Défis et limites de la détection de fausses informations prononcées en émissions de télévision

Jérémie Nicey, Frédéric Rayar

Défis et limites de la détection de fausses informations prononcées en émissions de télévision

Jérémie NICEY [(1)] et Frédéric RAYAR [(2)]

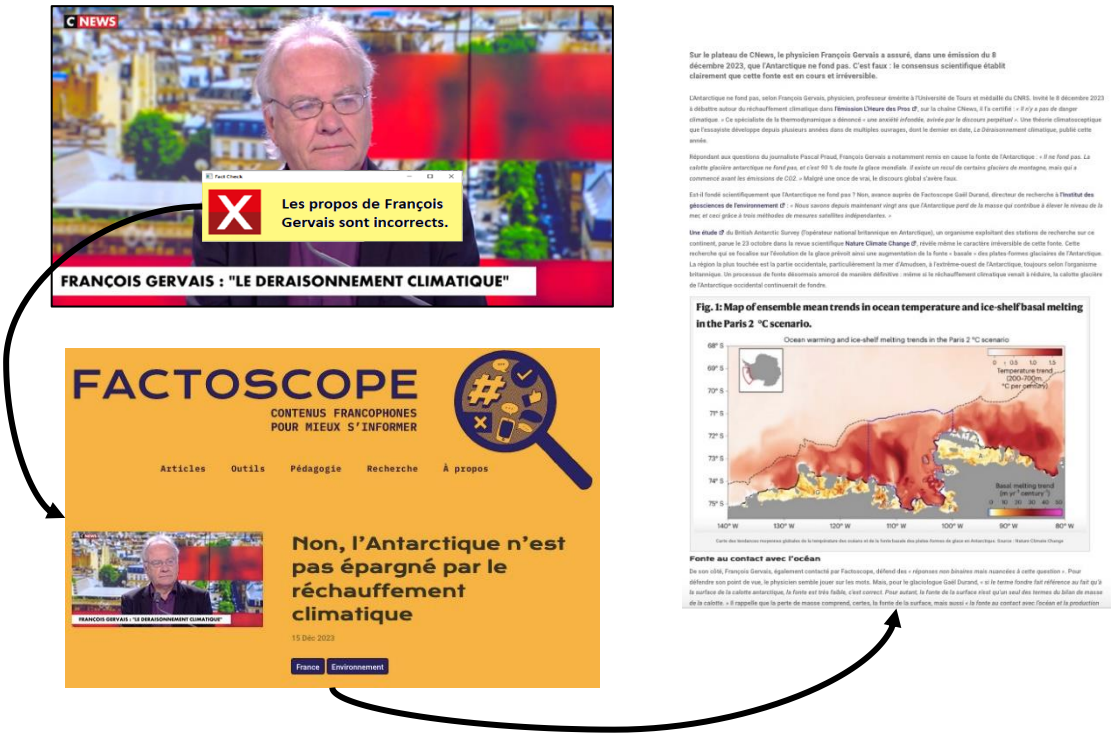
[(1)] PRIM, EA 7503, Université de Tours, jeremie.nicey@univ-tours.fr

[(2)] LIFAT, EA 6300, Université de Tours, frederic.rayar@univ-tours.fr

Cette proposition de communication s'inscrit dans la continuité d'un projet transdisciplinaire (mêlant sciences informatiques et sciences de l'information et de la communication) qui ambitionne le développement d'un dispositif numérique visant, à terme, la détection semi-automatique et multimodale de fausses informations en télévision, en particulier dans les émissions, interviews et débats politiques, toujours très consommés malgré l'existence de nombreuses alternatives proposées par les espaces numériques, et plus encore à l'approche d'échéances électorales.

Nous avions antérieurement (lors de la précédente – et première – édition de ce même *workshop*, en mars 2023) présenté la conceptualisation, les fondements techniques et l'intérêt aussi bien citoyen (pour les téléspectateurs) que professionnel (pour les équipes journalistiques : Nakov *et al.*, 2021 ; Bigot, 2019) d'un tel outil. Quel(s) qu'en soi(en)t le(s) futur(s) *design* (application sur smartphone, *plugin* sur un terminal de lecture, ou fenêtre *pop-up* sur l'écran même du téléviseur), cet outil permettrait au moment de la diffusion d'une émission de télévision (autant en linéaire qu'en *replay*) : *(i)* d'une part de détecter les thèmes sensibles, douteux ou ambigus évoqués par les interlocuteurs et ayant déjà fait l'objet d'articles de vérifications (Shaar *et al.*, 2020) par les journalistes *fact-checkers*, de plusieurs médias et agrégés dans une base de données ; *(ii)* d'autre part d'identifier les interlocuteurs déjà repérés comme approximatifs ou mensongers dans leurs déclarations antérieures. Nous avions à cette occasion souligné l'objectif de guider, *via* une alerte visuelle, les utilisateurs vers les contenus vérifiés et adéquats, constitutifs de ladite base de données.

Nous proposons lors de la nouvelle édition du *workshop* d'interroger cette fois les défis et limites d'un tel dispositif, au regard des connaissances déjà établies concernant les mécanismes des fausses informations.



Défis et limites techniques

La conception de notre prototype nécessite de lever plusieurs verrous scientifiques et technologiques. Il conviendra de discuter de l'harmonisation et de l'incrémentalité d'une base de données de contenus de *fact-checking* depuis laquelle sera effectué le *matching* avec les énoncés politiques douteux en télévision, de même que la nature des algorithmes dudit *matching* et leur interopérabilité entre vocal et texte.

En outre, les complexités liées à la superposition vocale (habituelle lors des débats en France) ou aux dimensions temporelles (détection et liaison attendues en temps rapide sinon réel ; mais aussi enjeux de chronologie et d'actualisation des vérifications archivées) seront discutées. Elles seront complétées par la prise en compte d'aspects langagiers retors, pour certains propres à la construction du discours politique, tels que les formulations rhétoriques, l'utilisation de la double négation (grammaticale) ou de l'ironie.

Ces réflexions, bénéficiant notamment de l'apport d'autres travaux, que ce soit sur la technicité multimodale (Akhtar *et al.*, 2023), sur les mots signifiants et récurrents mobilisés dans les contenus de désinformation (Maine *et al.*, 2023) ou sur les expérimentations de détection du « vague lexical » (Icard *et al.*, 2023), seront ainsi mises en perspective avec les leçons issues de la littérature académique internationale de ces dernières années sur les biais des tentatives d'automatisation du *fact-checking* (Graves, 2018 ; Adair, 2021) et sur la nécessaire intervention humaine complémentaire (Nakov, 2021).

Défis et limites légales/réglementaires

Quelles que soient la pertinence et l'efficience de notre futur dispositif de détection et de contextualisation des fausses informations télévisées, sa mise en place ne peut s'affranchir du cadre légal, à la fois aux niveaux national et supranational. Ainsi, nous évoquerons brièvement dans cette partie les dimensions liées aux droits d'auteurs (notamment des productions de *fact-checking*). Nous soulèverons à cette occasion les enjeux de « coopétition » et de « coalition » des unités de *fact-checking* – françaises voire désormais francophones – c'est-à-dire le nécessaire partenariat des médias professionnels (déjà existant dans le cadre de plusieurs initiatives) afin d'enrichir et de pérenniser la base de données des faits pré-vérifiés.

À une échelle différente, nous conduirons également la réflexion sur les récentes dispositions réglementaires européennes portant sur l'intelligence artificielle (AI Act : avril 2021 et décembre 2023) ou sur la place des services numériques (Digital Services Act – DSA – publié en octobre 2022 et entré en application en août 2023), incluant notamment la responsabilité des plateformes faisant circuler de l'information.

Défis et limites éditoriales, éthiques et d'appropriation par les publics

Notre projet se veut utile aux citoyens et aux équipes de télévision mais, pour ne pas se révéler contre-productif, il doit s'appuyer sur les connaissances concernant les biais et effets environnant la réception des fausses informations, fruits des diverses sciences humaines et sociales. Nous interrogerons ici, entre autres, la place que doivent prendre nos alertes visuelles proposées aux téléspectateurs : dans leur calibre et leur *design* (en envisageant des étapes de *feedback* des utilisateurs) et surtout dans leur fréquence (en tenant compte du degré de lassitude ou de "pollution attentionnelle" vécues face aux vérifications d'information).

De même, nous devons penser la signification de la détection, et ses effets contre-intuitifs : par exemple le postulat du mensonge politique généralisé, ou la position de surplomb/d'évaluation permanente, ou encore les risques liés à la non-exhaustivité des productions de *fact-checking* (un mensonge qui ne pourrait pas être relié à la base de données car n'ayant pas fait l'objet d'une vérification antérieure, ne générerait pas d'alerte et risquerait ainsi de passer pour vrai : Pennycook *et al.*, 2019). Autre biais, majeur : la rationalité et le *fact-checking* ne suffisent pas à changer les croyances (Barrera *et al.*, 2020), voire comportent un « effet boomerang » (l'incitation à lire des articles qui corrigent/contredisent un candidat peut en réalité être fortement rejetée par ses sympathisants, voire les renforce dans leur position : Christenson *et al.*, 2021). Ces limites et ces contraintes, et d'autres encore, nous conduiront à interroger l'indispensable accompagnement éditorial, à terme, de notre dispositif.

Bibliographie indicative :

ADAIR Bill, 2021 (16 juin), « The lessons of Squash, Duke's automated fact-checking platform », *Poynter Institute*, https://www.poynter.org/fact-checking/2021/the-lessons-of-squash-the-first-automated-fact-checking-platform/

AKHTAR Mubashara, SCHLICHTKRULL Michael, GUO Zhijiang, COCARASCU Oana, SIMPERL Elena, VLACHOS Andreas, 2023, « Multimodal automated fact-checking: A survey », *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): Findings*, https://arxiv.org/pdf/2305.13507.pdf

ALTAY Sacha, BERRICHE Manon, ACERBI Alberto, 2023, « Misinformation on Misinformation: Conceptual and Methodological Challenges », *Social Media + Society*, pp. 1-13, https://journals.sagepub.com/doi/pdf/10.1177/20563051221150412

BARRERA Oscar, GURIEV Sergei, HENRY Emeric, ZHURAVSKAYA Ekaterina, 2020, « Facts, alternative facts, and fact checking in times of post-truth politics », *Journal of Public Economics*, Vol. 182, article 104123, https://www.sciencedirect.com/science/article/pii/S0047272719301859

BIGOT Laurent, 2019, *Fact-checking vs. fake news. Vérifier pour mieux informer*, INA Editions.

CHRISTENSON Dino, KREPS Sarah, KRINER Douglas, 2021, « Contemporary presidency. Going public in an era of social media: tweets, corrections and public opinion », *Presidential Studies Quarterly*, Vol. 51, n°1, p. 151-165, https://onlinelibrary.wiley.com/doi/abs/10.1111/psq.12687

GRAVES Lucas, 2018, « Understanding the promise and limits of automated fact-checking », in *Reuters Institute for the Study of Journalism Factsheets*, Oxford University, https://ora.ox.ac.uk/objects/uuid:f321ff43-05f0-4430-b978-f5f517b73b9b

ICARD Benjamin, GUÉLORGET Paul, GADEK Guillaume, GAHBICHE Souhir, FAYE Géraud, GATEPAILLE Sylvain, ATEMEZING Ghislain, CLAVEAU Vincent, ÉGRÉ Paul, 2023, « Détection du vague lexical et identification par apprentissage profond des fausses informations », Workshop *Infox-sur-Seine 2023*.

MAINE François, BANCILHON François, GANASCIA Jean-Gabriel, PUJOLLE Guy, 2023, « Catégorisation automatique d'infox par apprentissage supervisé : expérimentations sur texte seul », Workshop *Infox-sur-Seine 2023*.

NAKOV Preslav, CORNEY David, HASANAIN Maram, ALAM Firoj, ELSAYED Tamer, BARRÓN-CEDEÑO Alberto, PAPOTTI Paolo, SHAAR Shaden, DA SAN MARTINO Giovanni, 2021, « Automated fact-checking for assisting human fact-checkers ». *CoRR*, abs/2103.07769.

NICEY Jérémie, BIGOT Laurent, 2020, « Le soutien de Google et de Facebook au *fact-checking* français : entre transparence et dépendance », *Sur le journalisme*, Vol. 9, n°1, pp. 188-203, http://www.surlejournalisme.kinghost.net/rev/index.php/slj/article/view/417

NICEY Jérémie, BIGOT Laurent, 2019, « Un pour tous, tous pour un ? Les pratiques inédites de "coalition" des journalistes *fact-checkers* français durant la campagne présidentielle de 2017 », in A. Theviot (dir.), *Médias et élections. Les campagnes présidentielle et législatives de 2017*, Ed. Septentrion, pp. 121-141.

PENNYCOOK Gordon, BEAR Adam, COLLINS Evan T., RAND David G., 2019, « The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings », *Management Science*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384

RAYAR Frédéric, 2024, "Fact-Checked Claim Detection in Videos Using a Multimodal Approach". In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-679-8, ISSN 2184-4321, pages 540-546.

RAYAR Frédéric, DELALANDRE Mathieu, LE Van-Hao, 2022, « A large-scale TV video and metadata database for French political content analysis and fact-checking », in Association for Computing Machinery (dir.), *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing* (*CBMI '22)*, pp. 181-185.

SAUVAGEAU Florian, THIBAULT Simon, TRUDEL Pierre (dir.), 2018, *Les fausses nouvelles : nouveaux visages, nouveaux défis*, Presses de l'Université Laval.

SHAAR Shaden, BABULKOV Nikolay, DA SAN MARTINO Giovanni, NAKOV Preslav, 2020, « That is a known lie: Detecting previously fact-checked claims », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3607-3618.