of FP7, H2020 and other projects to combat misinformation and disinformation including WeVerify [6, 7], SocialTruth [8], PHEME [9, 10], EUNOMIA [11] Fandango [12, 13] and the European Digital Media Observatory (EDMO) [14]. Many other international

organisations have also identified misinformation and disinformation as a threat and have increased efforts to combat it. These include the United Nations through its Verified platform [15] and the World Health Organisation [16]. More can be read about these initiatives in the Poynter Institute's guide to national and international efforts to combat misinformation and disinformation around the world [17].

Provenance is a H2020 project[1], however it differs from many of the above as it is a user orientated intermediary-free solution to help consumers identify misinformation and disinformation as they browse the Internet and social media. It is also designed to improve media literacy skills by equipping consumers with the tools, knowledge and know-how to face this challenge now and into the future.

3. Related Work

This review of related work will focus on comparable browser plugins designed to provide users with warning notifications about disinformation or other problematic content and which are currently active or maintained. The purpose of this review is to establish how Provenance advances the state of the art.

NewsGuard [18] provides 'nutrition' labels for news websites based on nine journalistic criteria. What differentiates it from many of the other fake news and bias detection browser plugins is that it does not use automated algorithms to assess news websites but rather relies on a team of journalists to conduct reviews. It comes as standard with Microsoft Edge, but a subscription is needed for other Internet browsers. Its notification icons appear as a browser extension in the upper right corner and within third party search engines and social media platforms. Clicking on its browser icon opens a nutrition label pane where users can quickly see whether the news website passes or fails any of the nine criteria. A link is also available for users to see a more detailed report. Visually, NewsGuard employs simple but effective white ✓ on a green shield and red **x** iconography to denote when a website has passed or failed. NewsGuard's transparent methodology has resulted in their datasets being used for research [19]. While expert led analysis has its merits, it also has issues with scalability, personal biases, and response times. Aker also maintains that much of the credibility and transparency scoring provided by NewsGuard could be automated [20].

Décodex [21] created by Le Monde originally started as an online search facility for users to check URLs against a list of known websites which spread misinformation and disinformation. They have since released a Facebook bot for users to directly chat to and a browser plugin that provides red, orange or blue notifications to

denote whether a website regularly disseminates false information, whose reliability is doubtful, or if they are a parody website. When installed, the Décodex icon becomes active when the website being viewed is listed in their database. It also produces a colour-coded popup with one of three standard warnings. Users cannot access detailed information about warnings, nor does it appear to be integrated with well-known search engines, social media platforms or discussion boards. Décodex's allow/deny list approach means that scalability is difficult and the warnings it provides are based on the historical publication record of the website, not the content currently being viewed. Transparency is also limited. While still available, its development appears to be in stasis.

Media Bias Fact Check (MBFC)[2] [22] is an extensive media bias resource curated by a small team of journalists and lay researchers who have undertaken detailed assessments of over 4000 media outlets. A transparent assessment methodology means that their datasets have been used for several research projects [23, 20]. Their team of researchers undertake in-depth analyses of news organisations and assess them using a standardised methodology, with some subjective judgement, to calculate a left/right bias score using their published formula. They also calculate scores for factual reporting and credibility. These reports are published on their website and updated from time to time. Each news website in their database is categorised as: left bias, left-centre bias, least biased, right-centre bias, right bias, pro-science, conspiracy-pseudoscience, fake news, or satire. While their browser extension conveys limited details, further information about each news source is available on their website. It draws on this dataset to inform users when they click on the notification icon as to which of these nine categories the news website they are viewing belongs to, including a brief explanation of the category. It also provides a link to the detailed MBFC report. The browser extension also provides Facebook and Twitter support by displaying a visual left/right bias scale on news articles that appear in users feeds with links to the MBFC detailed report and Factual Search[3] so that the user can investigate the topic further. While a valuable resource with considerable detail, MBFC's expert evaluations are based on the historical publication record of the news website and not an evaluation of the content the user is looking at. It is also a labour intensive and time consuming process.

Stopaganda Plus[4][24] is a browser extension that adds accuracy and bias decals to Facebook, Twitter, Reddit, DuckDuckGo and Google. These visual indicators extend the functionality of MBFC (who determine the scores) to these common information portals so that users may more easily choose high-quality information resources. It should be noted that this extension is not designed to provide users with detailed warning notifications when viewing a news website and thus is not directly comparable to the other systems or Provenance. It is included here due to its use of MBFC, the fact that it conveys limited visual information/warnings before the user visits an information source, and for plenitude.

3.1. No Longer Active

Many other projects and services related to this work, which have been reviewed in the literature, c.f. [25, 26, 27, 11, 28, 29, 30], now no longer appear to be active or working. This is concerning as despite the fact that misinformation and disinformation have been recognised as a threat to democracy and social cohesion, and the fact that browser plugins are one of the few citizen-orientated direct interventions which can help solve the problem at source while increasing long term media literacy, very few of the proposed solutions have been actively promoted or maintained. The main reason for this appears to be the fact that many of these plugins were developed by individuals or small teams, or even as part of a hackathon, and were thus lacked the resources to be actively maintained or updated to deal with changing technology such as browser updates or the rapidly evolving threats posed by misinformation and disinformation. The following present those related projects found in the literature, but which now no longer appear to be actively maintained, though some are still available to install. URLs have been included for posterity where possible as many do not have peer-reviewed publications.

B.S Detector[5] relied on matching the URLs of content in the news feed to a known allow/deny list of sources of fake news and misinformation.

AreYouFakeNews.com[6] utilised Natural Language Processing (NLP) and deep learning to identify patterns of bias on websites.

Fake News Detector AI[7] claimed to use a neural network to detect similarity between submitted URLs and known fake news websites.

Fake News Detector[8] was designed to learn from webpages flagged by users to detect other similar fake news webpages.

Trusted News[9] is a browser plugin that was designed to assess the objectivity of news articles. Its functionality was limited to 'long form' news articles and it does not work with social media content.

Fake News Guard[10] claimed to combine linguistic and network analysis techniques to identify fake news, however this can no longer be verified.

FiB[11] A browser extension built in a hackathon which was reviewed several times in the literature as a comparable system [31].

TrustedNews[12] Trusted News used AI to help users evaluate news articles by scoring their objectivity [32]. However, it does not work on social media and has issues with analysing webpages that require scrolling.

Trusty Tweet [26] was designed to help users deal with fake news tweets and to increase media literacy. Their transparent approach is designed to prevent reactance and increase trust. Early user evaluations showed promise.

Check-It [33] was designed to analyse a range of signals to identify fake news. It was focused on user privacy with computation undertaken locally. Their approach used a combination of linguistic models, fact checking, and website and social media user allow/deny lists.

3.2. Out of Scope Approaches

Some misinformation and disinformation detection tools which have been reviewed in other papers have not been included in this literature review. This is because they are not a browser plugin or they are a paid for b2b service (Fakebox [34]; AreYouFakeNews [35]), they are focused on an aligned but separate issue e.g., detection of bias or detection of reused and or manipulated images (Ground.News [36]; SurfSafe [37]), they are specifically for fact checking (BRENDA [38], CredEye [39]), they have pivoted into a B2B platform (FightHoax [40]), they are not user orientated (Credible News [41, 42]), or they are research systems and have not been made available to the public [30, 43]. While relevant to combating disinformation, these are not directly comparable to Provenance.

3.3. Advancing the State of the Art

This review demonstrates that browser plugins are a common user-orientated approach to combat misinformation and disinformation. However, Provenance adopts a significantly more advanced and granular methodology than current or previous efforts in the domain. The warnings provided by earlier plugins are often based on the news website's history of publishing misinformation and disinformation. Thus, they are limited to providing a coarse-grained retrospective analysis of the news website's publication history. In contrast, Provenance's fine-grained approach is designed to analyse the content of the news webpage or users' social media feeds and,

where necessary, provide an easy to understand warning to the user when the content they are viewing may be problematic or symptomatic of disinformation. In the cases where linguistic analysis or other machine learning approaches have been utilized, the results are not presented to the user in an explainable or transparent way. Some of these methods have also proven susceptible to adversarial attacks, whereby text may be augmented slightly to fool pretrained models [44, 45].

Two factors differentiating Provenance from the plugins described above are their limited reach and scalability. Many of the above plugins do not provide any information for some heavily trafficked news websites such as the LA Times, Al Jazeera, and the Independent.co.uk. This is likely due to limiting factors of time and labour of including humans in the disinformation judgement process. While no one doubts the benefits of highly trained expert judgement, the size and nature of the rapidly evolving media landscape, especially in regard to misinformation and disinformation in which publishers are prone to rapid growth, failure and re-branding, means that providing human ratings is a never ending game of whack-a-mole. Current solutions are only partially succeeding in providing judgements of some news agencies. None have attempted to analyse the millions of pieces of content they publish daily. Unlike each of the plugins described above, Provenance does not require a human-in-the-loop, nor does it need to be backed by human-generated allow/deny lists. Its architecture supports fully automated and intermediary free analysis of news content.

The ability to evaluate news articles against seven criteria and provide users with visual notifications and deeper explanations is also a significant advancement on the state of the art and a direct benefit to users in three ways. First, and most importantly, users will be made aware of individual issues with the content they are consuming and can thus decide whether they will continue viewing it or look for alternative sources. Second, it will help develop users' media literacy skills by making them aware of the different caution worthy indicators and how to check them, making them less susceptible to misinformation and disinformation in the future. Third, the nature of these systems means that they cannot be properly examined. In contrast, a full description of Provenance's system architecture is provided below. It is also currently undergoing evaluation and testing and the results will be published in time.

4. Architecture Overview

The system architecture for Provenance is shown in Figure 1. The components and services use REST APIs serving JSON for easy, reliable, and fast data exchanges across internal subsystems.

---

[10]http://fakenewsguard.com/

[11]https://projectfib.azurewebsites.net/
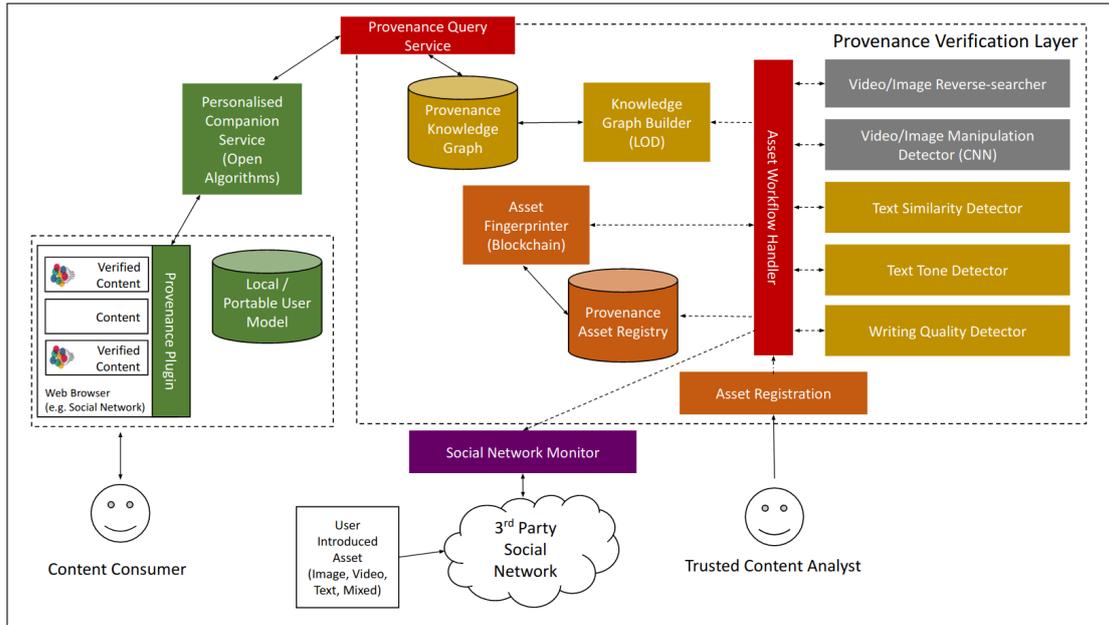
[12]https://trusted-news.com/

Figure 1: Provenance System Architecture: Dashed lines denote REST API calls, solid lines denote local access.

Data in the form of webpages or social media content is ingested by Provenance either through the *Social Network Monitor* or by a *Trusted Content Analyst* (e.g., a journalist or fact checker). The *Social Network Monitor* service discovers content using NewsWhip's[13] social network monitoring platform. The introduced asset is enriched with social engagement data (e.g., likes and shares) and is forwarded to the *Asset Workflow Handler* service.

The *Asset Workflow Handler* separates the incoming data (e.g., a news webpage) into individual assets such as images, video, text, etc. These assets are registered with the *Asset Fingerprinter* before being disseminated to the analytical components (*Video/Image Reverse-searcher*, *Video/Image Manipulation Detector*, *Text Similarity Detector*, *Text Tone Detector*, and *Writing Quality Detector*) to determine if they exhibit any features which normally characterise misleading, questionable, or unsubstantiated information. The output of each analytical service, and the initial data passed from the *Social Network Monitor* are combined and sent to the *Knowledge Graph* where they are stored.

The Knowledge Graph may be queried by the *Provenance Query Service* to retrieve the results of analysis for a given webpage. The Provenance plugin, installed in the user's browser, leverages this query service to retrieve information about webpages that a user is currently viewing. If the webpage has been analysed by Provenance, and exhibits questionable features, the plugin will issue a warning to the user, indicating that they may want

to further investigate the claims made in the article's content. The *Personalised Companion Service* is used to determine how this information should be presented for an individual user.

4.1. Key Components

4.1.1. Social Network Monitor

The *Social Network Monitor* communicates with NewsWhip's *Social Network API* to identify assets which should be ingested by Provenance. Finding assets involves querying Newswhip's API with a parameterized search request. The call to NewsWhip's *Social Network API* is automatically invoked periodically to maintain an updated record of trending news articles and social media posts. Assets detected by NewsWhip are enriched through social scoring. The URL, titles, summaries, images and videos (if any), along with the enrichment data, is extracted from the article and provided to Provenance. Assets composed only of text, for example, are registered in fragments consisting of news feed/article title, the summary, and user engagement data.

4.1.2. Asset Registration

A dedicated *Asset Registration* web interface also allows *Trusted Content Analysts* to add assets into the *Asset Workflow Handler*. *Trusted Content Analysts* are stakeholders such as journalists and other representatives of news agencies and wire services, fact checkers, de-

---

[13]https://www.newswhip.com

bunkers, and original content creators who may want to register their multimedia content assets. In future, this facility will be made more widely available to allow the general public to send content directly to Provenance. It may also be integrated with news publication platforms and content management systems so that content is automatically added. The primary task of this component is to enable third-parties to register assets that have not been discovered by the *Social Network Monitor*.

4.1.3. Asset Workflow Handler

The *Asset Workflow Handler* is the component of the *Provenance Verification Layer* that is responsible for orchestrating the components and data within the layer. This component's primary task is to distribute assets to different components for further processing. It invokes the service interfaces and handles the data flow between the services. By utilising the *Asset Workflow Handler*, components are loosely coupled, thus mitigating direct component-to-component communications. This will enable Provenance to work with the variety of APIs exposed from the existing tools/components. Moreover, the APIs can be adjusted to meet Provenance's specific needs. Due to this modular design, new components can be easily added to the *Provenance Verification Layer* (e.g., detection of bias [46], tabloidization [47], and hate speech [48]), and connected to the *Asset Workflow Handler*.

4.1.4. Video/Image Reverse Searcher

The *Video/Image Reverse Searcher* is a key component for creating a large-scale annotated dataset for detecting manipulated visual content. The dataset consists of three distinct parts. The first part includes 45,000 images, each captured by a unique device (i.e., 45,000 different cameras have been used). Half of these images are real, and the other half has been digitally manipulated by applying a random image processing operation to a local area of the image. Since the sensor pattern noise present in images is unique to each sensor (i.e., camera), this dataset introduces large diversity, such as noise. The second part of the dataset uses imaging software in cameras to introduce a large diversity of artefacts in images. Commonly available camera brands and models were identified and used to collect a dataset of 50,000 images. Half of these images were digitally manipulated using an advanced image editing method based on Generative Adversarial Networks (GAN) [49]. Finally, the third part of the dataset consists of 2,000 images downloaded from the Internet representing "real-life" (uncontrolled) manipulated images created by random people. For all of the manipulated samples collected for the third part of the data, the matching unmanipulated image was also collected. This component's primary task is to enable search operation for videos and images.

4.1.5. Video/Image Manipulation Detector

The Provenance *Video/Image Manipulation Detector* identifies if an image or video has been manipulated in comparison to its source. This work is based on the PIZZARO[14] project. It utilises recent developments achieved by deep learning-based methods to enable an instant detection of manipulations in visual content. In addition, use of the latest technologies based on Convolutional Networks will lead to tangible enhancements in integrity verification in visual content. The *Video/Image Manipulation Detector* increases trust and improves governance. The solution is designed to build a web-based system to assess visual content in a real-world setting. The *Video/Image Manipulation Detector* will further support the development of user skills in detecting false visual information themselves by providing a world-class image forensic technology. The *Video/Image Manipulation Detector* has a special focus on developing a solution that will be intuitive and easy to understand and interpret for end-users, thereby increasing its uptake by the public and its impact on the information system. This component's primary task is to detect if the image and video are manipulated by comparing them with previously registered images and videos in the system.

4.1.6. Asset Fingerprinter and Asset Registry

The *Asset Fingerprinter and Asset Registry* provide traceability of registered content. It is based on Blockchain technology, making content immutable and enabling the verification of the sources and alterations to the content. Registered assets are handed to the *Asset Fingerprinter* via the *Asset Workflow Handler*. Due to the General Data Protection Regulation (GDPR) and the size of some assets, the hash of the data is stored on Blockchain. Azure Storage is used as the Blockchain, and the assets themselves, including large files, are stored using an off-line storage service available to store multimedia files. Blockchain is used due to its innate data integrity which is important to prove the traceability of registered content if the tool was ever targeted as part of a combined disinformation and hacking campaign. This component's primary task is the traceability of registered content via Blockchain.

4.1.7. Text Similarity Detector

News is regularly republished nationally and locally from international wire services such as Reuters, Agence France-Presse (AFP) and Associated Press (AP). In a bid to lower costs, many news agencies who are not in competition negotiate deals to republish each other's content.

---

[14]http://zoi.utia.cas.cz/node/180/0459504

Similarly, less trustworthy news outlets often put 'spins' on existing articles, where correct articles are modified to contain false information.

To combat this, the *Text Similarity Detector* in Provenance attempts to verify the textual content of an article by comparing it to similar articles published elsewhere. A backlog of trustworthy articles is stored in an Elasticsearch database with a BM25 similarity index [50]. As BM25 under-performs with very long documents [51], only the title and first 10 sentences are used in the index. Once similar articles have been found the component searches for facts given in the query article in the similar ones. Facts in an article are found by taking sentences with a low subjectivity from TextBlob's sentiment analysis model [52]. The similarity of two facts is the cosine similarity of the vector embedding of both, which is provided by Google's multilingual text model [53]. If enough of the article's factual content cannot be verified, the plugin displays a warning.

4.1.8. Text Tone Detector

Intuitively, one would expect that impartial news sources would use impartial, unemotive language to convey the facts of a story. Recent research has shown that emotions such as fear, anger, sadness, doubt, and the absence of joy and happiness are indicative of misinformation and disinformation [54, 55, 56]. Provenance's *Text Tone Detector* is designed to identify emotions in text which may indicate that the news source is unreliable. Threshold values are used to determine whether caution should be shown, and the degree of caution is determined by how far the calculated value deviates from the threshold value.

4.1.9. Writing Quality Detector

Provenance's *Writing Quality Detector* computes a writing quality score (WQS) for the textual content the user is viewing and provides a warning when it falls below a threshold value. Writing quality is closely related to cohesion and coherence [57]. Within the context of news, high quality writing is indicative of paid professional journalism from mainstream, independent, and to a lesser degree, alternative news agencies, whereas low quality writing is indicative of amateur or unprofessional news production processes [58]. This high/low quality differentiation is also apparent in other domains such as academia, publishing, commercial, and blogs and information websites. While NLP techniques exist to derive writing quality [59], and others have called for it to be used to identify misinformation and disinformation [60, 61], only two examples of systems could be found in the literature which actually calculate writing quality [62, 63].

To calculate WQSs for Provenance, a dataset of news articles, blog posts, and other website content, much of which had characteristics symptomatic of disinformation, was annotated in a crowdsourced study to identify terms and phrases indicative of low quality writing. A WQS for each piece of content was then derived using a standard formula. This was subject to testing and expert evaluation to ensure the WQS the formula produced accurately reflected each piece of content. Models were then trained on the dataset which showed that the WQS could be automatically generated with a high degree of accuracy. These models and the overall process are currently undergoing formal evaluation.

4.1.10. Knowledge Graph and Knowledge Graph Builder

The Provenance *Knowledge Graph* stores a record of all the articles introduced to Provenance via the *Social Network Monitor* service or via Asset Registration from a *Trusted Content Analyst*. It is also a record of all analysis performed on said assets.

The content is organised according to concept, categories and topics. For example, a news article discussing politics can be categorised according to the left/right political spectrum followed by the topics discussed as shown in Figure 2. Each node at the article level is split according to text, image and video.

The output of the *Video/Image Reverse Searcher* includes the N most similar images/videos, distance measures and geometric validation results. The data from the *Video/Image Manipulation Detector* includes the probability of manipulations and the area of polygons. These are sent as JSON objects to the *Knowledge Graph* where they are stored as entities in a triplestore.

Modelling of Provenance data is achieved using a combination of the RDF Data Cube vocabulary [64] to store statistical information such as the outputs from the various analytical components, and the Dublin Core/BIBO vocabularies [65] to model bibliographic information about the assets themselves. Some use is also made of the FOAF[15] vocabulary to model information such as content publishers, which are naturally represented as `foaf:Agent` entities.

The *Knowledge Graph Builder* is responsible for exposing a REST API which the *Asset Workflow Handler* may use to upload assets as JSON, and then transforming the JSON into triples which are stored in a triplestore. In Provenance, this is achieved using JOPA [66]: a Java library which can be used to map POJOs to triples. Using Spring Boot[16], a REST API accepting JSON is exposed. The uploaded JSON is serialized into POJOs using Spring Boot's built-in version of Jackson. JOPA is then used to serialize the triples out to an RDF4J[17] instance.

---

[15]http://xmlns.com/foaf/spec/
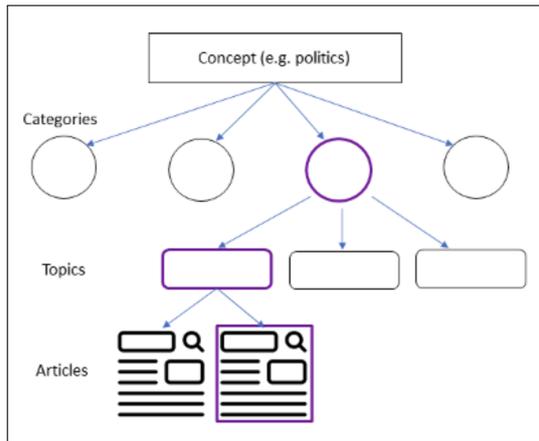[16]https://spring.io/projects/spring-boot
[17]https://rdf4j.org/

Figure 2: Knowledge Graph categorisations of assets.

The same serialization process works in reverse, allowing the *Provenance Query Service* to expose both a JSON REST endpoint which can produce JSON objects from the results of a canned SPARQL query exposed via a Spring Boot REST endpoint, and a much lower level raw SPARQL endpoint from the triplestore, for those who want a high level of control over their queries.

4.1.11. Provenance Query Service

The *Provenance Query Service* is the interface to the *Verification Layer* and offers external trusted services with the means to request verification information about a webpage or article. It will also allow trusted services with a means to identify the relatedness of content (through similarity and the *Knowledge Graph*) and determine if content has been modified. As the results of all analysis are stored in the *Knowledge Graph*, the *Provenance Query Service* is effectively a proxy between the user-facing front-end, and the query interface to whatever storage medium is used to implement the Knowledge Graph.

As mentioned in Section 4.1.10, the Provenance Query Service exposes both a raw SPARQL endpoint and a REST API which provides endpoints for a number of canned SPARQL queries which return JSON objects. It is envisioned that the vast majority of user cases will be covered by the REST API, making it easier for developers to access data that is helpful to users. However, it is worthwhile to allow lower level access to the KG's contents in the event of unforeseen requirements being placed on the KG.

4.1.12. Personalised Companion Service

The *Personalised Companion Service* manages the Provenance verification indicator, the minimal user model, and user scrutability and control. The verification indicator is implemented as a Chrome Extension and works on the Facebook and Twitter platforms and with articles published by news agencies. The *Personalised Companion Service* uses the user's interests, domain knowledge, digital literacy, and the warning preferences stored in the Minimal User Model to determine whether to highlight caution or show the verification indicator without caution. The *Personalised Companion Service* uses the data provided by the *Asset Fingerprinter*, the *Video/Image Reverse Searcher* and *Video/Image Manipulation Detector*, and the *Text Similarity, Tone and Writing Quality Detector* components to create the set of icons that are presented to users, who can explore the levels of verification presented through the visual iconography.

5. Provenance in Action

The Provenance browser plugin is designed to provide users with easy to understand, granular and cautionary warnings about the content they are consuming. These warnings are provided via an in-browser icon beside the address bar when the user is browsing the Internet, or within their Facebook and Twitter social media feeds beside the content they are viewing. Figures 3 - 6 show how Provenance and its visual warnings appear to a user - who has the Provenance plugin installed - within their Facebook social media feed. The Provenance icon appears as a small blue square with a white P above each content item that it has checked. When the icon background turns red (with a small exclamation mark), it indicates to the user that the content item is worthy of a cautionary warning. The following presents the four main states of Provenance which a user will see.

Figure 3 shows a user's Facebook feed who has the Provenance browser plugin installed. The Provenance icon is visible at the top of each news article in the user's feed. In this image, the icon is blue which indicates that there are no warnings with this particular news item.

In Figure 4, the background of the Provenance icon within the user's news feed has turned red to indicate that this news item is worthy of one or more cautionary warnings. A small black exclamation mark has been added to the top right of the icon for colour blind users.

In Figure 5, the user has clicked on the red Provenance icon. A window has appeared beneath the Provenance icon to show the user which of the seven criteria the news article was checked against that Provenance has detected an issue with. In this example, the red background and exclamation mark beneath the *Writing Quality* icon indicates that this aspect of the news article is worthy of caution. The user may click on the downward arrow beneath each icon for further information. In this example, the *Tone* icon is greyed out indicating that this could not be assessed by Provenance in this instance.
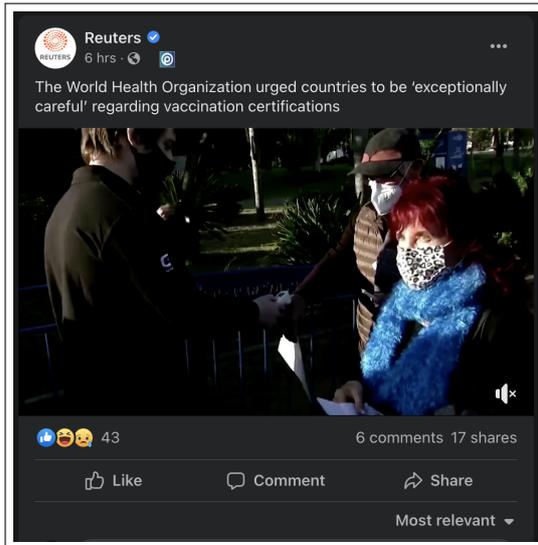
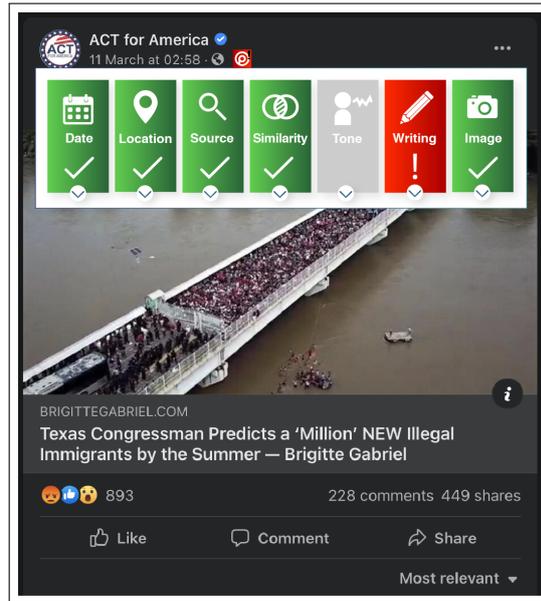**Figure 3:** A user's Facebook feed showing the Provenance icon in blue indicating that there are no warnings.

Figure 4: The Provenance icon in red (with exclamation mark) indicating that this article has one or more issues which are worthy of caution.
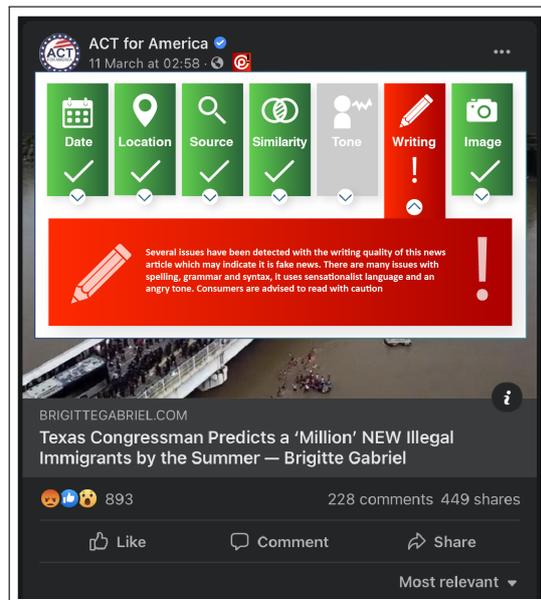
Figure 6 shows a detailed explanation of the *Writing Quality* warning after the user clicked on the option to expand it. It contains further information about how *Writing Quality* score is calculated and why low quality writing is indicative of misinformation and disinformation.

Figure 5: An initial explanation pane appears when then user clicks on the Provenance icon in their social media feed.

Figure 6: A detailed explanation pane appears when the user clicks on any of the seven categories Provenance analyses the news item under.

6. Use Cases: Provenance Plugin

6.1. Social Media Timeline

On the recommendation of a friend, Mary installed the Provenance browser plugin due to increased concerns

about the spread of misinformation and disinformation. The instructional video on the Provenance Chrome Extension webpage explained that Provenance uses seven criteria to verify digital content on the Internet and social media feeds. After installing the Provenance plugin, she notices that the news items in her Facebook timeline now display the Provenance icon beside the publisher's name. For most of the news stories, the Provenance icon shows a white P inside a white circle on a blue background. When she clicks on the blue Provenance icon, it opens a notification pane showing the seven verification criteria, all of which display a green background with a white ✓.

She is able to click on each of the seven verification icons to read a detailed explanation for each criterion, why failing the criterion is an indication that the webpage or social media post may be misinformation or disinformation, and how the warning is derived. As all of the icons are green, she is reassured about the origin, veracity and overall quality of the news article. For some news items displayed on her timeline, she notices that the blue background of the Provenance icon has turned red. When she clicks on it, the same information pane displaying the same verification criteria appears, except one or more of the seven verification criteria now display a red background with an exclamation mark beneath. When she clicks on these, an additional detailed explanation pane appears underneath them to explain why it has failed. Reading through each warning including their detailed description, she gains a better understanding of how to identify misinformation and disinformation. In both instances, Mary has become more aware of the need to critically check the news she consumes and more aware of good media literacy habits in general.

6.2. News Websites

Mary regularly visits news websites to inform herself of current affairs. Usually, the Provenance icon, which is visible to the right of her browser's address bar, displays a white P inside a white circle on a blue background. However, recently when she was visiting news websites to read more about a story relating to Covid 19 vaccination, she noticed that the background of the Provenance icon would sometimes turn red. When she clicked on the icon, the verification criteria information pane showed that Provenance had detected a problem with the image used in the news article she was reading. Clicking on the arrow to open the drop-down explanation pane, she reads that Provenance has detected that the image has been used before in another article. The image in question shows a picture taken at a conference of the World Health Organisation. Looking closely, she sees a credit to the Associated Press (AP). She knows that AP is an international news wire service, and that local and national news agencies republish their articles, including

the images. As this is just an image of a press conference, she is confident that its use by multiple news agencies is not an issue.

7. Evaluation

Provenance is under development and will shortly be undergoing human evaluation. Currently, five of the seven news analysis functions have been implemented and have been integrated with the platform. These are undergoing technical evaluation while the final two analysis tools are being completed. When the tool is fully completed, a series of technical tests and human evaluation tests will be undertaken to evaluate basic functionality and to ensure that it is providing the right warnings at the appropriate time. Following this, a series of experiments will be undertaken to evaluate its effect on user behaviour. This will include the likelihood of reading and sharing news articles that have cautionary warnings beside them. We will also be analysing unintended effects of the tool. Finally, a series of long term studies are planned to evaluate its effect on users' media literacy.

8. Conclusions

Misinformation and disinformation are significant issues that have negatively affected public discourse, politics and social cohesion. The Internet and especially social media are the primary conduits for its growth and spread. Existing user-orientated browser plugins have limited capabilities and only provide users with an historical rating of a website's propensity to publish misinformation and disinformation. They are also not capable of detailed analysis of the content of news webpages or social media feeds. The Provenance browser plugin significantly improves upon existing user orientated solutions by providing intermediary free analysis of webpage and social media content using seven criteria, and where necessary providing cautionary warnings to users. The user can then check the detailed explanatory warning notifications to make their own judgement. This will improve users' media literacy and reduce susceptibility to misinformation and disinformation long term.

9. Acknowledgements

References

[1] G. Rehm, An infrastructure for empowering internet users to handle fake news and other online media phenomena, in: G. Rehm, T. Declerck (Eds.), Language Technologies for the Challenges of the Digital Age, Lecture Notes in Computer Science, Springer International Publishing, 2018, p. 216–231. doi:10.1007/978-3-319-73706-5_19.

[2] E. Commission, Action plan against disinformation (2018). URL: https://ec.europa.eu/digital-single-market/en/news/action-plan-against-disinformation.

[3] E. Commission, Tackling online disinformation, 2017. URL: https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation.

[4] J. Bayer, N. Bitiukova, P. Bard, J. Szakács, A. Alemanno, E. Uszkiewicz, Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and its Member States, 2019. URL: https://papers.ssrn.com/abstract=3409279.

[5] 2021. URL: https://twitter.com/vonderleyen/status/1354030170789834755.

[6] A. Aker, A. Sliwa, F. Dalvi, K. Bontcheva, Rumour verification through recurring information and an inner-attention mechanism, Online Social Networks and Media 13 (2019) 100045. doi:10.1016/j.osnem.2019.07.001.

[7] Z. Marinova, J. Spangenberg, D. Teyssou, S. Papadopoulos, N. Sarris, A. Alaphilippe, K. Bontcheva, Weverify: Wider and enhanced verification for you project overview and tools, in: 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2020, p. 1–4. doi:10.1109/ICMEW46912.2020.9106056.

[8] M. Choraś, M. Pawlicki, R. Kozik, K. Demestichas, P. Kosmides, M. Gupta, Socialtruth project approach to online disinformation (fake news) detection and mitigation, in: Proceedings of the 14th International Conference on Availability, Reliability and Security, ARES '19, Association for Computing Machinery, 2019, p. 1–10. URL: https://doi.org/10.1145/3339252.3341497. doi:10.1145/3339252.3341497.

[9] L. Derczynski, K. Bontcheva, Pheme: Veracity in digital social networks (2014) 4.

[10] P. K. Srijith, M. Hepple, K. Bontcheva, D. Preotiuc-Pietro, Sub-story detection in twitter with hierarchical dirichlet processes, Information Processing & Management 53 (2017) 989–1003. doi:10.1016/j.ipm.2016.10.004.

[11] L. Toumanidis, R. Heartfield, P. Kasnesis, G. Loukas, C. Patrikakis, A prototype framework for assessing information provenance in decentralised social media: The eunomia concept, in: S. Kat- sikas, V. Zorkadis (Eds.), E-Democracy – Safeguarding Democracy and Human Rights in the Digital Age, Communications in Computer and Information Science, Springer International Publishing, 2020, p. 196–208. doi:10.1007/978-3-030-37545-4_13.

[12] D. Martín-Gutiérrez, G. Hernández-Peñaloza, J. M. Menéndez, F. Álvarez, A multi-modal approach for fake news discovery and propagation from big data analysis and artificial intelligence operations (2020) 3.

[13] D. Martín-Gutiérrez, G. Hernández-Peñaloza, A. B. Hernández, A. Lozano-Diez, F. Álvarez, A deep learning approach for robust detection of bots in twitter using transformers, IEEE Access 9 (2021) 54591–54601. doi:10.1109/ACCESS.2021.3068659.

[14] L. Ginsborg, P. Gori, Report on a survey for fact checkers on COVID-19 vaccines and disinformation, 2021. URL: https://cadmus.eui.eu//handle/1814/70917, accepted: 2021-04-26T08:57:47Z.

[15] U. N. S. Verified, Shareverified, 2021. URL: https://shareverified.com/en.

[16] W. H. Organisation, 1st who infodemiology conference, who infodemic management, 2020. URL: https://www.who.int/teams/risk-communication/infodemic-management/1st-who-infodemiology-conference.

[17] T. P. Institute, A guide to anti-misinformation actions around the world, 2021. URL: https://www.poynter.org/ifcn/anti-misinformation-actions/.

[18] 2021. URL: https://www.newsguardtech.com/.

[19] J. Nørregaard, B. D. Horne, S. Adalı, Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles, Proceedings of the International AAAI Conference on Web and Social Media 13 (2019) 630–638.

[20] A. Aker, V. Kevin, K. Bontcheva, Credibility and transparency of news sources: Data collection and feature analysis (2019) 6.

[21] Le Monde.fr (2017). URL: https://www.lemonde.fr/les-decodeurs/article/2017/01/23/le-decodex-un-premier-premier-pas-vers-la-verification-de-masse-de-l-information_5067709_4355770.html.

[22] 2021. URL: https://mediabiasfactcheck.com/.

[23] V. Kevin, B. Högden, C. Schwenger, A. Şahan, N. Madan, P. Aggarwal, A. Bangaru, F. Muradov, A. Aker, Information nutrition labels: A plugin for online news evaluation, ACL, 2018. doi:10.18653/v1/W18-5505.

[24] 2020. URL: https://browserextension.dev/blog/stopagandaplus-helps-understanding-media-biases/.

[25] P. Nordberg, J. Kävrestad, M. Nohlberg, Au-

tomatic detection of fake news, in: Proceedings of the 6th International Workshop on Socio-Technical Perspective in IS Development (STPIS 2020), CEUR-WS, 2020, p. 168–179. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-19356.

[26] K. Hartwig, C. Reuter, Trustytweet: An indicator-based browser-plugin to assist users in dealing with fake news on twitter (2019).

[27] A. Giełczyk, R. Wawrzyniak, M. Choraś, Evaluation of the existing tools for fake news detection, in: K. Saeed, R. Chaki, V. Janev (Eds.), Computer Information Systems and Industrial Management, Lecture Notes in Computer Science, Springer International Publishing, 2019, p. 144–151. doi:10.1007/978-3-030-28957-7_13.

[28] A. Školkay, J. Filin, A comparison of fake news detecting and fact-checking ai based solutions, Studia Medioznawcze 20 (2019) 365–383.

[29] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (2017) 22–36. doi:10.1145/3137597.3137600.

[30] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, I. Gurevych, A retrospective analysis of the fake news challenge stance-detection task, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, 2018, p. 1859–1874. URL: https://www.aclweb.org/anthology/C18-1158.

[31] A. Goel, ProjectFib - GitHub Repo, 2016. URL: https://github.com/anantdgoel/ProjectFib.

[32] Eyeo, 2020. URL: https://chrome.google.com/webstore/detail/trusted-news/nkkghpncidknplmlkgemdoekpckjmlok?hl=en.

[33] D. Paschalides, C. Christodoulou, R. Andreou, G. Pallis, M. D. Dikaiakos, A. Kornilakis, E. Markatos, Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web, in: 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2019, p. 298–302.

[34] V. Inc, Fakebox, 2021. URL: https://machinebox.io/.

[35] Z. A. Estela, N2ITN/are-you-fake-news, 2021. URL: https://github.com/N2ITN/are-you-fake-news.

[36] 2021. URL: https://ground.news/.

[37] A. Bhat, SurfSafe, 2021. URL: https://chrome.google.com/webstore/detail/surfsafe-join-the-fight-a/hbpagabeiphkfhbboacggckhkkipgdmh?hl=en.

[38] B. Botnevik, E. Sakariassen, V. Setty, Brenda: Browser extension for fake news detection, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Ma-

chinery, 2020, p. 2117–2120. URL: https://doi.org/10.1145/3397271.3401396.

[39] K. Popat, S. Mukherjee, J. Strötgen, G. Weikum, Credeye: A credibility lens for analyzing and explaining misinformation, in: Companion Proceedings of the The Web Conference 2018, WWW '18, International World Wide Web Conferences Steering Committee, 2018, p. 155–158. URL: https://doi.org/10.1145/3184558.3186967. doi:10.1145/3184558.3186967.

[40] FightHoax, Fighthoax - unlock your programmatic advertising, 2021. URL: http://34.253.212.69/.

[41] M. Hardalov, I. Koychev, P. Nakov, In search of credible news, in: C. Dichev, G. Agre (Eds.), Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Computer Science, 2016. doi:10.1007/978-3-319-44748-3_17.

[42] M. Hardalov, mhardalov/news-credibility, 2019. URL: https://github.com/mhardalov/news-credibility.

[43] X. Zhou, A. Jain, V. V. Phoha, R. Zafarani, Fake news early detection: A theory-driven model, Digital Threats: Research and Practice 1 (2020) 12:1–12:25. doi:10.1145/3377478.

[44] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep learning models in natural language processing: A survey, arXiv:1901.06796 [cs] (2019). URL: http://arxiv.org/abs/1901.06796, arXiv:1901.06796.

[45] Z. Zhou, H. Guan, M. M. Bhat, J. Hsu, Fake news detection via nlp is vulnerable to adversarial attacks, Proceedings of the 11th International Conference on Agents and Artificial Intelligence (2019) 794–800. doi:10.5220/0007566307940800, arXiv:1901.09657.

[46] B. Spillane, S. Lawless, V. Wade, The impact of increasing and decreasing the professionalism of news webpage aesthetics on the perception of bias in news articles, in: Proceedings of the 22nd International Conference On Human-Computer Interaction, Lecture Notes in Computer Science, Springer, 2020. doi:https://doi.org/10.1007/978-3-030-49059-1_50.

[47] B. Spillane, I. Hoe, M. Brady, V. Wade, S. Lawless, Tabloidization versus credibility: Short term gain for long term pain, in: CHI '20: The ACM Conference on Human Factors in Computing Systems, ACM, 2020. URL: https://dl.acm.org/doi/abs/10.1145/3313831.3376388. doi:http://dx.doi.org/10.1145/3313831.3376388.

[48] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, 2017,

p. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[50] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: SIGIR'94, Springer, 1994, pp. 232–241.

[51] Y. Lv, C. Zhai, When documents are very long, bm25 fails!, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, Association for Computing Machinery, 2011, p. 1103–1104. URL: https://doi.org/10.1145/2009916.2010070. doi:10.1145/2009916.2010070.

[52] S. Loria, textblob documentation (2020). URL: https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf, release 0.16.0.

[53] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic retrieval, 2019. arXiv:1907.04307.

[54] S. B. Parikh, V. Patil, P. K. Atrey, On the origin, proliferation and tone of fake news, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2019, p. 135–140. URL: https://ieeexplore.ieee.org/document/8695387/. doi:10.1109/MIPR.2019.00031.

[55] J. Paschen, Investigating the emotional appeal of fake news using artificial intelligence and human contributions, Journal of Product & Brand Management 29 (2019) 223–233. doi:10.1108/JPBM-12-2018-2179.

[56] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, K. Shu, Mining dual emotion for fake news detection, Proceedings of the Web Conference 2021 (2021) 3465–3476. doi:10.1145/3442381.3450004, arXiv: 1903.01728 version: 1.

[57] I. Singh, D. P., A. K., On the coherence of fake news articles, in: I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, O. Özgöbek, R. P. Ribeiro, R. Gavaldà, et al. (Eds.), ECML PKDD 2020 Workshops, Communications in Computer and Information Science, Springer International Publishing, 2020, p. 591–607. doi:10.1007/978-3-030-65965-3_42.

[58] M. Chung, N. Kim, When i learn the news is false: How fact-checking information stems the spread of fake news via third-person perception, Human Communication Research 47 (2021) 1–24. doi:10.1093/hcr/hqaa010.

[59] V. Klyuev, Fake news filtering: Semantic approaches, in: 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2018, p. 9–15. doi:10.1109/ICRITO.2018.8748506.

[60] M. Spradling, J. Straub, J. Strong, Protection from 'fake news': The need for descriptive factual labeling for online content, Future Internet 13 (2021) 142. doi:10.3390/fi13060142.

[61] N. Fuhr, A. Giachanou, G. Grefenstette, I. Gurevych, A. Hanselowski, K. Jarvelin, R. Jones, Y. Liu, J. Mothe, W. Nejdl, et al., An information nutritional label for online documents, ACM SIGIR Forum 51 (2018) 46–66. doi:10.1145/3190580.3190588.

[62] C. Fan, Classifying fake news, 2017. URL: https://www.conniefan.com/wp-content/uploads/2017/03/classifying-fake-news.pdf, connie Fan.

[63] E. S. Jo, A. Muhamed, S. Nuthakki, A. Singhania, DeepNews: Detecting Quality in News, 2018.

[64] W. W. W. Consortium, et al., The rdf data cube vocabulary (2014).

[65] D. C. M. Initiative, et al., Dublin core metadata element set, version 1.1 (2012).

[66] M. Ledvinka, P. Kremen, Jopa: Accessing ontologies in an object-oriented way., in: ICEIS (2), 2015, pp. 212–221.