Designing Social Machines for Tackling Online Disinformation

Antonia Wild, Andrei Ciortea, Simon Mayer
antonia.wild@student.unisg.ch
{andrei.ciortea, simon.mayer}@unisg.ch
Institute of Computer Science, University of St.Gallen

ABSTRACT

Traditional news outlets as carriers and distributors of information have been challenged by online social networks with regards to their gate-keeping function. Users are now left with the difficult task of assessing the credibility of information provided to them, which facilitates the spread of disinformation. At the same time, current human- and machine-based approaches to tackle disinformation are operating in isolation from one another, each with its respective weaknesses. We believe that only a combined effort of people and machines will be able to curb so-called "fake news" at scale in a decentralized Web. In this paper, we propose an approach to design social machines that coordinate human- and machine-driven credibility assessment of information on a decentralized Web. To this end, we defined a fact-checking process that draws upon ongoing efforts for tackling disinformation on the Web, and we formalized this process as a multi-agent organisation for curating W3C Web Annotations. We present the current state of our prototypical implementation in the form of a browser plugin that builds on the Hypothesis annotation platform and the JaCaMo multi-agent platform. Our social machines can span across the Web to enable collaboration in form of public discourse, thereby increasing the transparency and accountability of information on the Web.

KEYWORDS

Disinformation, Social Machines, Web Architecture, Linked Data, Multi-Agent Systems, Decentralization

ACM Reference Format:

1 INTRODUCTION

Online social networks, as carriers and distributors of information, have challenged traditional news outlets in their gate-keeping function [11]. This allows for a more diverse, immediate and, unfiltered access to information, but at the same time leaves users with the difficult task of assessing the credibility of provided information, which eases the spread of disinformation. Already in his talk at WWW'94, Sir Tim Berners-Lee was raising the challenge of ensuring the quality of information in a system as open as the Web – and envisioned the use of annotations as a suitable mechanism. Such Web Annotations, now a W3C Recommendation [1], construct a metadata layer on top of existing resources and without requiring their modification. As such, they can be regarded as a connective fabric that allows users to address statements within Web pages and to bind these statements to metadata – in the domain of online disinformation, for instance, regarding their provenance, truth assessments by others, or opposing views from other sources.

Academics and practitioners have tested human- and machine-based approaches that make use of annotations to tackle online disinformation (see Section 2.1.2). However, to-date no solution seems sufficient to efficiently and reliably address the issue. We believe that only a combined effort of both humans and machines will be able to curb so-called "fake news" at Web scale. We refer to such collaborative efforts of humans and machines as social machines, following the definition of [16]. We pursue the following research question: how to design social machines that coordinate humans and autonomous agents to perform a transparent and guided credibility analysis of information on the open, decentralized Web – at Web scale and in a timely manner?

In this paper, we take a first step towards addressing this question. We defined a fact-checking process that draws upon existing efforts of the W3C Credible Web Community Group7 and EUFACTCHECK8. Then, building upon results from research on autonomous agent and multi-agent systems (MAS) [19], we conceptualized this fact-checking process as a multi-agent organization for curating W3C Web Annotations. To validate our approach, we are currently in the process of implementing a prototypical social machine using Hypothesis9, an annotation platform that conforms to the W3C Web Annotation recommendations, and JaCaMo [5], a platform for the development of MAS.

The rest of this paper is structured as follows. Section 2 discusses related work on tackling online disinformation, social machines and...
We give an overview of existing fact-checking approaches in Section 2.1. We argue that human- and machine-driven fact-checking efforts by themselves are not sufficient to cope with disinformation at scale and with accuracy. Consequently, in Section 2.2 we present the concepts of social machines and MAS that allow to design and implement hybrid fact-checking campaigns.

2.1 Efforts directed towards combating disinformation online

To help users assess the credibility of information at the point of consumption, manifold human- and machine-driven approaches have been developed and tested in both academia and practice. Efforts to analyze the content and its metadata can be mapped according to the dimensions of automation – i.e., to which degree software drives the credibility assessment, and abstraction – i.e., to which degree the analysis abstracts from the content itself and considers external information provided in an open-world scenario.

2.1.1 Human-driven Fact-checking. At a low degree of automation, organizations and groups of experts (notably journalists) have formed manual fact-checking initiatives. While manual fact-checking enables a high degree of abstraction and to a certain extent ensures accountability (through disclosure of the fact-checking organization) and transparency (through disclosure of the fact-checking process), it however cannot scale to the size of the Web: the Washington Post alone publishes on average 500 stories and videos per day. As a consequence, the majority of fact-checking sites has limited coverage regarding topics, languages, and geography. Finally, manual fact-checking exhibits a slower response time; ideally, however, claims would be verified already before going viral.

2.1.2 Automated Fact-checking. Fact-checking approaches with a higher automation degree mostly rely on natural language processing (NLP) to extract, retrieve, and analyze information [3]. Based on the processed text corpora, they then evaluate the truthfulness of the content either by using information that is contained within the content itself and its immediate metadata – i.e., adopting a closed-world approach – or by additionally incorporating outside knowledge [17]. Closed-world approaches are however prone to deception, since, for instance, semantically correct linguistics do not guarantee that an expressed claim is true [17]. To counteract this, automated open-world approaches work in tandem with (semi-)structured databases.

Despite the progress, automated fact-checking however still lacks accuracy, and the underlying models are often trained with topic-, language- or culture-specific content [7], which undermines the adaptability of these systems. Additionally, current solutions are not yet able to grasp the intent of the text, e.g., irony or metaphors [18], and are vulnerable to bias [15]. Lastly, currently automated approaches do not provide an end-to-end solution for reliably determining the truthfulness of an article as they often focus on one aspect of credibility (e.g., the language style or the publisher's credibility) and thus demand for integration within a larger system that assesses credibility along all its dimensions.

In summary, the manual checking of online information cannot be performed in a timely manner (but is mostly accountable and transparent), while automated approaches lack the facilities to reliably detect disinformation (but can potentially work at scale). Systems encompassing collaborative efforts of human and autonomous agents might be able to unify the desirable properties of each of these approaches. First ideas to create such systems have been envisioned in [14], which presents an ontology for a hybrid fact check infrastructure that could be embedded into the Web, based on the W3C Web Annotation Recommendation.

2.2 Social Machines on the Web

Social machines could enable such large-scale collaborations among humans and autonomous agents on the Web. Over the past couple of years, the World Wide Web Consortium (W3C) – and in particular the W3C Web Annotation Working Group and the W3C Social Web Working Group – finalized standards that weave social features into the very fabric of the Web. Particularly, the W3C Annotation Protocol by the W3C Web Annotation Working Group builds upon the Linked Data Platform which is part of SOLID, thus strengthening the decentralized features of the Web. These standards unlock new opportunities for the development and deployment of social machines on the Web.

Some researchers have identified Multi-Agent Systems (MAS) as a suitable means to conceptualize and engineer social machines [6, 12, 13]. In distributed artificial intelligence, MAS are systems conceptualized in terms of (human and non-human) agents that are situated in a shared environment and interact with one another to achieve their objectives [10, 19]. In this context, an autonomous agent is commonly defined as "a computer system, situated in some environment, that is capable of flexible autonomous action in order to meet its design objectives" [10]. Autonomy is central to this definition and refers to the agent’s ability to operate on its own, without the need of direct intervention from humans or other agents.

Drawing upon MAS research, in [12] the authors propose to design social machines for crowd-sourcing software development by complementing the social compute unit – a model for ad-hoc human worker teams – with interaction protocols expressed in lightweight social calculus. Interaction protocols are also central to [6]: the authors define social protocols as behavioral standards based on the expectations of (human and non-human) participants towards interactions, which are then reflected in a computational model and can be enacted in a decentralized manner. Decentrality was also identified in [13] as one of the three key challenges for making social machines easily implementable on the Web – such that many people, rather than only a handful of powerful platforms, can benefit from them. This early research presents encouraging...
results, but the use of MAS to conceptualize social machines is still insufficiently investigated.

3 CONTRIBUTION

We introduce an approach that uses multi-agent organizations to conceptualize social machines for tackling online disinformation. Our objective is to tackle disinformation in a decentralized Web at scale and in a transparent manner. At the same time, we aim to avoid undue censorship and help agents make informed decisions, allowing the system to adapt to changes (in the underlying user community and information). Finally, we maintain openness in terms of the information being examined and the participating agents to benefit from new (technological) developments.

To this end, we designed a multi-faceted fact-checking process that considers several credibility indicators (see Section 3.1). Based thereon, we formalized this process as a multi-agent organization that coordinates the individual fact-checking efforts of humans and autonomous agents (see Section 3.2). To separate domain knowledge from operational processes – and to support the re-usability of our concepts and approach – we defined the Disinformation Tackler Ontology (see 3.3, which extends the W3C Web Annotation Vocabulary [2].

3.1 Disinformation Tackling

In our system, the tackling of disinformation consists of 47 distinct questions that are assigned to six processes of varying length (see Figure 1). Each process focuses on a different dimension of evaluating the credibility of an online information resource. These processes guide users through content evaluation along the four credibility indicators defined by the W3C Credible Web Community Group (CredWeb CG): inspection – assessing the content itself, corroboration – identifying claims and checking for verification by outer sources, reputation – assessing the credibility of the information provider, and lastly transparency – evaluating the self-declaration of publisher and author [21]. To each dimension we assigned corresponding elements of the disinformation tackling process of EUIFACTCHECK. While EUIFACTCHECK served as a starting point for our fact-checking process, we abstracted from its domain-specificity and reliance on experts by aligning with the efforts of the W3C CredWeb CG. Users at different instances of the process are required to provide evidence for their answers, for instance, by highlighting emotionalism within the text, linking to opposing sources, or tagging the central claim of the article.

3.2 Organizational Model

As the dissemination and forms of online information evolve, so do the mechanisms of disinformation producers and the tactics of malevolent system users. To cope with these dynamics, our social machine is expected to change, thus needing to be adaptive. Further, the system should be open with regards to participant agents and the integration of technical solutions. Lastly, we require some sort of normative power within our system to combat malevolent influences and to foster self-regulation in the system. Thus, we are in need for a formal representation of our organization geared towards tackling disinformation. Having a formal representation, we follow Hendler and Berners-Lee, who demand for technology...
Figure 3: Main concepts of DisinformationTackler ontology

the agents’ behavior [9]. The concept of roles allows for decoupling
the fact-checking process from individual persons and autonomous
agents, who can thus freely enter or leave these (open) organiza-
tions. Furthermore, this allows new goal coordination schemes to be
deployed at run time, which enables the dynamic reconfiguration
of the fact-checking process. Such reconfiguration could be deployed,
for instance, with regards to changes in the underlying content
(e.g., with regards to the language the information is displayed in).

3.2.2 A MOISE Organization for Tackling Disinformation. We used
MOISE to formalize our fact-checking process described in Sec-
tion 3.1 as a multi-agent organization. MOISE organizations are
defined declaratively, and we represent the specification of our
organization in RDF using the MOISE ontology defined in [20].
MOISE organizations can be instantiated on a per-needs basis to
meet the needs of tackling disinformation at scale in a decentral-
ized Web. Both people and autonomous agents can reason on their
specifications (e.g., to decide to join a fact-checking organization,
what goals to achieve, etc.).

The functional dimension in Figure 2 is composed of six schemes
which correspond to the six processes of our fact-checking pro-
cess, where all sub-goals can be achieved in parallel. Our structural
dimension highlights the different groups that agents within the
system can join. Each group in our organization can contain one of
two roles, the active user role (part of active user group), the
admin role (part of system functionality group), or the control
role (part of control group). The active user role encompasses our (human or autonomous agent) fact checkers. With
regards to the normative dimension, the active user role is permitted
to perform fact-checking tasks, such as the evaluation of content.

3.3 Vocabulary

To enable unification and re-usability of our domain concepts, we
conceptualize the Disinformation Tackler ontology. Figure 3 depicts
the core concepts of our ontology.

3.3.1 Weaving W3C Web Annotations into the Web. We rely on
annotations as our method of choice for Web-based credibility as-
essment. Since these annotations are independent of the informa-
tion being annotated and have their own resource identifiers and
representations, an entity that owns an information resource does
not own the dialogue about it. This is crucial for the independent
assessment of resources. Standardization efforts of the W3C Web
Annotation Working Group allow annotations to be weaved into the
Web in the future.

3.3.2 The Disinformation Tackler Ontology. We defined Disinfor-
mation Tackler (cf. Figure 3), an OWL ontology that allows the
uniform representation of annotations following from our fact-
checking process. Our ontology is aligned with the W3C Web An-
notation Vocabulary [2], which allows for inclusion of information
about annotations on a more technical level (e.g., talking about
different information content types such as audio or video). We
anchor our ontology via the concepts of oa:Annotation12 and
oa:Motivation13. The concept of oa:Annotation is central to our
ontology. Annotations are created by dt:Agents and motivated by
different oa:Motivations - e.g., a false claim a user has spotted
in the text. The concept of Agents subsumes both human and
non-human autonomous agents, treating both as first-class citizens in
our process. A dt:Agent can take on different dt:Roles, such as
dt:ActiveUser, which can dt:annotate a dt:InformationResource,
for instance through dt:highlights and dt:tags.

The Disinformation Tackler ontology also contains extension
points for other ontologies. The dt:Agent class is aligned with
similar concepts from the FOAF and PROV-O ontologies14 to
include network effects and provenance information into our system
in the future. This is of importance, for instance, when assessing
whether a user is a bot, based on previous Web-activity or his net-
work. Lastly, dt:Role serves as an anchor point for the MOISE
ontology15.

4 PROTOTYPICAL IMPLEMENTATION

In the following, we present the current state of our prototypical
implementation of our social fact-checking machines. Section 4.1
presents an architectural overview of our system. We discuss the
limitations of our prototype in Section 4.2.

4.1 Architectural Overview

The current progress with our prototypical implementation demon-
strates the feasibility of our approach. We have designed and im-
plemented a front end that allows humans to self-assess the cred-
ibility of online information. The resulting annotations, in turn,
can be discovered by other agents. Additionally, the implications
of the annotations are clustered and aggregated to give humans a
summarized overview of the evaluation results. A MOISE-based
intermediary keeps track of the organization’s status – tracking
level of goal fulfillment, committed agents, etc. – while our back
end processes and stores annotations.

To evaluate and annotate online information in conformance
with the workflow process defined in Section 3.1, we developed
a browser plugin based on Hypothesis. This browser plugin al-

12https://www.w3.org/TR/annotation-vocab/annotation
13https://www.w3.org/TR/annotation-vocab/motivation
14http://xmlns.com/foaf/spec/, https://www.w3.org/TR/prov-o/
15https://www.emse.fr/~zimmermann/Ontologies/moise.owl
We designed the intermediary component using JaCaMo—a platform for the development of MAS that integrates the MOISE organizational platform—to define and deploy multi-agent organizations. In our MOISE organizations, people are proxied by simple software agents: the agents enact roles in the organizations on behalf of humans in order to signal the achievement of goals, and they forward to humans all their permissions and obligations within the organization (e.g., goals to be achieved). The intermediary component in our prototype is used to enrich a back end that conforms to the W3C Web Annotation protocol. However, these components could also be used independently from one another. We discuss current limitations and future extensions of our prototypical implementation in the next section.

4.2 Discussion and Limitations

Our prototype uses the Hypothesis back end, which conforms to the W3C Web Annotation Protocol [1]. However, Hypothesis uses OAuth 2, which requires the centralized authentication and authorization of HTTP requests. To further support decentralization, we intend to extend this component to implement the SOLID [17] open specifications. In addition, the MOISE framework we use in our implementation is centralized: an organization provides an explicit and objective standard of behavior, but then it is necessary to monitor the behavior of participants against the defined standard. However, because we use formal explicit representations of organizations in RDF, we could potentially reuse and instantiate organizations as needed to cope with tackling disinformation at scale. We currently implement the organizations using an intermediary component that enriches the Hypothesis back end, but in the future we intend to further decouple these components to have (i) components concerned with the running organizations, and (ii) components concerned with storing and managing W3C Web Annotations. Clients could then discover relevant components at run time via hypermedia. To further support decentralization, in future research we intend to complement our work by investigating additional mechanisms that are conceptually decentralized, such as the use of social protocols [6].

5 CONCLUSION

To tackle disinformation at scale in a decentralized Web in a transparent, adaptable and open manner, we designed and implemented a system that enables and coordinates fact-checking efforts of both humans and machines. To this end, we distilled ongoing efforts for tackling disinformation into an overall fact-checking process that is split into six sub-processes. We used MOISE to conceptualize and formalize this process as a multi-agent organization, and defined the Disinformation Tackler Ontology to allow for unification and reuse of domain concepts.

The proposed contribution opens the door for several interesting problems. For instance, efforts should be directed towards investigating mechanisms to keep user communities healthy. Introducing malevolent user detection or discriminative power at the system level (e.g., for verified fact-checkers) could be promising avenues to explore. Our system aims to put forward the vision of the Web as a collaborative knowledge base that is not mainly generating value as a dissemination mechanism for content, but rather through the discussions this content sparks. In such a Web, the gate-keeper function regarding content quality is distributed on the shoulders of both humans and machines, allowing the individual to knowlegedly navigate through the online information jungle.

REFERENCES