SocialFairness: Assessing Fairness in Digital Media

L. Alfonso Ureña-López[1], M.Teresa Martín-Valdivia[1], Estela Saquete Boró[2] and Patricio Martínez Barco[2]

[1]Department of Computer Science, Advanced Studies Center in ICT (CEATIC), Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[2]Departament of Computer Science. Universidad de Alicante. Campus San Vicente del Raspeig. Building EPS-IV. Alicante, Spain

Abstract

The proliferation of hoaxes on the Internet and toxic messages (with offensive or very negative content and high dissemination rates) constitute a current problem on the road to truthful and respectful information. Due to the enormous influence that social media have in the generation of opinion and as a channel of information for society, the efforts that various companies, organisations and institutions are making to detect and counteract the high volume of disinformation circulating on the networks are important. This project deals with the implementation of a proof of concept of a system for analysing the fairness of messages published through social media, built on the basis of various methods and algorithms from human language technologies. These methods and algorithms are the result of research that the participating groups have been working on for the last few years and are promising solutions for the determination of different levels of quality of publications in two fundamental aspects: their veracity and their toxicity. To address the proof of concept, activities aimed at the definition and integration of these technologies and their evaluation by stakeholders are proposed. This will make it possible to establish the responsiveness of these technologies to the needs of society and industry, as well as their viability to work towards higher levels of technological maturity.

Keywords

Natural Language Processing, NLP, human language technologies, language modeling, machine learning, offensive language and hate speech, toxicity, trustworthiness, misinformation

1. Introduction

Today, the power that misinformation strategies exert over public opinion is a reality that threatens the foundations of democracy. The proliferation of false accounts and the misrepresentation of news seek the propagation of hoaxes that contribute to generate manipulated thoughts. The effectiveness of these strategies has been demonstrated in electoral processes, such as the US presidential election and the election of Donald Trump, or referendums with profound repercussions, such as the exit of the United Kingdom from the European Union in the so-called "Brexit". Today, the injection of hoaxes into social networks is common practice by radical parties throughout Europe. Messages inciting extreme nationalism, hatred of immigrants, and the questioning of social policies, among many others, proliferate without the control of those who listen, but with a clear intention on the part of those who spread them.

We propose a Proof Of Concept (PoC) to address these issues. This work has been partially supported by projects SocialTOX (PDC2022-133146-C21) and SocialTRUST (PDC2022-133146-C22), funded by MCIN/AEI/10.13039/501100011033, and by the "European Union NextGenerationEU/PRTR". Additionally, the technology to which this proof of concept is intended to be applied is the result of research carried out in the coordinated project of the Ministry of Science and Innovation, LivingLang with reference RTI2018-094653-B-C21/C22, titled Human Language Technologies for Living Digital Entities.

The technology proposed in this project consists of a modular system capable of analyzing different characteristics of a publication made in any digital media, including the reliability of such publication, as well as the toxicity of its textual content. The output of the system will be a report related to the disinformation derived from a news item and the level of toxicity. Specifically, the proposed system will have, in this PoC, two synergistic modules: detection of the trustworthiness of the news (SocialTrust) and detection of the toxicity of the news (SocialTox). These two aspects are fundamental to measure the degree of impact that a hoax reaches on the Internet.

For this proof of concept we are going to focus on two modules related to the misinformation of the news published in the media, both the news provided in traditional digital media and its repercussion in social media. On the one hand, one of the modules (SocialTrust) will be in charge of determining the trustworthiness of a news pub-

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.

laurenas@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M.Teresa Martín-Valdivia); stela@dlsi.ua.es (E. S. Boró); patricio@dlsi.ua.es (P. M. Barco)
0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M.Teresa Martín-Valdivia); 0000-0002-6001-5461 (E. S. Boró); 0000-0003-4972-6083 (P. M. Barco)

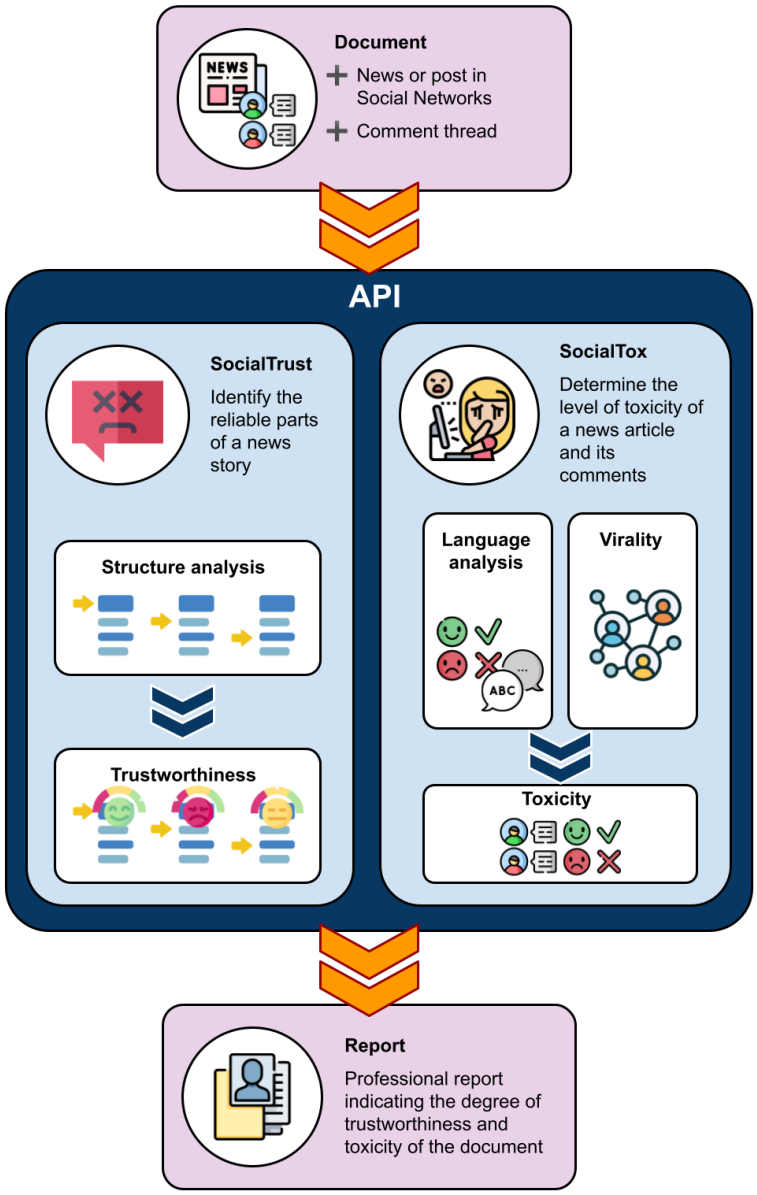

Figure 1: Architecture of the proposed tool

lished in a digital media and with a journalistic format with the novelty of not only characterizing the news, but also indicating which parts of the news text are or are not trustworthy. The research from which this project is derived is based on the fact that misinformation is currently going viral at high speed, intermingling true and false information to confuse the user of such information. Given this premise, the proposed technology will be able to detect the main parts of a news item written under the journalistic style of the inverted pyramid (the most relevant information is presented at the beginning of the news item) and within each part, the tool will detect the relevant information by marking the 5W1H. Once the parts of the news and the essential information are identified, the tool will categorize those elements into a reliability value based on a language model previously

generated with a training corpus. Furthermore, depending on the reliability of the elements, the whole news item will also take an overall reliability value.

On the other hand, and given that the toxic and/or constructive content of a news item, as well as its comments can be indicators of whether or not a news item is false, a module (SocialTox) will be developed to automatically determine the degree of toxicity of a publication. To do this, an analysis of the language and virality of the content will be performed. In previous studies, it has been identified that negative content tends to be more viral than positive content, which could also be an important feature to determine that a news item is false, given the rapid spread of misinformation.

Both modules will be integrated in a common architecture, whose interface will be a backend with APIs. Through calls to these API endpoints it will be possible to interact with the system taking, for example, as input a text (from a news item, a comment or comment thread together with an associated news item, a post in some social network, including or not also the comment thread...), a professional report will be returned indicating the degree of veracity, as well as the toxicity of the document (see Figure 1).

2. Goals

As the origin of this PoC, we will start from an experimental system generated in the origin project that uses a very limited set of training data, but that obtained very satisfactory and competitive results. Thus, the main objective of the PoC is to have a tool that applies different language models for the detection of the degree of confidence and toxicity of the information. This tool will allow, in future projects, its applicability to the value chain of different sectors. With this proof of concept we intend to validate its incorporation in a real way in productive sectors such as journalism, politics, public sector, among others. To achieve this general objective it will be necessary to address the following specific objectives. These objectives are shared in a symbiotic way by the two research teams, being the UJA team in the toxicity aspects focusing on the implementation and development of the SocialTox module, while the UA team will focus on the trustworthiness aspects oriented to the implementation and development of the SocialTrust module:

- Objective 1: Build a dataset large enough to generate the most powerful language models adapted to the domain and language possible for the development of both the SocialTrust module (UA) and the SocialTox module (UJA).
- Objective 2: Train the tool with the dataset. As a result of this training, new language models will be obtained and adjusted to obtain an optimized framework for both the SocialTrust module (UA) and the SocialTox module (UJA).
- Objective 3: Integrate in a single cloud service the algorithms and models that compose a system for measuring the honesty (trustworthiness and toxicity) of a message published in digital media (UJA, UA).
- Objective 4: Define the evaluation framework of the proposed service by identifying the stakeholders, the necessary data and evaluation metrics, which enable the validation of the solution according to the needs of these groups (UJA, UA).
- Objective 5: Perform an analysis of the tests carried out and obtain a final evaluation. This is a primary objective, as it represents the expected goal of any proof of concept (UJA, UA).
- Objective 6: Determine the aspects that can be highlighted for subsequent technological maturation processes. The aim is to identify improvements that facilitate the evolution of the product to higher TRL levels (UJA, UA).

The tool proposed for this PDC is applicable to any productive and social sector and, in some specific cases, could be a support tool for various sectors (journalism, education, administration, etc.), since its application is to support the detection of reliable information, as well as toxic content in the text. In its current state, it is able to give with high precision a value of veracity not only of the news itself, but also of the various information contained in it. In this way, following the annotation structure already successfully defined in its current state, after an enrichment of the generated datasets with veracity and toxicity features and after a training process with a much larger dataset, it will quickly allow to obtain an evidence report of the veracity and toxicity of the information. The initial trainings of the tool have been in Spanish and therefore currently that would be its language of application, but being a tool that is not language dependent, in the future, with a training set in other languages it could be applied to any other language.

3. Scientific and Technical Impact

One of the objectives of this project is the integration of advances achieved in real-world environments. Different types of collaborations with external entities will be explored to assess the possibilities of transferring the generated products and their impact. In essence, the proposed proof of concept has real-world scope, and its application in various scenarios, such as digital media or monitoring information on social networks, will constitute a disruptive technology that enhances access to information.

Natural language is the primary means of interaction in human societies. The rapid development of the Internet in terms of volume and diversity of information, coupled with user access to this data, poses a significant challenge for retrieving and analyzing factual and subjective content for specific purposes and representing them to gain knowledge. These new types of texts are highly subjective, and their automated treatment requires specific methods from Human Language Technologies (HLT).

Another factor to consider is that digital media has created an ecosystem of spaces where content is created and consumed at increasing quantity and speed. However, this ubiquitous environment of interaction among members of current society harbors certain omnipresent content that negatively affects the quality and freedom of information. Digital media has become a space where misinformation, hate speech, or abusive behaviors proliferate, among other contents that can directly harm individual users and society as a whole. Thus, this project will provide the modeling of digital content behavior and the availability of a tool that, by applying different language models, returns the reliability and toxicity level of information. It will contribute, on one hand, to the detection, mitigation, and prevention of harmful digital content, towards cleansing social media on the Internet, and on the other hand, to characterizing beneficial and trustworthy content, thus contributing to ensuring a respectful, safe, and reliable communication environment.

For all these reasons, the project is expected to have scientific-technical impact (both nationally and internationally) in various fields, such as the creation and use of advanced resources or the development, implementation, and integration of specific methods and tools. All these developments will have a strong impact on the scientific community and society. Medium-term transferable results are also expected, working with real practical cases.

4. Social and Economic Impact

The project has a clear ethical and social orientation, proposing the study of automated measures to combat toxicity in digital communications. Given the societal concern generated by certain contents, leading to the adoption of political and legal measures for their treatment, our project will decisively contribute in the medium and long term to the construction of safer digital environments, more beneficial for everyone.

4.1. Description of Impact and Social Benefits

The product can be commercialized from two different perspectives:

1. As a licensing product: Any consulting company providing content management solutions for news (news portals or online journalistic editions) can benefit from marketing the tool as part of their business solution. In this case, the tool provides a quality filter ensuring the reliability of content.
2. As a SaaS (Software as a Service) product: Online advertising agencies marketing their advertising portals can benefit from having a quality seal that guarantees the prestige of the advertising space. In this case, marketing would be done per service provided.

4.2. Description of the Project's Proximity to the Market or Target End Users

The issue of misinformation is not exclusive to a particular sector; rather, it affects virtually all productive and social sectors. Sectors such as politics, public administration, healthcare, industry, brands, advertising, and culture have ample examples of detrimental consequences of misinformation caused by fake news when they enter their sphere of action. However, the primary affected sector and the origin of problems for other sectors is the journalistic industry. The inclusion of misinformation in a media outlet's editing, often caused by echoing unreliable sources, affects the credibility of the outlet and its editorial, seriously impacting its advertising capacity due to the discrediting that may occur to a commercial brand appearing alongside news proven to be unreliable or toxic. The possible inclusion of misinformation in traditionally considered reliable media by their audience poses a significant risk that the publishing industry tries to combat by dedicating considerable time to verifying sources and information. However, in the current state of this sector, time works against the business, as news loses its value within a few hours of being published, and a media outlet that delays publication becomes outdated. The traditional industry believed that news could have a lifespan of at least 24 hours, the time it took to edit a new newspaper, but the digitization of news and its availability on the Internet dramatically reduces the news cycle. For this reason, the industry needs automated tools to detect unreliable or toxic information, reducing the verification time for news by professional journalists while increasing precision in determining content reliability.

Under this premise, this proposal of PoC is primarily framed in the journalistic industry sector. According to the report "The Newspaper Publishing Industry" [1] from the European Union (2012), the newspaper industry has considerably altered its value chain due to digitization and Internet growth, leading to a shift in business from print journalism to online. In this new scenario, advertising goes from representing 43% of revenue

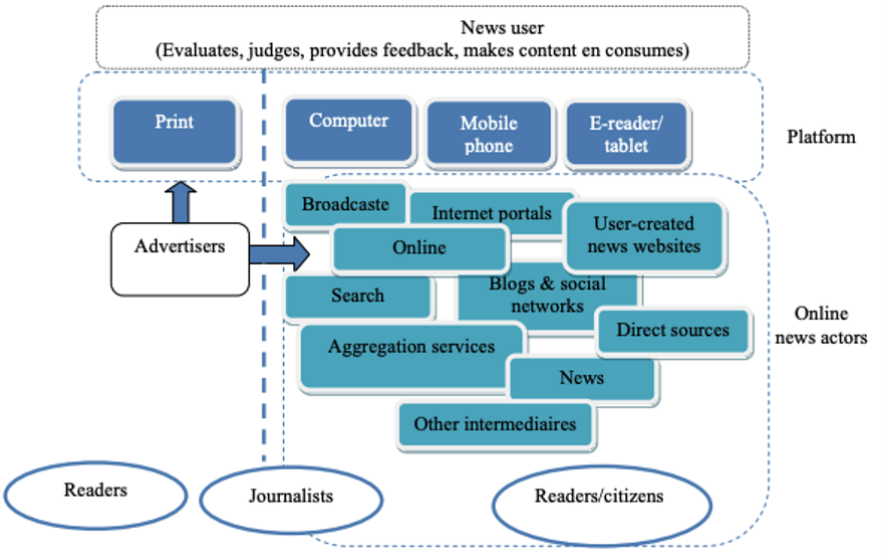

Figure 2: Value chain in the new newspaper industry (Source: [1]).

sources (in 2010) to being the main source of revenue today. However, according to this report, online advertising is cheaper, and the audience is more fragmented and pays less attention to advertising than in the case of "written" advertising. All of this leads media outlets to the need for greater efforts in finding and retaining advertisers. This is one of the determining aspects to care for the article's quality and thus ensure an advertiser a reliable space for their brand using tools like the one proposed in this PoC.

According to this source, the different actors participating in this sector, and therefore potential clients of our PoC, include especially the journalistic sector: print and broadcast media, and online media: web news, portals and news aggregators, user-generated news (blogs), social networks, and platforms.

The Figure 2 illustrates the value chain of the journalistic industry, as per the referenced report, highlighting the involvement of new actors. In this scenario, users take on a more participatory role, not only as assessors and judges of information but also as content creators. The proposed tool gains significance in this new context by acting as a filter for information whose origin is beyond the editor's control, providing assurance for advertisers seeking to associate their product with a prestigious space.

Alternatively, another productive sector that could benefit from the tool is the food industry. Social networks and other media portals daily accumulate a vast collection of articles analyzing the merits and drawbacks related to the consumption of both generic and commercialized food products. These articles, widely circulated, create a network of influencers with a high number of followers who entrust their diet to the judgment of these analysts, causing radical changes in consumption habits. This landscape is further influenced by commercial wars and other market interests [2]. Specifically, the II Study on Health Hoaxes [3], conducted by SaludsinBulos and Doctoralia, determines that hoaxes about food constitute 57% of false beliefs detected by doctors in consultations.

The proposed tool will allow the evaluation of the reliability of articles published about commercial products, generating effective reliability and toxicity reports against negative articles with the brand. It can also enhance the dissemination of quality articles, enabling companies to strategically manage communication with potential clients, where silence is not an option.

Acknowledgments

This work has been partially supported by projects **SocialTOX** (PDC2022-133146-C21), **SocialTRUST** (PDC2022-133146-C22) funded by MCIN/AEI/ 10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"

References

[1] A. Leurdijk, M. Slot, O. Nieuwenhuis, The newspaper publishing industry. statistical, ecosystems and

competitiveness analysis of the media and content industries, 2012.
[2] I. Lorenzo, Cómo afrontar la desinformación en la alimentación: Organizaciones sectoriales, empresas e instituciones se enfrentan al reto de las fake news, con estrategias de educación tecnológica y comunicación reputacional, Distribución y consumo 29 (2019) 62–67.
[3] A.-A. de Investigadores en eSalud, Ii estudio sobre bulos en salud. encuesta a profesionales de la salud de españa, 2019.