Automatic and Manual Web Annotations in an Infrastructure to handle Fake News and other Online Media Phenomena

Georg Rehm, Julian Moreno-Schneider, Peter Bourgonje
DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
{georg.rehm, julian.moreno_schneider, peter.bourgonje}@dfki.de

Abstract

Online media are ubiquitous and consumed by billions of people globally. Recently, however, several phenomena regarding online media have emerged that pose a severe threat to media consumption and reception as well as to the potential of manipulating opinions and, thus, (re)actions, on a large scale. Lumped together under the label "fake news", these phenomena comprise, among others, maliciously manipulated content, bad journalism, parodies, satire, propaganda and several other types of false news; related phenomena are the often cited filter bubble (echo chamber) effect and the amount of abusive language used online. In an earlier paper we describe an architectural and technological approach to empower users to handle these online media phenomena. In this article we provide the first approach of a metadata scheme to enable, eventually, the standardised annotation of these phenomena in online media. We also show an initial version of a tool that enables the creation, visualisation and exploitation of such annotations.

Keywords: Tools, Systems, Applications; Linked Data; Semantic Web

1. Introduction

The amount of online news content is constantly growing. In addition to typical online news outlets, news are more and more consumed through social media channels. While about five or ten years ago only a fraction of the global population got their news online, nowadays online media and social networks are ubiquitous and used by a significant part of the population. By now we have an incredibly high amount of internet users who consume most or all of their news online. While there is still no clear consensus regarding the question whether the outcome of the 2016 US presidential election (Allcott and Gentzkow, 2017) has been manipulated through highly focused social media ads and large bot nets, we, nonetheless, now live in an age in which successful manipulations of online content (news, advertisements, likes, clicks, viral campaigns etc.) can have significant consequences in the real world. It is, however, important to state that there are different types of "fake news" – *false news* is a more appropriate term – and they all come with their own specific intentions and purposes (Rehm, 2018). Often the impact of false news can be harmless (April Fools jokes) or maybe less harmless but still limited in scope (pupils spreading false rumours to bully their peers).

Many online news outlets apply the same journalistic principles that have been in use for newspapers for decades, especially factchecking. However, there are also websites whose primary mission is not to provide high-quality journalistic reporting but, basically, any type of content as long as it produces as many clicks as possible, generating revenue through online advertisements. At first glance, online articles of this second type look just like normal journalistic content. In addition, there are other types of content whose design also mimicks news websites, for example, satire pages. This optical uniformity coupled with the fact that many users only scan the headlines, immediately taking them for fact, makes the World Wide Web and its users susceptible for manipulations and deceptions. This is why online users need to be equipped with additional tools and technologies, they need to be empowered to handle modern online media phenomena, most importantly by helping them to assess the quality of a piece of content, its accuracy, trustworthiness and reputation. Checking the facts and assessing the trustworthiness of online content are increasingly left to the reader. For this purpose, distributed automatic but also semi-automatic approaches can be applied.

In order to empower users to handle the online phenomena mentioned above, in an adequate way, several different approaches could be realised. A consensus seems to emerge that fully automatic means are most likely insufficient properly to address the issue, i. e., we need to combine automatic tools and the wisdom of the crowd through manual or, rather, intellectual assessments in the form of annotations that users attach to a piece of content. Annotations can be made by human users but also by automatic filters, classifiers and watchdogs. In a follow-up step, users can be informed about any issues that the automatic tools or human peers have with the content (manipulated, satirical, imposter content, etc.). In this article we focus on the first steps towards the definition of an annotation scheme to be used both by humans and by machines so that the needed metadata can be added to arbitrary pieces of online content in the form of annotations. We also demonstrate the current version of a browser plugin that allows the creation, visualisation and, eventually, exploitation of the different annotations.

The remainder of this article is structured as follows. First, Section 2. describes related work, while Section 3. briefly sketches the overall infrastructural concept. Section 4. provides an initial draft of a false news annotation schema. Section 5. illustrates the annotation tool. Finally, Section 6. concludes the article.

2. Related Work

Despite the recent increase in research in this area, an effective technological antidote against false news is yet to be found (Rehm, 2018; Rubin et al., 2015). One common denominator of all related work is that they address specific aspects of the broad set of content phenomena. The EU

project Pheme (Derczynski and Bontcheva, 2014), for example, focused on modeling, identifying, and verifying online rumours (Srivastava et al., 2017), which they call "phemes" (internet memes with added truthfulness or deception), as they spread across media, languages, and social networks. (Conroy et al., 2015) looks into different veracity assessment methods emerging from two major categories, i. e., linguistic cues (through machine learning) and network analysis. Martinez-Alvarez (2017) notes in this context: "Fake news is a too general and too vague problem to address directly.", which is why he is splitting it up into smaller, more approachable problems: fact checking, source credibility and trust, news bias and misleading headlines.

Factchecking is a key characteristic of high-quality news content and journalism in general. A large number of factchecking initiatives is active all over the world (Mantzarlis, 2017) but they mostly rely on human expertise and, thus, do not scale (Martinez-Alvarez, 2017; Dale, 2017). The small number of automated fact checking initiatives is fragmented, unreliable and not efficient (Babakar, Mevan and Moy, Will, 2016).

Identifying bias in online articles is another fundamental challenge. Watanabe (2017) analyses the influence of the Russian government on ITAR-TASS during the Ukraine crisis using the state-owned news agency, while Yeo et al. (2017) studied the effect of uncivil comments in online news articles in order to avoid bias interpretations. apply neural networks to identify polarity in news. Valdeón (2017) examines the impact of bias introduced in translations.

Clickbait is often subsumed under the label of "fake news" (Bourgonje et al., 2017). Wei and Wan (2017) identifies misleading headlines – supposed to generate clicks – using class sequential rules to exploit structure information in ambiguous headlines. The BuzzFeed Marketing Challenge (Cowley, 2017) encourages the creation, publication and promotion of an article for generating 1,000 article views in one week. An overview of clickbait analysis approaches was published by Chen and Rubin (2017). Important related characteristics are also the dissemination (Maheshwari, 2016) and spreading (Giglietto et al., 2016) false news exhibit.

Satirical articles are in stark contrast to false news whose objective often is, in the severe cases, to misinform and to manipulate. Rubin et al. (2016) show that online satire often mimics the format and style of journalistic reporting. In addition to false news, other online phenomena need to be taken into account such as abusive language (Nobata et al., 2016; Bourgonje et al., 2018) and hatespeech (Warner and Hirschberg, 2012; Djuric et al., 2015; Schmidt and Wiegand, 2017).

A large number of industrial approaches are focusing on detecting fake news. In early 2017, Facebook announced that they collaborate with the factchecking group Correctiv[1] in Germany (Reuters, 2017). Fakeblok[2] is a Chrome plugin that aims to sanitize the Facebook newsfeed from fake news sites using a curated, factchecked and monitored list of links curated by a group of independent media professionals. Another relevant tool is Fakenews Dataset Analysis, a machine learning system that analyses online news and provides a user-friendly visualisation.[3] Hoaxy[4] visualises how reported claims – and checks of those claims – spread online through social networks. Facebook, Google and seventeen French news organisations joined forces to combat fake news through an initiative called CrossCheck, which uses tools such as CrowdTangle or Spike.[5] There are also several startups in this space like, for example, Factmata, who use NLP and IR algorithms.[6] Among the factchecking initiatives are FactCheck.org,[7] PolitiFact[8] and Fact Checker.[9] These initiatives rely on the claim of Baker (2017): "Human-led fact-checking is the most obvious (and longstanding) weapon against misinformation."

3. Infrastructure Concept

In (Rehm, 2018) we define an infrastructure for the handling and processing of fake news and related phenomena. Here, we give a brief overview of the envisioned system (Figure 1).

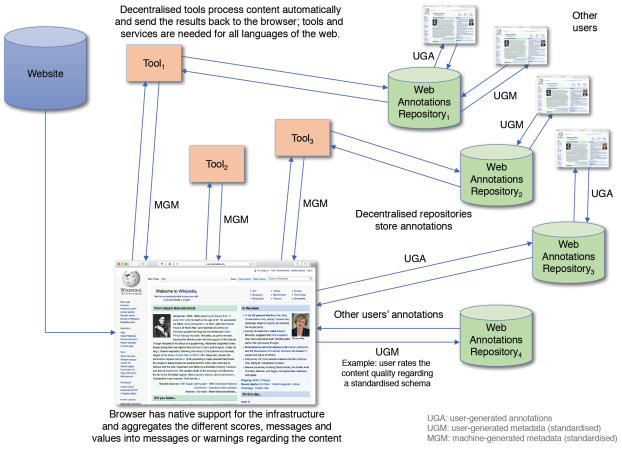

Figure 1: Simplified architecture of the infrastructure

The infrastructure is to be natively embedded into the architecture of the World Wide Web itself. It should rely on standards and be endorsed and supported not only by all browser vendors but also by all content and media providers. The services must be unobtrusive and cooperative, their recommendations and warnings must be clearly understandable. Several pieces are already in place: Web Annotations, standardised by W3C in early 2017 (Sanderson et al., 2017a; Sanderson et al., 2017b; Sanderson, 2017), enable users to annotate arbitrary pieces of web content, creating an independent layer on top of the regular web. They are *the* natural mechanism to enable users to work with content and to include feedback and assessments. Their content can be automatically mined but there are still limitations. Content providers need to enable Web Annotations. Federated

[1] https://correctiv.org
[2] https://fakeblok.com
[3] https://github.com/melphi/fakenews-analysis
[4] http://hoaxy.iuni.iu.edu
[5] https://crosscheck.firstdraftnews.com
[6] http://factmata.com
[7] http://factcheck.org
[8] http://www.politifact.com
[9] https://www.washingtonpost.com/news/fact-checker/

sets of annotation stores are not yet foreseen, neither are native controls in browsers that provide aggregated feedback. Browsers should, ideally, enable free-text annotations and simple flagging of problematic content, e. g., "content pretends to be factual but is of dubious quality". Annotations could be aggregated and presented to new readers to provide guidance and indicate issues. Automatic tools and services can also make use of Web Annotations, e. g., simple classifiers (e. g., regarding abusive language), or sophisticated NLU components that attempt to fact-check statements against knowledge bases. The results can be made available as globally accessible Web Annotations. Also needed is an agreed upon metadata schema (Babakar, Mevan and Moy, Will, 2016) to be used in manual or automatic annotation scenarios. Its complexity should be as little as possible so that key content characteristics can be adequately captured and described by humans or machines. W3C published standards to represent the provenance of objects (Groth and Moreau, 2013; Belhajjame et al., 2013a; Belhajjame et al., 2013b). An alternative approach is Schema.org's ClaimReview markup.[10] Furthermore, the architectural setup must be federated and decentralised to prevent abuse. Annotations must be stored in decentral repositories. These will soon also include more complex data, information and knowledge that tools and services will make use of, e. g., for fact checking. Crowd-sourced knowledge graphs such as Wikidata or DBpedia will continue to grow, the same is true for semantic databases, usually available as Linked Open Data. Already now we can foresee more sophisticated methods of validating and fact-checking content using systems that make use of knowledge graphs, e. g., through entity recognition and linking, relation and event extraction. Finally, we need to be able to aggregate manual and automatic annotations.

4. Annotation Approach and Schema

One of the next steps towards a first prototype is the definition of an annotation schema so that online content can be marked up, both by humans and by machines. We work with an ontology composed of a set of classes and relations that allow automatically processing the annotated data. As the infrastructure is meant to be natively embedded into the web technology stack, we work with the W3C standard for Web Annotations (Sanderson et al., 2017a), which is ideally suited to address the phenomena discussed in this article. The annotation schema includes properties and classes taken from the Provenance Ontology (Prov-O) (Belhajjame et al., 2013a), created for the annotation of the provenance of annotations.

The infrastructure needs to be able to process the following three different types of annotations (Rehm, 2018).

Machine-Generated Metadata (MGM) are automatically generated by a specific service that analyses and annotates a piece of text or multimedia content (image, video, etc.) accordingly, e. g., by assigning a respective score for a given content dimension such as political bias or veracity.

User-Generated Metadata (UGM) are manually annotated by a user through an interface. The user manually assigns a set of predefined tags or scores to express an opinion about the content with the help of a controlled vocabulary, for example, "content is not factual and intentionally misleading".

User-Generated Annotations (UGA) are free text annotations added by a user, i. e., essentially a natural language comment regarding a piece of content.

The annotation schema is defined in an experimental ontology, FANE (Fake News Ontology), which makes use of the Web Annotation standard combined with relevant existing ontologies, such as the Prov-O ontology for provenance information. At the current stage of the implementation the main goal of this experimental schema is to illustrate the overall approach and to demonstrate technical feasiblity. The ontology is not meant to be complete or all-encompassing, for example, currently we are studying the inclusion/mapping of Schema.org's ClaimReview.

In these ontologies we already have all the necessary mechanisms to make annotations in texts or multimedia content in the Web. Therefore, we only need a formal definition of how to annotate the different types of fake news and the values (degree of membership in a fake news type) associated with each of these annotations. Therefore, FANE defines several additional classes and relations (cf. Figure 2).

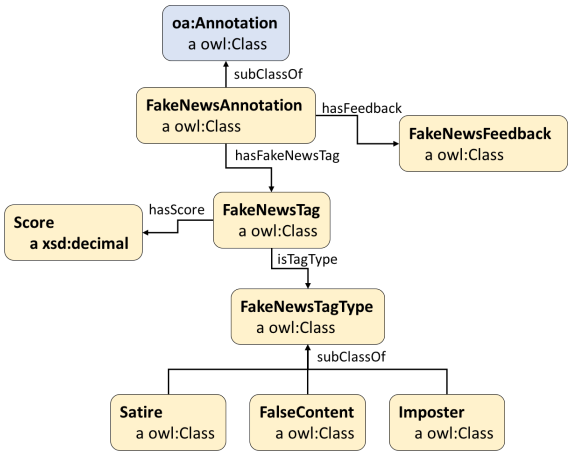

Figure 2: Classes and relations defined in the experimental Fake News Ontology

To simplify the notation, from now on we summarise the URLs, replacing *http://persistence.dfki.de/ontologies/fane-core#* with *"fane:"*. The different classes and properties defined in the ontology are summarised in Table 1.

There are currently seven classes in the ontology (see Table 2). The three relations are described in Table 3.

These classes and relations allow the annotation ("FakeNewsAnnotation") of arbitrary web content in order to tag it ("FakeNewsTag") as different types of false news. The three currently implemented different types ("Satire", "Imposter" and "FalseContent") are defined through the "FakeNewsTagTypes" class allowing the assignment of multiple false news types to the same annotation and are meant to be illustrative examples only.

"FakeNewsFeedback" is a class and mechanism included to allow users to provide feedback on such annotations, which is meant to be a solution for the misuse of such annotations, i. e., a "false news" annotation that has the objective of discrediting real news can be encountered through the feedback

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

Name	Definition
FakeNewsAnnotation	A piece of content annotated by a human user or by an automatic process, on a general level, as false news (can be either UGA, MGM or UGM)
FakeNewsTag	Defines FakeNewsTag associated with the annotation
FakeNewsTagType	Defines the type of the FakeNewsTag
FakeNewsFeedback	Defines feedback about fake news annotations provided by users
Satire	Type Satire
FalseContent	Type False Content
Imposter	Type Imposter
...	*Any additional types of false news*
hasFakeNewsTag	References the FakeNewsTag associated with a false news annotation
isTagType	References the false news type associated with the content
hasScore	Specifies a score associated with the annotation (either MGM or UGM)

Table 1: Classes and properties defined in the Fake News Ontology

of others users affirming that the news is real and that the annotation is false (or vice versa, of course).
In order further to explain the schema, we are going to exemplify it together with an annotated news example (cf. Listing 1) selected from the satirical website The Onion.[11]
The example combines annotations from different ontologies (defined in the namespace part) allowing the use of web annotations and provenance information in a simple way. The main element, "**ex:anno1**", is a web annotation ("**oa:Annotation**") and also a false news annotation ("**fane:FakeNewsAnnotation**"). This annotation is associated with:

- the content defined in "**ex:target1**", which is at the same time associated with a source ("<http://goo.gl/pD9gVE>"), a selector defining the concrete part of text that was annotated and a person who generated the content ("**ex:person2**")
- the person who created the annotation ("**ex:person1**")
- the annotation itself in "**ex:body1**" containing value, format and language
- a creation date ("2017-09-28T16:48:00Z")
- the annotation activity through it was generated "**ex:annotationActivity**"
- the fake news annotations tags "**ex:fnTag1**" and "**ex:fnTag2**"

The fake news information is annotated in "**ex:fnTag1**" and "**ex:fnTag2**", which are fake news tags associated with a fake news tag type ("**fane:isTagType**") and a decimal score or ranking ("**fane:hasScore**"). In this example, the first tag is a "**fane:Satire**" tag, while the second is a "**fane:Imposter**" tag.
Regarding the provenance of the content and annotations, "**ex:person1**" and "**ex:person2**" are physical persons ("**foaf:Person**" and "**prov:Agent**") belonging to companies ("**ex:dfki_gmbh**" and "**ex:the_onion**") who generated content (text) and annotations.

```
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dc:   <http://purl.org/dc/elements/1.1/> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix ex:   <http://example.org/> .
@prefix fane: <http://persistence.dfki.de/ontologies/fane#> .

ex:anno1 a oa:Annotation,prov:Entity,fane:FakeNewsAnnotation ;
  oa:hasBody ex:body1 ;
  oa:hasTarget ex:target1 ;
  oa:motivatedBy oa:describing ;
  dcterms:created "2017-09-28T16:48:00Z" ;
  prov:wasGeneratedBy ex:annotationActivity ;
  prov:wasAttributedTo ex:person1 ;
  fane:hasFakeNewsTag ex:fnTag1, ex:fnTag2 .

ex:body1 a oa:TextualBody ;
  rdf:value "This affirmation is completely false." ;
  dc:format "text/plain" ; dc:language "en" .

ex:target1 a prov:Entity ;
  oa:hasSource <http://goo.gl/pD9gVE> ;
  oa:hasSelector ex:target1_selector ;
  prov:wasAttributedTo :person2 .

ex:target1_selector a oa:TextPositionSelector ;
  oa:start 257 ; oa:end 303 ;
  oa:exact "Obamacare is collapsing under its own weight," .

ex:fnTag1 a fane:FakeNewsTag ;
  fane:hasScore 0.9^^xsd:decimal ;
  fane:isTagType ex:fnTagType1 .

ex:fnTagType1 a fane:FakeNewsTagType, fane:Satire .

ex:fnTag2 a fane:FakeNewsTag ;
  fane:hasScore 0.7^^xsd:decimal ;
  fane:isTagType ex:fnTagType2 .

ex:fnTagType2 a fane:FakeNewsTagType, fane:Imposter .

ex:person1 a foaf:Person, prov:Agent ;
  foaf:givenName "Julian" ;
  prov:actedOnBehalfOf ex:dfki_gmbh .

ex:dfki_gmbh a foaf:Organization, prov:Agent ;
  foaf:name "DFKI GmbH" .

ex:annotationActivity a prov:Activity ;
  prov:wasAssociatedWith ex:person1 ;
  prov:startedAtTime "2011-07-14T01:01:01Z"^^xsd:dateTime ;
  prov:used ex:target1 ;
  prov:endedAtTime "2011-07-14T02:02:02Z"^^xsd:dateTime .

ex:person2 a foaf:Person, prov:Agent ;
  foaf:givenName "Article author" ;
  prov:actedOnBehalfOf ex:the_onion .

ex:the_onion a foaf:Organization, prov:Agent ;
  foaf:name "The Onion" .
```

Listing 1: Content annotation example

5. Current Prototype

As described above and, in detail, in Rehm (2018), our goal is to use a set of decentralised automatic tools and services (that add MGM to content) in tandem with information added by users (UGM, UGA). The respective annotations are stored in decentralised Web Annotation repositories. Whenever a user retrieves online content to be rendered in

[11] http://www.theonion.com

Name	URI	Label	SubClass of
FakeNewsAnnotation	fane:FakeNewsAnnotation	Fake News Annotation	http://www.w3.org/ns/oa#Annotation
FakeNewsTag	fane:FakeNewsTag	Fake News Tag	
FakeNewsTagType	fane:FakeNewsTagType	Fake News Tag Type	
FakeNewsFeedback	fane:FakeNewsFeedback	Fake News Feedback	
Satire	fane:Satire	Satire	fane:FakeNewsTagType
FalseContent	fane:FalseContent	False Content	fane:FakeNewsTagType
Imposter	fane:Imposter	Imposter	fane:FakeNewsTagType

Table 2: Classes in the Fake News Ontology

Name	URI	Label	Domain	Range
hasFakeNewsTag	fane:hasFakeNewsTag	has FN Tag	fane:FakeNewsAnnotation	fane:FakeNewsTag
isTageType	fane:isTagType	is Tag Type	fane:FakeNewsTag	fane:FakeNewsTagType
hasScore	fane:hasScore	has Score	fane:FakeNewsTag	xsd:decimal

Table 3: Relations in the Fake News Ontology

the browser, the browser then retrieves the available information about the content (MGM, UGM, UGA) from the currently configured repositories, aggregates them into easily consumable values and displays these values to the user, for example, through a traffic light metaphor or through a set of reputation and confidence scores.

The prototype relies on Web Annotations, which are not yet natively supported by all browsers. As soon as there is native support in all browsers, the solution we propose will develop its full potential, i. e., users will be able automatically to get clear signals and recommendations with regard to the content they are currently seeing in their browsers – for example, whether to trust it or to take it with a grain of salt. We are currently developing a prototype of such a browser feature in the form of a plugin, which offers the possibility of adding UGA and UGM to content. The implementation is mostly a technical challenge; the process of annotating web resources does not rely only on the graphical interface, but also on servers in the backend that enable automatic content processing and the generation of annotations (text classification, dealing with author/source information, storing the resulting annotations, etc.).

Most of these features are present in the Web Annotation infrastructure provided by Hypothes.is.[12] The Hypothes.is ecosystem not only allows the annotation of online content, but also the annotation of comments and also annotations made by users (UGM, UGA) or machines (MGM). The Hypothes.is tools are currently tailored to Chrome but as soon as Web Annotations are natively available in all browsers, the solution we propose will be universally available without the cumbersome installation of needed plugins.

Our current prototype consists of three main components: the client, the server and the web extension (GUI). For now we use the Hypothes.is infrastructure as our client/server architecture. On the interface level we are making the necessary modifications so that users can not only add and visualise annotations based on the annotation schema (Section 4.), but also provide feedback. The interface (cf. Figure 3) is an adapted version of the Hypothes.is extension, through which users can automatically annotate web content, including numeric scores for different types of false news (Rehm, 2018).

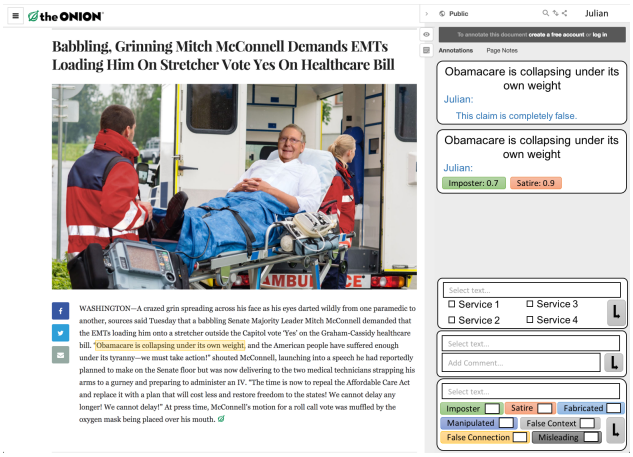

Figure 3: Annotated online content

As regards automatic classification services (MGM), we have conducted several experiments, especially for rumours, for clickbait content and for abusive language and hate-speech (Srivastava et al., 2017; Bourgonje et al., 2018; Bourgonje et al., 2017). We are currently working on attaching these experimental services to the backend so that respective analysis results are automatically shown to the user.

6. Conclusions and Future Work

There is currently a lot of interest in the wider Computational Linguistics, Language Technology and AI community on the general topic of fake news and related online content phenomena as can be seen by the high number of dedicated workshops, e. g., Fake News Challenge,[13] Abusive

[12] https://web.hypothes.is

[13] http://www.fakenewschallenge.org

Language Workshop,[14] Computational Fake News Analysis,[15] the SemEval task on Rumour Evaluation[16] and the Clickbait Challenge,[17] among others. This interest shows that there is a real need and high demand for solutions. The infrastructural approach suggested in (Rehm, 2018), further extended with practical next steps towards an annotation schema and prototype client application in this paper, is an attempt at providing an umbrella application scenario that is both practical, universally applicable in the real world by relying on W3C standards and actually usable and flexible enough to take on board many different types of decentralised classifiers and automatic services.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions.
The project "Digitale Kuratierungstechnologien" (DKT) is supported by the German Federal Ministry of Education and Research (BMBF), "Unternehmen Region", instrument Wachstumskern-Potenzial (no. 03WKP45). More information: http://www.digitale-kuratierung.de.

7. Bibliographical References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, May.

Babakar, Mevan and Moy, Will. (2016). The State of Automated Factchecking – How to make factchecking dramatically more effective with technology we have now. https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf.

Baker, T. (2017). Fake news: What is it, and how can we tackle it? http://www.nesta.org.uk/blog/fake-news-what-it-and-how-can-we-tackle-it.

Belhajjame, K., Cheney, J., Corsark, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013a). PROV-O: The PROV Ontology. W3C Recommendation, World Wide Web Consortium (W3C), April. https://www.w3.org/TR/2013/REC-prov-o-20130430/.

Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., and Zednik, S. (2013b). PROV Model Primer. W3C Working Group Note, World Wide Web Consortium (W3C), April. https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/.

Bourgonje, P., Schneider, J. M., and Rehm, G. (2017). From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In Octavian Popescu et al., editors, *Proceedings of the Second Workshop on Natural Language Processing meets Journalism – EMNLP 2017 Workshop (NLPMJ 2017)*, pages 84–89, Copenhagen, Denmark, September. 7. September.

Bourgonje, P., Schneider, J. M., and Rehm, G. (2018). Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication. In Georg Rehm et al., editors, *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 180–191, Cham, Switzerland, January. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Cavalcanti, R. D., Lima, P. M., De Gregorio, M., and Menasche, D. S. (2017). Evaluating weightless neural networks for bias identification on news. In *Networking, Sensing and Control (ICNSC), 2017 IEEE 14th International Conference on*, pages 257–262. IEEE.

Chen, Y. and Rubin, V. L. (2017). Perceptions of clickbait: A q-methodology approach. In *Proceedings of the 45th Annual Conference of The Canadian Association for Information Science/L'Association canadienne des sciences de l'information (CAIS/ACSI2017), Ryerson University, Toronto, May 31-June 2, 2017.*

Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, pages 82:1–82:4, Silver Springs, MD, USA. American Society for Information Science.

Cowley, S. W. (2017). The buzzfeed marketing challenge: An integrative social media experience. *Marketing Education Review*, 27(2):109–114.

Dale, R. (2017). NLP in a post-truth world. *Natural Language Engineering*, 23(2):319–324.

Derczynski, L. and Bontcheva, K. (2014). Pheme: Veracity in digital social networks. In Iván Cantador, et al., editors, *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014.*, volume 1181 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 29–30, New York, NY, USA. ACM.

Giglietto, F., Iannelli, L., Rossi, L., and Valeriani, A. (2016). In *Convegno annuale dell'Associazione Italiana di Comunicazione Politica (AssoComPol 2016), Urbino, Dec 15-17, 2016.*

Groth, P. and Moreau, L. (2013). PROV-Overview: An Overview of the PROV Family of Documents. W3C Working Group Note, World Wide Web Consortium (W3C), April. https://www.w3.org/TR/prov-overview/.

Maheshwari, S. (2016). How Fake News Goes Viral: A Case Study. https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html?mcubz=0.

Mantzarlis, A. (2017). There are now 114 fact-checking initiatives in 47 countries. https://www.poynter.org/news/there-are-now-114-fact-checking-initiatives-47-countries.

[14] https://sites.google.com/site/abusivelanguageworkshop2017

[15] http://www.sobigdata.eu/computational-fake-news-analysis-practical-workshop

[16] http://alt.qcri.org/semeval2017/task8/

[17] http://www.clickbait-challenge.org

Martinez-Alvarez, M. (2017). How can machine learning and ai help solving the fake news problem? https://miguelmalvarez.com/2017/03/23/how-can-machinelearning-and-ai-help-solving-the-fake-news-problem/.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Rehm, G. (2018). An Infrastructure for Empowering Internet Users to handle Fake News and other Online Media Phenomena. In Georg Rehm et al., editors, *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 216–231, Cham, Switzerland, January. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Reuters. (2017). Facebook germany announces tools to tackle fake news ahead of election. http://www.abc.net.au/news/2017-01-16/facebook-germany-says-it-will-start-tackling-fake-news-in-weeks/8184436.

Rubin, V. L., Chen, Y., and Conroy, N. J. (2015). Deception detection for news: Three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, pages 83:1–83:4, Silver Springs, MD, USA. American Society for Information Science.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17.

Sanderson, R., Ciccarese, P., and Young, B. (2017a). Web Annotation Data Model. W3C Recommendation, World Wide Web Consortium (W3C), February. https://www.w3.org/TR/2017/REC-annotation-model-20170223/.

Sanderson, R., Ciccarese, P., and Young, B. (2017b). Web Annotation Vocabulary. W3C Recommendation, World Wide Web Consortium (W3C), February. https://www.w3.org/TR/2017/REC-annotation-vocab-20170223/.

Sanderson, R. (2017). Web Annotation Protocol. W3C Recommendation, World Wide Web Consortium (W3C), February. https://www.w3.org/TR/2017/REC-annotation-protocol-20170223/.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Srivastava, A., Rehm, G., and Schneider, J. M. (2017). DFKI-DKT at SemEval-2017 Task 8: Rumour Detection and Classification Using Cascading Heuristics. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 477–481, Vancouver, Canada, August. Association for Computational Linguistics.

Valdeón, R. A. (2017). Political and sexist bias in news translation: Two case studies. *Trans. Revista de Traductología*, (11):231–243.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Watanabe, K. (2017). Measuring news bias: Russia's official news agency itar-tass'coverage of the ukraine crisis. *European Journal of Communication*, 32(3):224–241.

Wei, W. and Wan, X. (2017). Learning to identify ambiguous and misleading news headlines. *arXiv preprint arXiv:1705.06031*.

Yeo, S. K., Su, L. Y.-F., Scheufele, D. A., Brossard, D., Xenos, M. A., and Corley, E. A. (2017). The effect of comment moderation on perceived bias in science news. *Information, Communication & Society*, pages 1–18.