K + K = 120: Papers dedicated to L. Kálmán & A. Kornai on the occasion of their 60th birthdays, pp. 461–479

An infrastructure for empowering internet users to handle fake news and other online media phenomena

GEORG REHM
DFKI GmbH, Language Technology Lab
georg.rehm@dfki.de

KEYWORDS

fake news
online misinformation
annotation
language technology
computational linguistics

ABSTRACT

Online media, online news and online communication have an unprecedented and increasing level of social, political and also economic relevance. This article proposes an infrastructure to address phenomena of modern online media production, circulation and reception by establishing a distributed architecture that relies on automatic processing and human feedback.

1. Introduction

Usually lumped together under the "fake news" label, a bundle of novel topics around online media production, circulation, reception and their impact has emerged in recent years and, thus, is receiving a lot of attention from multiple stakeholders including politicians, journalists, researchers, industry, non-governmental organisations (NGOs) and civil society. In addition to the challenge of addressing and dealing with "fake news", "alternative facts" as well as "post-truth politics", there is an ever increasing amount of hate speech, abusive language and cyber bullying taking place online.[1]

Among the interested stakeholders are politicians who have begun to realise that, increasingly, major parts of public debates and social discourse are carried out online, on a small number of social networks. We have witnessed that not only online discussions but also the perception of trends, ideas, theories, political parties, individual politicians, elections and societal challenges can be subtly influenced and significantly rigged using targeted social media campaigns, devised at manipulating opinions to create long-term sustainable mindsets on the side of the recipients. We live in

[1] A revised version of this paper was published in Rehm & Declerck (2017).

a time in which online media, online news and online communication have an unprecedented level of social, political and also economic relevance.

Due to the sheer importance and visibility of the topic one cannot help but think about designing and deploying technologies to improve the situation, maybe even to solve the problem altogether – thanks to the recent breakthroughs in artificial intelligence (AI) (Metz 2016; Gershgorn 2016; Martinez-Alvarez 2017; Chan 2017) –, while at the same time *not* putting in place a centralised infrastructure that could be misused for the purpose of censorship, media manipulation or mass surveillance.[2]

This paper addresses key challenges of the digital age (section 2) by introducing and proposing a technological framework concept (section 3), which has been devised under the umbrella of a two-year research and technology transfer project, in which a research centre collaborates with four small and medium-sized enterprise (SME) partners that face the challenge of having to process, analyse and make sense of large amounts of digital content. The companies cover four different use cases and sectors (Rehm & Sasaki 2015) including journalism. For these partners we develop a platform that provides access to language and knowledge technologies Bourgonje et al. (2016a;b). The services are integrated by the SME partners into their own in-house systems or those of clients (Rehm et al. 2017). Among others, we currently develop services aimed at the detection and classification of abusive language and clickbait content.

2. Online media in 2017: status quo

The debate around online media is currently dominated by several topics and challenges. They share certain characteristics that make it possible to address them with the same technological approach.

A key prerequisite for the current situation is the existence of the World Wide Web itself: everybody is able to create content, to write an article on a certain topic. Until a few years ago the key challenge was to optimise the HTML code, linking and metadata to get into the top of the relevant search engine results pages for important keywords. Nowadays, however, content is no longer predominantly discovered through search engines but through social media platforms: users see interesting content, which is then shared to their own connections. Many users only read a

[2] An indicator for the relevance of the topic is the increasing number of "how to identify fake news" articles published online (Mantzarlis 2015; Bazzaz 2016; Rogers & Bromwich 2016; Wardle 2017; Walbrühl 2017).

headline, identify a certain relevance to their own life and then spread the content. When in doubt, users estimate the trustworthiness of the source: potentially dubious stories about which they are skeptical are shared anyway if the source or friend through whom the story was discovered is considered reliable or if the number of views is rather high, which, to many users, indicates legitimacy.

There is a tendency for very provocative, aggressive, one-sided, allegedly "authentic" (Marchi 2012) content. The idea is to make it as easy as possible to identify the stance of the article so that the reader's own world view is validated, implicitly urging the user to share the content. The hope of the content's originator is that a story will go viral, that it will be shared very quickly by many users and spread through multiple networks in order to establish a reach of millions of people. One sub-category of this type of content is "clickbait", articles with dubious factual content that are presented with misleading headlines, designed for the purpose of generating many clicks. The more extreme the virality, the higher the reach, the higher the click numbers, the higher the advertisement revenue. The term "clickbait" is usually associated with commercial intentions, but it can also refer to articles spreading political mis- or disinformation.

Content is, first and foremost, discovered through a small number of big social networks. While only a handful of search engines and online news outlets used to be the central points of information until a few years ago, the role of the centralised hub is now played by social networks that help content to be discovered and go viral (Barthel et al. 2016). All social networks have the same key feature, a feed or timeline, i.e., posts, news, tweets, photos that are presented to the user, starting at the most recent one. As there is simply too much content, all social networks introduced machine learning-based algorithms to determine which content to present to a certain user. They are continuously trained through interactions with the network, i.e., "liking" a post boosts the respective topic, checking the feed of a certain friend on the network boosts the connection to this friend. Some networks have even introduced more fine-grained sentiments that can be used in addition to the simple "like" (see, e.g., Facebook's reactions "love", "haha", "wow", "sad", "angry"). Through "likes" of topics, connections to friends and interactions with the site, the social network creates, and continuously updates, for every single user, an internal model of likes and interests. This model is used to select content to be presented on the timeline by only selecting content that is assessed as being relevant to the user's interests. Plus, algorithms typically favour content that is being "liked" or shared by those friends and connections that the user interacts with the most. This is the origin of the filter bubble phenomenon: users are only exposed to content that can also be described as “safe” – content shared by friends they know and like is considered content that matches a user’s interests. Controversial content that contradicts a user’s world view or that presents opposing information, that challenges their beliefs is *not* presented – according to the underlying user model it is not relevant.

In the digital age, we can no longer assume that everything that has been published is necessarily correct. While this has been true in some parts of the world for decades, this challenge has now also entered the Western part of the world. Since November 2016 it has been socially accepted, in some parts of the political spectrum, to categorise fact-checked articles, written by experienced journalists and published by respected news outlets, as “fake news” – not because the news are false but because the corresponding articles do not endorse and support the opinion and agenda of the reader. The age of post-factual politics creates an unprecedented tension and stimulates fundamental debates about the relationship between politics and the fourth estate in civil society and beyond.

Additionally, we are faced with the challenge that more and more content is produced and spread with the sole purpose of manipulating the readers’ beliefs and opinions by appealing to their emotions instead of informing them objectively. Rather, this type of opinionated, emotional, biased, often aggressive and far-right content is prepared and spread to reach specific goals, for example, to create support for controversial ideas or to destroy the reputation of a politician. These coordinated online marketing campaigns are often carried out by experts with in-depth knowledge of the underlying technologies and processes. They involve large numbers of bots and fake accounts as amplifiers Weedon et al. (2017) as well as large budgets for online advertisements in social media, clearly targeted at very specific demographic groups the originators want to influence and then to flip to reach a specific statistical threshold. The way news are nowadays spread, circulated, read and shared – with less and less critical thinking or fact checking – enables this type of content to gather a large number of readers (and sharers) quickly. The filter bubble acts like an echo chamber that can amplify any type of content, from genuine, factual news to emotionally charged, politically biased news, to false news to orchestrated disinformation campaigns, created with the specific purpose of large-scale manipulation. Content of the last two categories can be hard or very hard to identify even for human experts.

A key challenge is to separate objective, balanced content, be it journalistic or user-generated, from hateful, abusive or biased content, maybe produced with a hidden agenda. Even if fundamentally different in nature, nowadays both types of content share the same level of visibility, reach and exposure through the equalisation mechanisms of the social web, which can be easily manipulated. In the past the tasks of fact checking, critical thinking and unveiling hidden agendas have mostly been in the realm of journalism, but in the digital age they are more and more transferred to the actual reader and recipient of online content. The analysis, curation and assessment of content is no longer carried out by professional journalists or news editors – the burden of fact checking and content verification is left to the reader. This aspect is getting even more crucial because the number of people who state that social networks are their *only* source of news and information is growing steadily (Marchi 2012). The most prominent example from recent history is that social media manipulation can apparently even make or break a national election (Barthel et al. 2016; Rogers & Bromwich 2016; Marwick & Lewis 2017). It must be noted, though, that a large number of fact checking 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 are fragmented (Babakar & Moy 2016).

Several types of online content are often grouped together under the label "fake news". For example, Holan (2016) defines fake news as "made-up stuff, masterfully manipulated to look like credible journalistic reports that are easily spread online to large audiences willing to believe the fictions and spread the word." In reality, the situation is much more complex. Initially based on the classification suggested by Wardle (2017), Table 1 (overleaf) shows a first attempt at bringing together the different types of false news including selected characteristics and associated intentions. The table shows the complexity of the situation and that a more fine-grained terminology is needed to discuss the topic properly, especially when it comes to designing technological solutions that are meant to address one or more of these types of content.

An additional challenge is the proliferation of hateful comments and abusive language, often used in the comments and feedback sections on social media posts. The effects can be devastating for the affected individual. Many hateful comments on repeated postings by the same person, say, a pupil, are akin to cyberbullying and cybermobbing. There is also a clear tendency to aggressive comments on, for example, the social media pages of traditional news outlets, who have to ask the users more and more to behave in a civilised way.

Table 1: Characteristics and intentions associated with different types of false news (adapted from Wardle 2017; Walbrühl 2017; Rubin et al. 2015; Holan 2016; Weedon et al. 2017)

	Satire or parody	False connection	Misleading content	False context	Imposter content	Manipulated content	Fabricated content
Clickbait		X	X	?		?	?
Disinformation			X	X		X	X
Politically biased		?	X	?		?	X
Poor journalism		X	X	X			
To parody	X				?		X
To provoke					X	X	X
To profit	?	X			X		X
To deceive		X	X	X	X	X	X
To influence politics			X	X		X	X
To influence opinions			X	X	X	X	X

3. Technology framework: approach

Technically, online content is predominantly consumed through two possible channels, both of which rely substantially on World Wide Web technology and established web standards. Users either read and interact with content directly on the web (mobile or desktop versions of websites) or through dedicated mobile apps; this can be considered using the web implicitly as many apps make heavy use of HTML5 and other web technologies. The World Wide Web itself still is and, for the foreseeable future, will continue to be the main transport medium for online content. The suggested technology architecture is, hence, designed as an additional layer on top of the web. Nevertheless, we also have to be clear about the scope and ambition of the challenge: the infrastructure needs to be able to cope with millions of users, arbitrary content types, hundreds of languages and massive amounts of data. The goal is to empower and to enable users to balance out the network, echo chamber and filter bubble effects and to provide mechanisms to filter for abusive content.

3.1. Services of the infrastructure

In many cases the burden of analysing and fact checking online content has been shifted to the reader (section 2), which is why corresponding analysis and curation services need to be made available in an efficient and ubiquitous way. The same tools to be used by *content consumers* can and should also be applied by *content creators*, e.g., journalists and bloggers. Those readers who are interested to know more about what they are currently reading should be able to get the additional information as easily as possible, and the same applies to those journalists who are interested in fact-checking the content they are researching for the production of new content.

Readers of online content are users of the World Wide Web. They need, first and foremost, web-based tools and services with which they can process any type of content to get additional information on a specific piece, be it one small comment on a page, the main content component of a page (for example, an article) or even a set of interconnected pages (one article spread over multiple pages), for which an assessment is sought.

The provided services need to be designed to operate in and with the web stack of technologies, i.e., within the web ecosystem, they need to support users in their task of reading and curating content within the browser in a smarter and, eventually, more balanced way. This can be accomplished by providing additional, also alternative opinions and view points, by presenting other, indepedent assessments, or by indicating if content is dangerous, abusive, factual or problematic in any way. Fully automatic technologies (Rubin et al. 2015; Schmidt & Wiegand 2017; Horne & Adal 2017; Martinez-Alvarez 2017) can take over a subset of these responsibilities but, given the current state of the art, not all, which is why the approach needs to be based both on simple or complex automatic filters and watchdogs as well as human intelligence and feedback.[3]

The tools and services should be available to every web user without the need to install any additional third-party software. This is why these services, ideally, should be integrated into the browser on the same level as bookmarks, the URL field or the navigation bar, i.e., without relying on the installation of a plugin. The curation tools should be thought of as an inherent technology component of the World Wide Web, for which intuitive and globally acknowledged user-interface conventions can be established,

[3] A fully automatic solution would work only for a very limited set of cases. A purely human-based solution would work but required large amounts of experts and, hence, would not scale. This is why we favour, for now, a hybrid solution.

such as, for example, traffic light indicators for false news content (green: no issues found; yellow: medium issues found and referenced; red: very likely false news). Table 2 shows a first list of tools and services that could be embedded into such a system.[4] Some of these can be conceptualised and implemented as automatic tools (Horne & Adal 2017), while others need a hybrid approach that involves crowd-sourced data and opinions. In addition to displaying the output of these services, the browser interface needs to be able to gather, from the user, comments, feedback, opinions and sentiments on the current piece of content, further to feed the crowd-sourced data set. The user-generated data includes both user-generated annotations (UGA) and also user-generated metadata (UGM). Automatically generated metadata are considered machine-generated metadata (MGM).

Table 2: Suggested tools and services to be provided through the infrastructure (selection)

Tool or Service	Description	Approach
Political bias indicator	Indicates the political bias (Martinez-Alvarez 2017) of a piece of content, e.g., from far left to far right	automatic
Hate speech indicator	Indicates the level of hate speech a certain piece of content contains	automatic
Reputation indicator	Indicates the reputation, credibility (Martinez-Alvarez 2017), trustworthiness, quality (Filloux 2017) of a certain news outlet or individual author of content	crowd, automatic
Fact checker	Checks if claims are backed up by references, evidence, established scientific results and links claims to the respective evidence (Babakar & Moy 2016)	automatic
Fake news indicator	Indicates if a piece of content contains non-factual statements or dubious claims (Horne & Adal 2017; Martinez-Alvarez 2017)	crowd, automatic
Opinion inspector	Inspect opinions and sentiments that other users have with regard to this content (or topic) – not just the users commenting on one specific site, but all of them	crowd, automatic

[4] This list is meant to be indicative rather than complete. For example, services for getting background information on images are not included (Gupta et al. 2013). Such tools could help pointing out image manipulations or that an old image was used, out of context, to illustrate a new piece of news.

3.2. Characteristics of the infrastructure

In order for these tools and services to work effectively, efficiently and reliably, they need to possess several key characteristics, which are quintessential for the overall success of the approach.

Like the Internet and the World Wide Web, the infrastructure must be operated in a federated, i.e., de-centralised setup – a centralised approach would be too vulnerable for attacks or misuse. Multiple organisations, companies, research centres or NGOs should be able to set up, operate and offer services (section 3.1) and additional pieces of the infrastructure. The internal design of the respective algorithms and tools may differ substantially, but their output (MGM) should comply to a standardised metadata format. It is rather likely that political biases in different models meant to serve the same purpose cannot be avoided, which is especially likely for models based on large amounts of data, which, in turn, may inherently include a political bias. This is why users must be enabled to activate or deactivate as many of these tools as they want to get an aggregated value, for example, with regard to the level of hate speech in content or its political bias. Services and tools must be combinable, i.e., they need to comply to standardised input and output formats (Babakar & Moy 2016). They also need to be transparent (Martinez-Alvarez 2017). Only transparent, i.e., fully documented, checked, ideally also audited approaches can be trustworthy.

Access to the infrastructure should be universal and available everywhere, i.e., in any browser, which essentially means that, ideally, the infrastructure should be embedded into the technical architecture of the World Wide Web. As a consequence, access mechanisms should be available in every browser, on every platform, as native elements of the graphical user interface (GUI). These functions should be designed in such a way that they support users without distracting them from the content. Only if these tools are available virtually anywhere, can the required scale be reached.

The user should be able to configure and to combine multiple services, operated in a de-centralised way, for a clearly defined purpose in order to get an aggregated value. There is a danger that this approach could result in a replication and shift of the filter bubble effect (section 2) onto a different level but users would at least be empowered actively to configure their own personal set of filters to escape from any resulting bubble. The same transparency criterion also applies to the algorithm that aggregates multiple values, of course.

3.3. Building blocks of the infrastructure

Research in Language Technology and Natural Language Processing (NLP) currently concentrates on smaller components, especially watchdogs, filters and classifiers (see section 4) that could be applied under the umbrella of a larger architecture to tackle current online media phenomena (section 2). While this research is both important and crucial, even if fragmented and somewhat constrained by the respective training data sets (Rubin et al. 2015; Conroy et al. 2015; Schmidt & Wiegand 2017) and limited use cases, we also need to come to a shared understanding how these components can be deployed and made available. The suggestion consists of the following building blocks (see Figure 1).

3.3.1. Building block: natively embedded into the World Wide Web

An approach that is able to address modern online media and communication phenomena adequately needs to operate on a web-scale level. It should natively support cross-lingual processing and be technically and conceptually embedded into the architecture of the World Wide Web itself. It should be standardised, endorsed and supported not only by all browser vendors but also by all content and media providers, especially the big social networks and content hubs. Only if *all* users have *immediate* access to the tools and services suggested in this proposal can they reach its full potential. The services must be unobtrusive and cooperative, possess intuitive usability, their recommendations and warnings must be immediately understandable, and it must be simple to provide general feedback (UGM) and assessments on specific pieces of content (UGA).

3.3.2. Building block: web annotations

Several pieces of the proposed infrastructure are already in place. One key component are Web Annotations, standardised by the World Wide Web Consortium (W3C) in early 2017 (Sanderson 2017; Sanderson et al. 2017a;b). They enable users to annotate arbitrary pieces of web content, essentially creating an additional and independent layer on top of the regular web. Already now Web Annotations are used for multiple individual projects in research, education, scholarly publishing, administration and investigative journalism.[5] Web Annotations are *the* natural mechanism to

[5] See, for example, the projects presented at I Annotate 2015 (http://iannotate.org/2015/), 2016 (http://iannotate.org/2016/) and 2017 (http://iannotate.org/2017/).

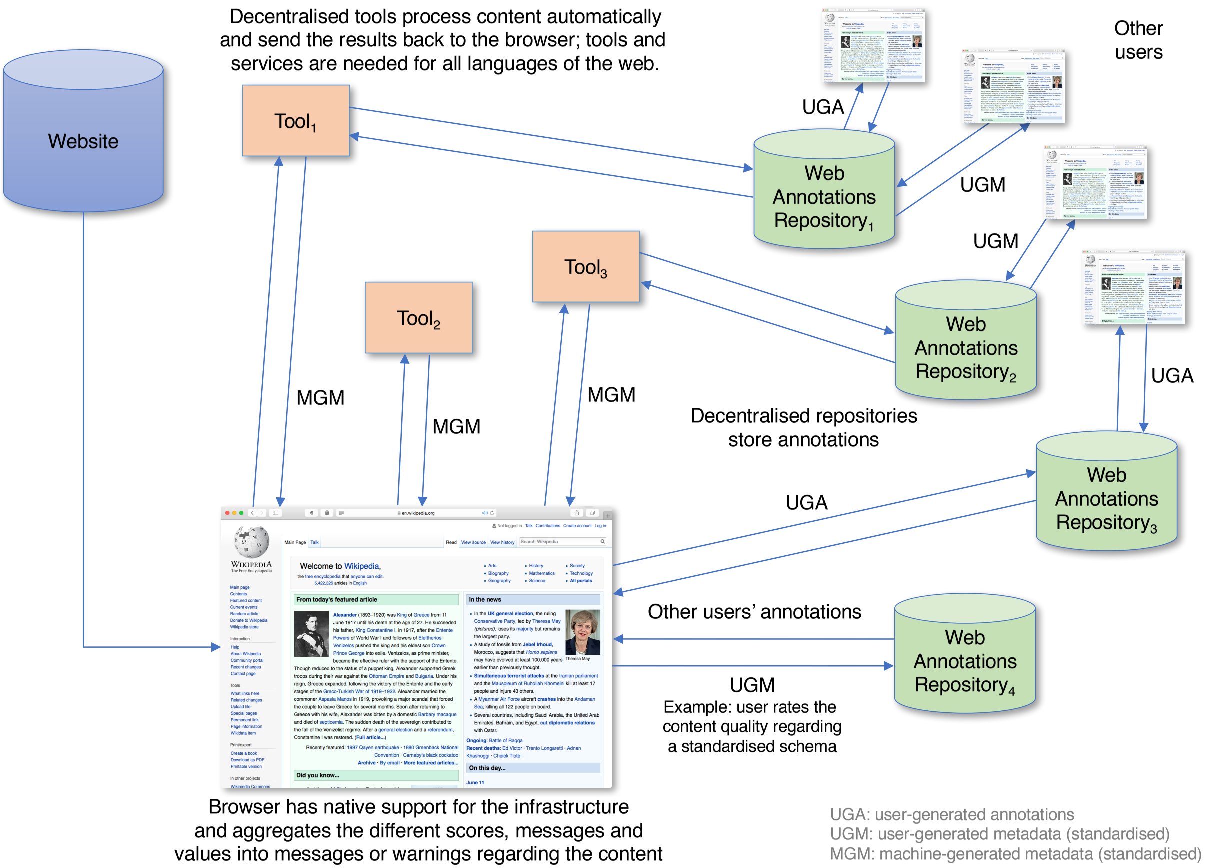
Decentralised tools process content automatically and send the results back to the browser; tools and services are needed for all languages of the web.
Other users
Website
Tool$_1$
Tool$_2$
Tool$_3$
UGA
UGM
UGM
UGA
Web Annotations Repository$_1$
Web Annotations Repository$_2$
Web Annotations Repository$_3$
Web Annotations Repository$_4$
MGM
MGM
MGM
Decentralised repositories store annotations
UGA
Other users' annotations
UGM
Example: user rates the content quality regarding a standardised schema
Browser has native support for the infrastructure and aggregates the different scores, messages and values into messages or warnings regarding the content
UGA: user-generated annotations
UGM: user-generated metadata (standardised)
MGM: machine-generated metadata (standardised)

Figure 1: Simplified architecture of the proposed infrastructure

enable users and readers interactively to work with content, to include feedback and assessments, to ask the author or their peers for references or to provide criticism. The natural language content of Web Annotations (UGA) can be automatically mined using methods such as sentiment analysis or opinion mining – in order to accomplish this across multiple languages, this needs to be done cross-lingually (Rehm et al. 2016). However, there are still limitations. Content providers need to enable Web Annotations by referencing a corresponding JavaScript library. Federated sets of annotation stores or repositories are not yet foreseen, neither are native controls in the browser that provide aggregated feedback, based on automatic (MGM) or manual content assessments (UGM, UGA). Another barrier for the widespread use and adoption of Web Annotations are proprietary commenting systems, as used by all major social networks. Nevertheless, services such as Hypothes.is enable Web Annotations on any web page, but native browser support, ideally across all platforms, is still lacking. A corresponding browser feature needs to enable both free-text annotations of arbitrary content pieces (UGA), but also very simple flagging of problematic content, for example, "content pretends to be factual but is of dubious quality" (UGM). Multiple UGA, UGM or MGM annotations could be aggregated and presented to new readers of the content to provide guidance and indicate any issues.

3.3.3. Building block: metadata standards

Another needed piece of the architecture is an agreed upon metadata schema Babakar & Moy (2016) to be used both in manual annotation scenarios (UGM) and also by automatic tools (MGM). Its complexity should be as little as possible so that key characteristics of a piece of content can be adequately captured and described either by humans or machines. With regard to this requirement, W3C published several standards to represent the provenance of digital objects (Groth & Moreau 2013; Belhajjame et al. 2013a). These can be thought of as descriptions of the entities or activities involved in producing or delivering a piece of content to understand how data was collected, to determine ownership and rights or to make judgements about information to determine whether to trust content (Belhajjame et al. 2013b). An alternative approach is for content publishers to use Schema.org's ClaimReview[6] markup in their websites after specific facts have been checked. The needed metadata schema can be based on the W3C provenance ontology and/or Schema.org. Additional metadata fields are likely to be needed.

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

3.3.4. Building block: tools and services

Web Annotations can be used by readers of online content to provide comments or to include the results of researched facts (UGA, UGM). Automatic tools and services that act as filters and watchdogs can make use of the same mechanisms (MGM, see section 3.1). These could be functionally limited classifiers, for example, regarding abusive language, or sophisticated natural language understanding (NLU) components that attempt to check certain statements against one or more knowledge bases. Regardless of the complexity and approach, the results can be made available as globally accessible Web Annotations (that can even, in turn, be annotated themselves). Services and tools need to operate in a decentralised way, i.e., users must be able to choose from a wide variety of automatic helpers. These could, for example, support users to position content on the political spectrum, either based on crowd-sourced annotations, automatic tools, or both.

3.3.5. Building block: decentralised repositories and tools

The setup of the infrastructure must be federated and decentralised to prevent abuse by political or industrial forces. Data, especially annotations, must be stored in decentral repositories, from which browsers retrieve, through secure connections, data to be aggregated and displayed (UGM, UGA, MGM, i. e., annotations, opinions, automatic processing results etc.). In the medium to long term, in addition to annotations, repositories will also include more complex data, information and knowledge that tools and services will make use of, for example, for fact checking. In parallel to the initiative introduced in this article, crowd-sourced knowledge graphs such as Wikidata or DBpedia will continue to grow. The same is true for semantic databases such as BabelNet and many other data sets, usually available and linkable as Linked Open Data. Already now we can foresee more sophisticated methods of validating and fact-checking arbitrary pieces of content using systems that make heavy use of knowledge graphs, for example, through automatic entity recognition and linking, relation extraction, event extraction and mapping etc. One of the key knowledge bases missing, in that regard, is a Web Annotation-friendly event-centric knowledge graph, against which fact-checking algorithms can operate.[7] Basing algorithms that are supposed to determine the truth of a statement on automatically extracted and formally represented knowledge creates both

[7] GDELT (Global Database of Events, Language, and Tone) comes close but is lacking with regard to its integratability, see http://www.gdeltproject.org.

practical and philosophical questions, among others, who checks these automatically extracted knowledge structures for correctness? How do we represent conflicting view points and how do algorithms handle conflicting view points when determining the validity of a statement? How do we keep the balance between multiple subjective opinions and an objective and scientific ground-truth?

3.3.6. Building block: aggregation of manual and automatic annotations

The final key building block of the proposed system relates to the aggregation of manual and automatic annotations, created in a de-centralised and highly distributed way by human users and automatic services (UGA, UGM, MGM). Already now we can foresee very large numbers of annotations so that the aggregation and consolidation will be a non-trivial challenge. This is also true for those human annotations that are not based on shared metadata vocabularies but that are free text – for these free and flexible annotations, robust and also multilingual annotation mining methods need to be developed.

4. Related work

Research on Computer-Mediated Communication (CMC) has a long tradition. Scholars initially concentrated on different types of novel communication media such as e-mail, IRC, Usenet newsgroups, and different hypertext systems and document types, especially personal home pages, guestbooks and, later, discussion fora. Early on, researchers focused upon the (obvious) differences between these new forms of digital communication and the traditional forms, especially when it comes to linguistic phenomena that can be observed on the text surface (smileys, emoticons, acronyms etc.). Several authors pointed out that the different forms of CMC have a certain oral and spoken style, quality and conceptualisation to them, as if produced spontaneously in a casual conversation, while being realised in a written medium (Haase et al. 1997).

If we now fast forward to 2017, a vastly different picture emerges. About half of the global population has access to the internet, most of whom also use the World Wide Web and big social networks. The internet is no longer considered fringe technology that is only used by scientists, early adopters and computer nerds, but it is mainstream. Nowadays the internet acts like an amplifier and enabler of social trends. It continues to penetrate and to disrupt our lives and social structures, especially our established traditions of social and political debates. The relevance of online media, online news and online communication could not be any more crucial. While early analyses of CMC, e.g., Reid (1991), observed that the participants were involved in the "deconstruction of boundaries" and the "construction of social communities", today the exact opposite seems to be case: not only online but also offline can we observe the (disturbing) trend of increased, intricately orchestrated, social and political manipulation, nationalism and the exclusion of foreigners, immigrants and seemingly arbitrary minorities – boundaries are constructed, social communities deconstructed, people are manipulated, individuals excluded.

There is a vast body of research on the processing of online content including text analytics (sentiment analysis, opinion and argument mining), information access (summarisation, machine translation) and document filtering (spam classification). Attempting to classify, among others, the different types of false news shown in Table 1 requires, as several researchers also emphasise, a multi-faceted approach that includes multiple different processing steps. We have to be aware of the ambition, though, as some of the "fake news detection" use case scenarios are better described as "propaganda detection", "disinformation detection", maybe also "satire detection". These are difficult tasks at which even humans often fail. Current research in this area is fragmented and concentrates on very specific sub-problems, see, for example, the Fake News Challenge, the Abusive Language Workshop, or the Clickbait Challenge.[8] What is missing, however, is a practical umbrella that pulls the different pieces together and that provides an approach that can be realistically implemented and deployed including automatic tools as well as human annotations.

5. Summary and conclusions

Humanity is transitioning into becoming a digital society, or at least a "digital first" society, i.e., news, media, facts, rumours (Zubiaga et al. 2016; Srivastava et al. 2017), information are created, circulated and disseminated online. Already now the right social media strategy can make or break an election or is able to influence if a smaller or larger societal or demographic group (city, region, country, continent) is in favour or against constructively solving a certain societal challenge. Social media and online communication can be extremely powerful tools to bridge barriers, to in-

[8] See http://www.fakenewschallenge.org, http://www.clickbait-challenge.org, https://sites.google.com/site/abusivelanguageworkshop2017/.

form people and to enable global communication. When abused, misused or infiltrated, they are a dangerous weapon.

The fields of Computational Linguistics, Language Technology and Artificial Intelligence should actively contribute solutions to this key challenge of the digital age. If we don't, there is a concrete danger that stakeholders with bad intentions are able to influence parts of the society to their liking, only constrained by their political, commercial, egotistical interests. Technologies need to be developed to enable every user of online media to break out of their filter bubbles and to inform themselves in a balanced way, taking all view points into account.

After dumb digital content, smart content and semantic content enrichment we now need to concentrate on content curation tools that enable *contextualised content*, i.e., content that can be, ideally, automatically cross-referenced and fact-checked, and for which additional background information can be retrieved in a robust way. This can involve assessing the validity of claims and statements made in the content as well as retrieving related texts, facts and statements, both in favour and against a certain piece of content.

Next steps include presenting this proposal in various different fora and communities, among others, researchers and technologists, standards-developing organisations (Babakar & Moy 2016) and national as well as international political bodies. At the same time, research needs to be continued and prototypes of the architecture as well as individual services developed, enabling organisations to build and to deploy decentralised tools early. While a universal, globally accessible, balanced and well maintained knowledge graph containing up-to-date information about entities and events would be handy to have, it is out of scope with regard to the initiative reported here; it is safe to assume that such a knowledge repository will be developed in parallel in the next couple of years. The proposed architecture can be used to link online content against this knowledge graph and to measure the directions of online debates.

The proposal introduced in this article is ambitious in its scope and implications, prevention of misuse will play a hugely important role. How can we make sure that a certain piece of technology is only used with good intentions? Recently it has been shown that a user's social media data can reliably predict if the user is suffering from alcohol or drug abuse (Ding et al. 2017). Will this technology be used to help people or to stigmatise them? Will an infrastructure, as briefly sketched in this paper, be used to empower users to make up their own minds by providing additional information about online content or will it be used to spy on them and to manipulate them with commercial or political intentions?

Acknowledgements

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.

References

Babakar, M. and W. Moy. 2016. The state of automated factchecking – How to make factchecking dramatically more effective with technology we have now. https://tinyurl.com/yd6cdoo5

Barthel, M., A. Mitchell and J. Holcomb. 2016. Many Americans believe fake news is sowing confusion. http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

Bazzaz, D. 2016. News you can use: Infographic walks you through 10 questions to detect fake news. http://www.seattletimes.com/education-lab/infographic-walks-students-through-10-questions-to-help-them-spot-fake-news/

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

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

Bourgonje, P., J. Moreno-Schneider, J. Nehring, G. Rehm, F. Sasaki and A. Srivastava. 2016a. Towards a platform for curation technologies: Enriching text collections with a semantic-web layer. In H. Sack, G. Rizzo, N. Steinmetz, D. Mladeni, S. Auer and C. Lange (eds.) The semantic web. Berlin & New York: Springer, no. 9989 in Lecture Notes in Computer Science. 65–68.

Bourgonje, P., J. M. Schneider, G. Rehm and F. Sasaki. 2016b. Processing document collections to automatically extract linked data: Semantic storytelling technologies for smart curation workflows. In A. Gangemi and C. Gardent (eds.) Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016). Edinburgh: The Association for Computational Linguistics, 13–16.

Chan, R. 2017. Artificial intelligence is going to destroy fake news – But A.I. can also cause the volume of fake news to explode. https://www.inverse.com/article/27723-artificial-intelligence-will-destroy-fake-news

Conroy, N. J., V. L. Rubin and Y. Che. 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology 52. 1–4.

Dale, R. 2017. NLP in a post-truth world. Natural Language Engineering 23. 319–324.

Ding, T., W. K. Bickel and S. Pan. 2017. Social media-based substance use prediction. https://arxiv.org/abs/1705.05633

Filloux, F. 2017. Quality for news is mostly about solving the reputation issue. https://mondaynote.com/quality-for-news-is-mostly-about-solving-the-reputation-issue-fdebd0dcc9e2

Gershgorn, D. 2016. In the fight against fake news, artificial intelligence is waging a battle it cannot win. https://qz.com/843110/can-artificial-intelligence-solve-facebooks-fake-news-problem/

Groth, P. and L. Moreau. 2013. PROV-Overview: An overview of the PROV family of documents. Tech. rep., World Wide Web Consortium (W3C). https://www.w3.org/TR/prov-overview/

Gupta, A., H. Lamba, P. Kumaraguru and A. Joshi. 2013. Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. In Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, 729–736.

Haase, M., M. Huber, A. Krumeich and G. Rehm. 1997. Internetkommunikation und Sprachwandel. In R. Weingarten (ed.) Sprachwandel durch Computer. Opladen: Westdeutscher Verlag. 51–85.

Holan, A. D. 2016. 2016 Lie of the year: Fake news. http://www.politifact.com/truth-o-meter/article/2016/dec/13/2016-lie-year-fake-news/

Horne, B. D. and S. Adal. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Proceedings of the 2nd International Workshop on News and Public Opinion at ICWSM.

Mantzarlis, A. 2015. 6 tips to debunk fake news stories by yourself. http://www.poynter.org/2015/6-tips-to-debunk-fake-news-stories-by-yourself/385625/

Mantzarlis, A. 2017. There are now 114 fact-checking initiatives in 47 countries. https://www.poynter.org/2017/there-are-now-114-fact-checking-initiatives-in-47-countries/450477/

Marchi, R. 2012. With Facebook, blogs, and fake news, teens reject journalistic "objectivity". Journal of Communication Inquiry 36. 246–262.

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

Marwick, A. and R. Lewis. 2017. Media manipulation and disinformation online. https://datasociety.net/output/media-manipulation-and-disinfo-online/

Metz, C. 2016. The bittersweet sweepstakes to build an AI that destroys fake news. http://www.wired.com/2016/12/bittersweet-sweepstakes-build-ai-destroys-fake-news/

Rehm, G. and T. Declerck (eds.). 2017. Sprachtechnologien für die Herausforderungen des digitalen Zeitalters – Language Technologies for the Challenges of the Digital Age: Proceedings of the GSCL Conference 2017. Berlin: Gesellschaft für Sprachtechnologie und Computerlinguistik. http://gscl2017.dfki.de

Rehm, G. and F. Sasaki. 2015. Digitale Kuratierungstechnologien – Verfahren für die effiziente Verarbeitung, Erstellung und Verteilung qualitativ hochwertiger Medieninhalte. In Proceedings of the 2015 international conference of the german society for computational linguistics and language technology. GSCL '15, 138–139.

Rehm, G., F. Sasaki and A. Burchardt. 2016. Web Annotations – A game changer for language technologies? Presentation given at I Annotate 2016, Berlin. http://www.slideshare.net/georgrehm/web-annotations-a-game-changer-for-language-technology; http://iannotate.org/2016/

Rehm, G., J. M. Schneider, P. Bourgonje, A. Srivastava, J. Nehring, A. Berger, L. König, S. Räuchle and J. Gerth. 2017. Event detection and semantic storytelling: Generating a travelogue from a large collection of personal letters. In T. Caselli, B. Miller, M. van Erp, P. Vossen, M. Palmer, E. Hovy and T. Mitamura (eds.) Proceedings of the events and stories in the news workshop. Vancouver, Canada: Association for Computational Linguistics.

Reid, E. M. 1991. Electropolis: Communication and community on Internet Relay Chat. Honours Thesis. University of Melbourne. http://www.aluluei.com/electropolis.htm

Rogers, K. and J. E. Bromwich. 2016. The hoaxes, fake news and misinformation we saw on election day. https://www.nytimes.com/2016/11/09/us/politics/debunk-fake-news-election-day.html

Rubin, V. L., Y. Chen and N. J. Conroy. 2015. Deception detection for news: Three types of fakes. Proceedings of the Association for Information Science and Technology 52. 1–4.

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

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

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

Schmidt, A. and M. Wiegand. 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. Valencia.

Srivastava, A., G. Rehm and J. M. Schneider. 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). Vancouver: Association for Computational Linguistics, 477–481.

Walbrühl, D. 2017. Das musst du wissen, um Fake News zu verstehen. https://perspective-daily.de/article/213/AhopoOEF

Wardle, C. 2017. Fake news. It's complicated. https://firstdraftnews.com/fake-news-complicated/

Weedon, J., W. Nuland and A. Stamos. 2017. Information operations and Facebook. https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

Zubiaga, A., M. Liakata, R. Procter, G. W. S. Hoi and P. Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 11. https://doi.org/10.1371/journal.pone.0150989