Ontologie - Sys. de Vérification de la Crédibilité de l'Information

Projet Final présenté à Petko Valchev, DIC-9335 Sciences du Web

Dominique S. Loyer

Département d'informatique, UQAM
loyer.dominique@courrier.uqam.ca

30 avril 2025

Plan de la Présentation

Le Défi de la Crédibilité de l'Information

- Prolifération exponentielle de l'information en ligne (Zhou et Zafarani 2020 ; Viviani et Pasi 2017).
- Difficulté croissante pour les utilisateurs à évaluer la fiabilité des sources et des contenus (Metzger, Flanagin et Medders 2010 ; Rieh 2010).
- Impact sociétal majeur de la désinformation (mésinformation, *fake news*) (Sharma et al. 2019 ; Souza et al. 2020 ; Oshikawa, Qian et Wang 2020).
- Complexité accrue par l'IA générative (GenAI) capable de créer des contenus synthétiques réalistes (K. Chen et Shu 2023 ; Loth, Kappes et Pahl 2024).
- Besoin de transparence et d'explicabilité des systèmes IA, notamment les LLMs (**Liao2024** ; **Mahari2023** ; **Zou2023** ; **Chaudhary2024** ; **DiakopoulosND**).

But Principal et Objectifs Spécifiques

But Principal : Développer une ontologie OWL 2 DL (**W3C2012OWL2**) pour modéliser formellement un système d'évaluation de la crédibilité, basé sur une modélisation UML préalable (**Loyer2025Rapport**).

Objectifs Spécifiques :

- Représenter le processus : `RequeteEvaluation` → Analyse → `RapportEvaluation`.
- Modéliser les acteurs : `User` (demandeur), `Expert` (configurateur).
- Capturer l'information : `InformationSoumise`, `Source`, `Author`.
- Intégrer l'approche hybride : `RegleVerification` (logique) + `ModeleIA` (sémantique).
- Définir l'évaluation : `VerificationCriterion`, `ResultatCritere`, `CredibilityLevel`.
- Inclure les données externes : `SystemeExterne`, `Evidence`.
- Permettre l'explicabilité et la classification inférée.
- Considérer les aspects de tracking et d'anonymisation (**Acar2014** ; **Yang2024** ; **Jaff2024** ; **Staab2024**).

Classes Principales

Processus & Acteurs

- RequeteEvaluation
- RapportEvaluation
- User / Expert
- SystemeExterne
  - MoteurRecherche
  - ApiLLM
  - BaseDeFaits

Information & Provenance

- InformationSoumise
- Source (+ sous-classes)
- Author
- Evidence (+ sous-classes)

Évaluation & Résultats

- VerificationMethod
  - RegleVerification
  - ModeleIA
- VerificationCriterion
- ResultatVerification
  - ResultatRegle
  - ResultatNLP
- ResultatCritere (*Nouveau*)
- CredibilityLevel
- InfoSourceAnalyse

Classes Définitionnelles

- InformationVerifiee
- InfoHauteCredibilite
- InfoMoyenneCredibilite
- InfoFaibleCredibilite

Propriétés Clés

Relations Principales (Propriétés d'Objet)

- *Processus :* `concernsInformation`, `submittedBy`, `producesReport`, `isReportOf`...
- *Provenance :* `hasOriginalSource`, `hasAuthor`, `originatesFrom`...
- *Évaluation :* `includesRuleResult`, `includesNLPResult`, `assignsCredibilityLevel`, `basedOnEvidence`...
- *Lien Méthode-Résultat :* `appliesRule`, `usesModel`...
- *Lien Méthode-Critère (Nouveau) :* `evaluatesCriterion`
- *Résultat par Critère (Nouveau) :* `hasCriterionResult`, `concernsCriterion`, `obtainedVia`...
- *Configuration :* `configuredByExpert`...

Attributs (Propriétés de Données)

- *Requête :* `requestStatus`, `submissionTimestamp`...
- *Information :* `informationContent`, `informationURL`...
- *Rapport :* `credibilityScoreValue`, `reportSummary`, `completionTimestamp`...
- *Résultats Règles/NLP :* `ruleResultValid`, `sentimentScore`, `coherenceScore`...
- *Résultat Critère (Nouveau) :* `criterionResultValue`, `criterionResultConfidence`...
- *Niveau Crédibilité :* `credibilityLevelValue`...

Processus de vérification

Flux Modélisé :

1. `User` $\xrightarrow{\texttt{submitsRequest}}$ `RequeteEvaluation`
2. `RequeteEvaluation` $\xrightarrow{\texttt{concernsInformation}}$ `InformationSoumise`
3. *Analyse Hybride (Règles + IA) utilisant* `SystemeExterne`
4. *Génération de* `ResultatRegle`, `ResultatNLP`, `ResultatCritere`
5. `RequeteEvaluation` $\xrightarrow{\texttt{producesReport}}$ `RapportEvaluation`
6. `RapportEvaluation` $\xrightarrow{\texttt{includes...}}$ *Résultats détaillés*
7. `RapportEvaluation` $\xrightarrow{\texttt{assignsCredibilityLevel}}$ `CredibilityLevel` (*ex : Niveau_Haut*)

Inspiration de l'Ontologie Subvention (Loyer2025OntoSubv) : La structure `RapportEvaluation` → `ResultatCritere` → `VerificationCriterion` est analogue à `EvaluationSommaire` → `NoteAttribuee` → `CritereEvaluation`, permettant une analyse fine par critère.

Approche Hybride et Classification

Modélisation de l'Approche Hybride :

- Classes distinctes : `RegleVerification` et `ModeleIA` (sous `VerificationMethod`).
- Propriété `evaluatesCriterion` lie explicitement méthode et critère.
- `ResultatCritere` agrège potentiellement les sorties via `obtainedVia`.
- L'ontologie structure les éléments, l'algorithme de combinaison est externe.

Classification par Inférence :

- Utilisation de `owl:equivalentClass` pour `InfoHauteCredibilite`, etc.
- Basée sur la valeur de `assignsCredibilityLevel` dans le `RapportEvaluation` associé.
- Utilisation de `owl:complementOf` pour assurer l'exclusivité (inspiré de (**Loyer2025OntoSubv**)).
- Permet au raisonneur OWL de déduire la catégorie de crédibilité.

Les classes et les relations de mon ontologie

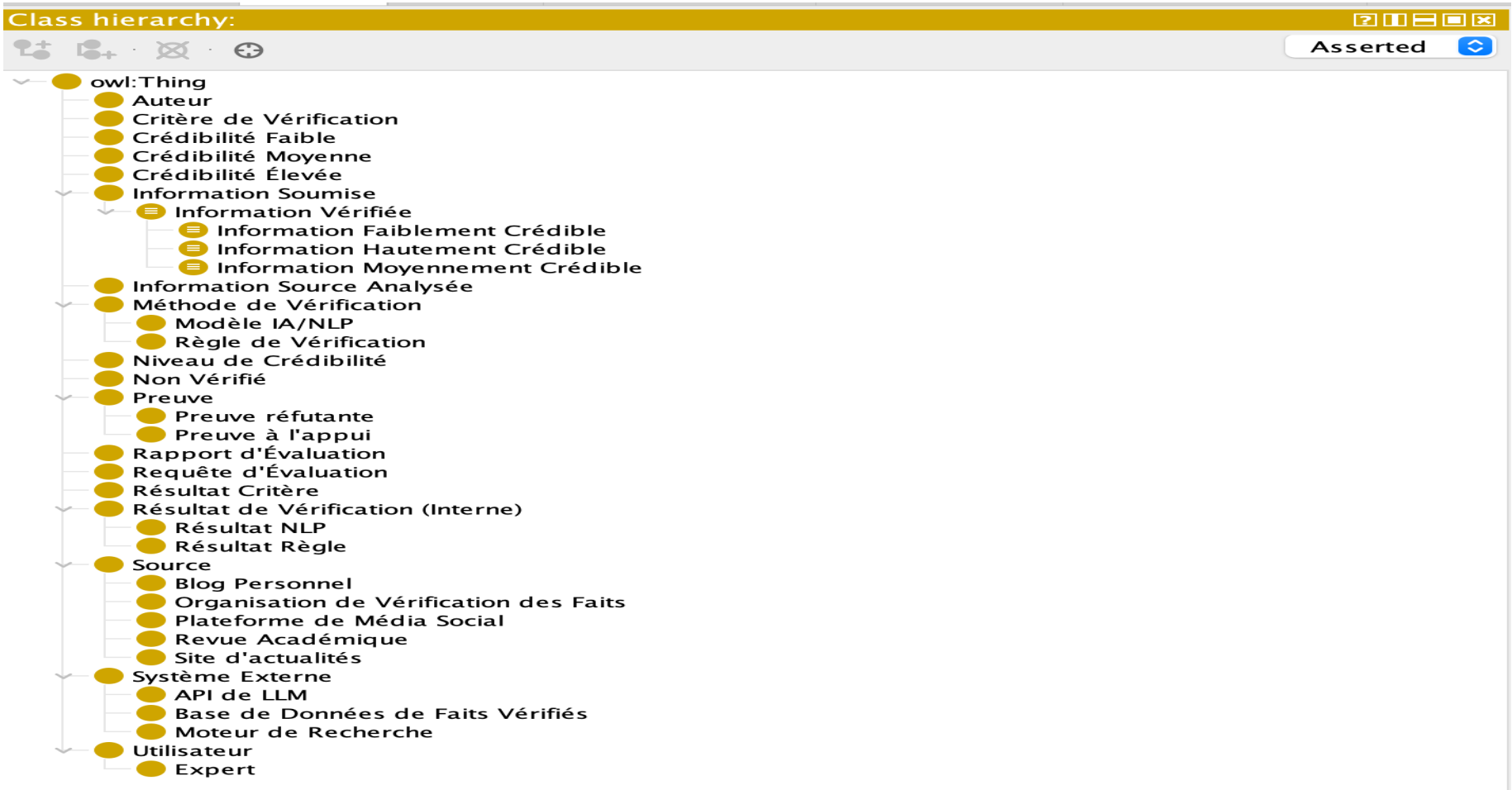
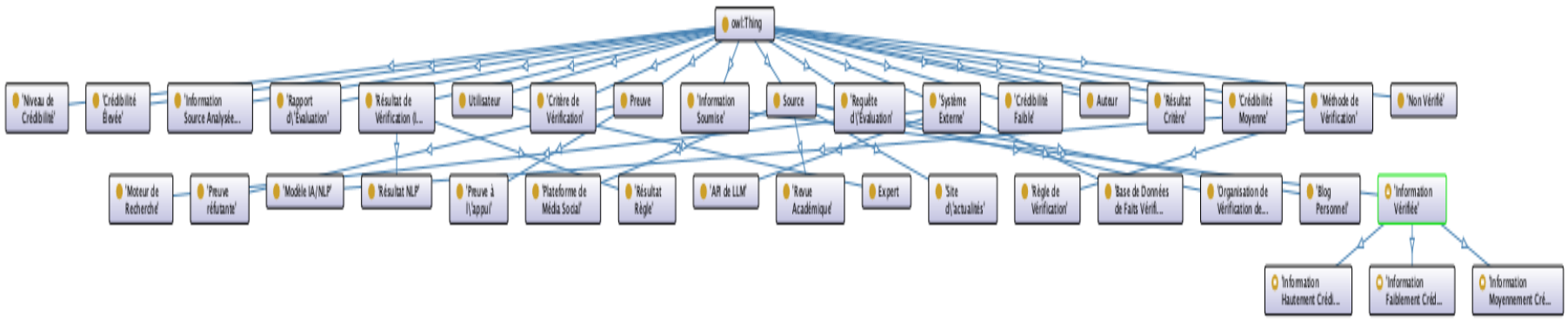


Figure 1 – Les classes de mon ontologie

La Taxonomie de mon Ontologie



Les propriétés des objets

Figure 3 – Les propriétés des objets

Les propriétés des données

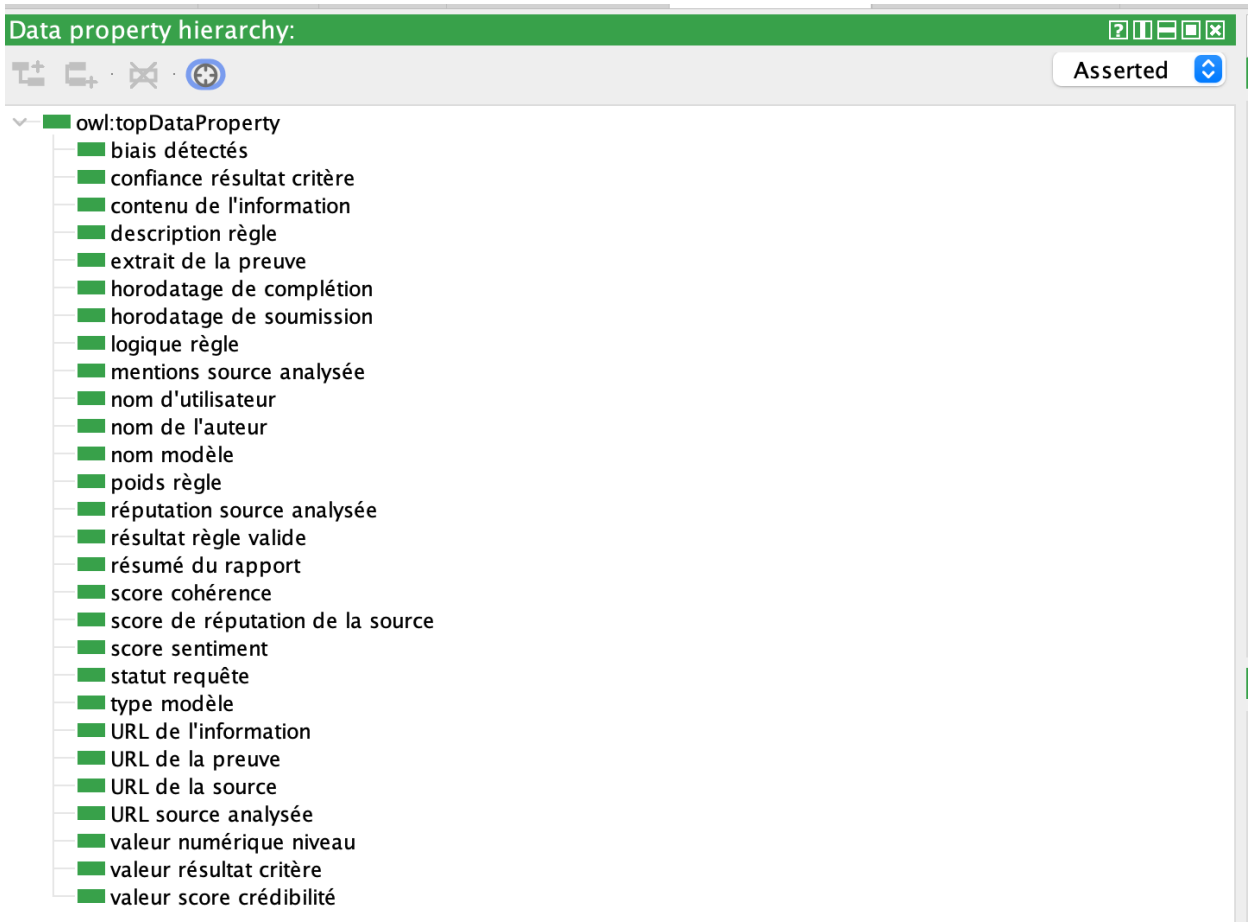


Figure 4 – Les propriétés des données

Utilisation avec Protégé

Compatibilité et Chargement :

- Syntaxe Turtle (TTL) (**TurtleSpec**) valide pour OWL 2 DL (**W3C2012OWL2**).
- Chargeable directement dans Protégé (**Protege2025**) (v5.x).
- Structure claire avec commentaires et labels en français.

Exploration et Raisonnement :

- Navigation dans les hiérarchies de classes (ex : `Source`, `VerificationMethod`) et de propriétés.
- Examen des axiomes :
  - Restrictions (cardinalité, valeur) sur les classes clés (`RequeteEvaluation`, `RapportEvaluation`, `ResultatCritere`...).
  - Disjonctions entre types (sources, niveaux, résultats...).
  - Définitions par équivalence (`InformationVerifiee`, `InfoHauteCredibilite`...).
- Instanciation possible pour tester des scénarios.
- Utilisation de raisonneurs (HermiT, Pellet) pour :
  - Vérifier la cohérence logique de l'ontologie.
  - Inférer les types basés sur les axiomes `owl:equivalentClass` (ex : classifier une instance d'`InformationSoumise`).

Bilan : Points Forts et Limitations

Points Forts du Modèle Ontologique :

- Alignement renforcé avec la modélisation UML (**Loyer2025Rapport**).
- Intégration de structures d'évaluation granulaires (via `ResultatCritere`) inspirées de (**Loyer2025OntoSubv**).
- Support explicite de l'approche hybride et de la configuration par l'`Expert`.
- Mécanisme de classification inférable basé sur le niveau de crédibilité final.
- Base sémantique améliorée pour l'explicabilité potentielle (**Liao2024** ; **Zou2023**).

Limitations Persistantes :

- Logique interne des règles/modèles et algorithme de scoring restent externes à l'ontologie.
- Modélisation des aspects dynamiques (flux d'appels API) limitée.
- Complexité accrue avec l'ajout de `ResultatCritere`.
- Nécessite des données d'instance bien formées pour que l'inférence de classification fonctionne.
- Ne modélise pas explicitement les mécanismes de tracking (**Acar2014**) ou d'anonymisation (**Yang2024** ; **Staab2024**).

Conclusion et Perspectives

Conclusion :

- L'ontologie v2.1 offre une représentation sémantique structurée pour la vérification de crédibilité.
- Elle combine la structure issue de l'UML (**Loyer2025Rapport**) avec des mécanismes d'évaluation inspirés par (**Loyer2025OntoSubv**).
- Elle fournit une base formelle pour un système d'évaluation, supportant la classification et l'explicabilité (**Chaudhary2024** ; **DiakopoulosND**).

Pistes d'Amélioration / Travaux Futurs :

- **Logique d'Agrégation :** Explorer comment modéliser la logique qui mène des `ResultatCritere` au `CredibilityLevel` final (SWRL ?).
- **Affiner Critères/Règles/Modèles :** Définir des sous-classes plus spécifiques.
- **Intégration Externe :** Aligner avec des ontologies existantes (Schema.org, etc.).
- **Gestion Incertitude/Confiance :** Modéliser `criterionResultConfidence`.
- **Validation Empirique :** Tester avec des données réelles.
- **Transparence LLM :** Intégrer concepts de (**Liao2024** ; **Mahari2023** ; **Zou2023**).
- **Vie Privée :** Modéliser aspects d'anonymisation/exposition (**Yang2024** ; **Jaff2024** ; **Staab2024**).

Références I

Adali, Sibel et al. (2015). "Measuring Behavioral Trust in Social Networks". In : *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, p. 153-160. doi : 10.1145/2808797.2808811.

Ahmed, Tanveer, Issa Traore et Sherif Saad (2024). "AI-Driven Hybrid Model for Fake News Detection : Integrating NLP, Sentiment Analysis, and Source Credibility to Combat Misinformation". In : *ResearchGate (Preprint)*. Accessed April 2025, DOI might be available later. url : https://www.researchgate.net/publication/390494580.

Baly, Ramy et al. (2020). "We Can Detect Your Bias : Predicting the Political Ideology of News Articles". In : *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6785-6791. doi : 10.18653/v1/2020.emnlp-main.552.

Barzilay, Regina et Mirella Lapata (2008). "Modeling Local Coherence : An Entity-Based Approach". In : *Computational Linguistics* 34.1, p. 1-34. doi : 10.1162/coli.2008.34.1.1.

Castillo, Carlos, Marcelo Mendoza et Barbara Poblete (2011). "Information Credibility on Twitter". In : *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, p. 675-684. doi : 10.1145/1963405.1963500.

Chen, Kai et Kai Shu (2023). *Combating Misinformation in the Age of LLMs : Opportunities and Challenges*. arXiv : 2311.05656 [cs.CL].

Chen, Wei-Fan et Alice Oh (2020). *Detecting Media Bias using Gaussian Mixture Models : An Unsupervised Approach*. arXiv : 2010.08096 [cs.CL].

Références II

Fowler, Martin (2003). *UML Distilled : A Brief Guide to the Standard Object Modeling Language*. 3rd. Boston, MA, USA : Addison-Wesley.

Garg, Sahaj et al. (2019). "Counterfactual Fairness in Text Classification through Robustness". In : *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, p. 219-226. doi : 10.1145/3306618.3314246.

Giachanou, Anastasia et Fabio Crestani (2016). "Like it or not : A survey of Twitter sentiment analysis methods". In : *ACM Computing Surveys (CSUR)* 49.2, p. 1-41. doi : 10.1145/2903992.

Google (s. d.[a]). *Custom Search JSON API*. Google Developers Documentation. Accessed April 2025. url : https://developers.google.com/custom-search/v1/overview.

— (s. d.[b]). *Fact Check Tools API*. Google Developers Documentation. Accessed April 2025. url : https://developers.google.com/fact-check/tools/api.

Hagag, Ben et Reut Tsarfaty (2023). "The Truth, The Whole Truth, and Nothing but the Truth : A New Benchmark Dataset for Hebrew Text Credibility Assessment". In : *Findings of the Association for Computational Linguistics : EMNLP 2023*. Singapore : Association for Computational Linguistics, p. 3850-3865. url : https://aclanthology.org/2023.findings-emnlp.251.

He, Li, Siyi Hu et Ailun Pei (2023). *Debunking Disinformation : Revolutionizing Truth with NLP in Fake News Detection*. arXiv : 2308.16328 [cs.AI]. url : https://arxiv.org/abs/2308.16328.

Kedzie, Christopher, Kathleen McKeown et Fernando Diaz (2018). "Content-Driven Detection of False News". In : *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Santa Fe, New Mexico, USA : Association for Computational Linguistics, p. 1771-1783.

Références III

Krieger, Felix M. et al. (2021). "DA-RoBERTa : Domain-adaptive RoBERTa for detecting media bias". In : *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2675-2681. doi : 10.18653/v1/2021.eacl-main.232.

Larman, Craig (2004). *Applying UML and Patterns : An Introduction to Object-Oriented Analysis and Design and Iterative Development*. 2nd. Upper Saddle River, NJ, USA : Prentice Hall.

Li, Jiwei et Eduard Hovy (2014). "A Model of Coherence Based on Distributed Sentence Representation". In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2039-2048. doi : 10.3115/v1/D14-1214.

Loth, Alexander, Martin Kappes et Marc-Oliver Pahl (2024). *Blessing or curse ? A survey on the Impact of Generative AI on Fake News*. v2, 27 Dec 2024. arXiv : 2404.03021 [cs.CL].

Lounis, Hakim (2025). *Modélisation des problèmes complexes en sciences cognitives*. Présentations et diapositives du cours DIC-9251, UQAM.

Menzner, Philipp et Jochen L. Leidner (2024). *BiasScanner : Detecting Biased Statements in News Articles*. arXiv : 2401.01793 [cs.CL].

Metzger, Miriam J., Andrew J. Flanagin et Ryan B. Medders (2010). "Social and Heuristic Approaches to Credibility Evaluation Online". In : *Journal of Communication* 60.3, p. 413-439. doi : 10.1111/j.1460-2466.2010.01488.x.

Microsoft (s. d.). *Bing Search APIs documentation*. Microsoft Azure Documentation. Accessed April 2025. url : https://docs.microsoft.com/en-us/bing/search-apis/.

Références IV

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In : *arXiv preprint arXiv :1301.3781*.

Nabozny, Aleksandra et al. (2021). "Active Annotation in Evaluating the Credibility of Web-Based Medical Information : Guidelines for Creating Training Data Sets for Machine Learning". In : *JMIR Medical Informatics* 9.11, e26065. doi : 10.2196/26065. url : https://medinform.jmir.org/2021/11/e26065/.

Osborne, Francesco et al. (2024). ***CimpleKG : A Knowledge Graph for Linking Claims with Explanations and Context***. ISWC 2024 Resources Track Submission. url : https://oro.open.ac.uk/101150/1/iswc2024_resources_track_cimplekg_cr.pdf.

Oshikawa, Ray, Jing Qian et William Yang Wang (2020). "A Survey on Natural Language Processing for Fake News Detection". In : *arXiv preprint arXiv :1811.00770*. v3.

Pennington, Jeffrey, Richard Socher et Christopher D. Manning (2014). "Glove : Global Vectors for Word Representation". In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532-1543. doi : 10.3115/v1/D14-1162.

Pescuma, Vincenzo N. et al. (2025). "Source Credibility Assessment : A Comprehensive Survey". In : *International Journal of Interactive Multimedia and Artificial Intelligence*. T. TODO, TODO. doi : 10.9781/ijimai.2025.01.002. url : https://www.ijimai.org/journal/sites/default/files/2025-01/ip2025_01_002.pdf.

Prabhakaran, Vinodkumar, Ben Hutchinson et Margaret Mitchell (2019). ***Perturbation Sensitivity Analysis to Detect Unintended Model Biases***. arXiv : 1910.04210 [cs.CL].

Références V

Premaratne, Pasan et al. (2012). "Consensus Algorithms for Credibility Assessment of Soft Information". In : *Proceedings of the 4th International Conference on Applied Human Factors and Ergonomics (AHFE)*. San Francisco, CA, USA.

Rieh, Soo Young (2010). "Credibility assessment of online information in context". In : *Information Research* 15.3. paper 445. url : http://InformationR.net/ir/15-3/paper445.html.

Sattar, Fariha et al. (2020). "ClaimEval : Integrated Framework for Joint Source Credibility Estimation and Claim Evaluation". In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 34. 01, p. 1024-1031. doi : 10.1609/aaai.v34i01.5456.

Schema.org Community (s. d.). *ClaimReview - schema.org Type*. Schema.org Documentation. Accessed April 2025. url : https://schema.org/ClaimReview.

Shah, Bhushan Santosh, Deven Santosh Shah et Vahida Attar (2025). *Decoding News Bias : Multi Bias Detection in News Articles*. v1, 5 Jan 2025. arXiv : 2501.02482 [cs.CL].

Sharma, Karishma et al. (2019). "Combating Fake News : A Survey on Identification and Mitigation Techniques". In : *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.3, p. 1-42. doi : 10.1145/3305260.

Souza, Daniel F. de et al. (2020). "A survey on the automation of fake news detection". In : *Information Fusion* 62, p. 1-26. doi : 10.1016/j.inffus.2020.04.002.

Thibault, Camille et al. (2025). *A Guide to Misinformation Detection Data and Evaluation*. v2, 19 Mar 2025. arXiv : 2411.05060 [cs.SI].

Références VI

Viviani, Marco et Gabriella Pasi (2017). "Credibility in social media : opinions, news, and health information—a survey". In : *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery* 7.5, e1209. doi : 10.1002/widm.1209.

W3C Credibility Community Group (oct. 2018). *Credibility Signals (Draft Community Group Report)*. W3C Community Group Draft Report. url : https://www.w3.org/2018/10/credibility-tech/.

Willems, Reinout (jan. 2025). "Modeling Thematic Coherence : An Interpretable Approach for Analyzing Fake and LLM-Generated News". MSc AI Thesis. Mém. de mast. Utrecht, The Netherlands : Utrecht University.

Zhou, Xinyi et Reza Zafarani (2020). "A Survey of Fake News : Fundamental Theories, Detection Methods, and Opportunities". In : *ACM Computing Surveys (CSUR)* 53.5, p. 1-40. doi : 10.1145/3395046.

Merci de votre attention !

Discussion