Modélisation d'un Système d'Évaluation de la Crédibilité de l'Information

Projet 1 de modélisation, DIC-9251

Dominique S. Loyer

UQAM
loyer.dominique@courrier.uqam.ca

28 avril 2025 (Présentation)

Plan de la Présentation

Contexte : La Surcharge Informationnelle

- Augmentation massive de l'information ("infobésité").
- Difficulté croissante à distinguer l'information fiable.
- Polarisation des opinions alimentée par la désinformation. Metzger et al. (12)
- Nécessité d'outils pour aider à l'évaluation critique.

Le Problème : Désinformation et "Fake News"

- Prolifération de désinformation, mésinformation et "fake news". Oshikawa et al. (14); Zhou and Zafarani (22); Pescuma et al. (15)
- Impact négatif sur l'opinion publique, la confiance et la démocratie. Loth et al. (10); Zhou and Zafarani (22)
- Illustration tragique lors de la pandémie de COVID-19. Loth et al. (10); Chen and Shu (4); Metzger et al. (12)

Un Nouveau Défi : L'IA Générative

- Émergence de l'Intelligence Artificielle Générative (GenAI).Loth et al. (10); Chen and Shu (4)
- Création facile et rapide de contenus synthétiques réalistes. Loth et al. (10); Willems (21); Thibault et al. (19)
- Exacerbation de la crise informationnelle.
- Difficulté de distinction entre contenu authentique et synthétique.

Le Défi de l'Évaluation de la Crédibilité

- Manque de transparence sur l'origine et la fiabilité des sources. Metzger et al. (12); W3C Credibility Community Group (20); Pescuma et al. (15); Rieh (16)
- Évaluation complexe nécessitant temps et compétences spécifiques.
- Nature multidimensionnelle de la crédibilité (fiabilité, expertise, exactitude, objectivité, qualité, actualité). Pescuma et al. (15); Metzger et al. (12)

Solution Proposée : Système Hybride d'Évaluation

- **Objectif :** Fournir des outils (métriques, analyse sources, opinions divergentes) pour aider l'utilisateur à développer son esprit critique.
- **Approche Hybride** Ahmed et al. (1) :
  - **Règles Logiques Prédéfinies :** Encodage de connaissances explicites (réputation source, marqueurs linguistiques) Kedzie et al. (8); W3C Credibility Community Group (20).
  - **IA / Traitement Langage Naturel (NLP) :** Analyse de nuances (sentiment, cohérence, biais), adaptation Ahmed et al. (1); Loth et al. (10); Chen and Shu (4).
- **Motivation :** Pallier les limites des systèmes purement algorithmiques face à la complexité et la contextualité Nabozny et al. (13).
- **Buts Spécifiques :** Lutter contre infobésité, détecter désinformation, évaluer sources, présenter métriques simplement, encourager pensée critique, assurer traçabilité, contribuer à la transparence.

Acteurs du Système

Table – Acteurs du Système d'Évaluation de Crédibilité.

Acteur	Description	Buts Principaux
Utilisateur	Personne souhaitant vérifier la crédibilité.	Soumettre requête, Consulter rapport, Fournir feedback.
Expert	Personne qualifiée (config. système).	Ajuster règles/paramètres, Analyser logs.
Système Externe	Source de données tierce (API, DB).	Fournir données brutes, Fournir métadonnées.
Système	Le système d'évaluation.	Traiter requêtes, Interroger syst. ext., Appliquer règles/IA, Générer rapports.

L'acteur "Expert" est crucial pour la maintenance et l'adaptation continue face à l'évolution des menaces Nabozny et al. (13); Loth et al. (10); Chen and Shu (4).

Diagramme des Cas d'Utilisation

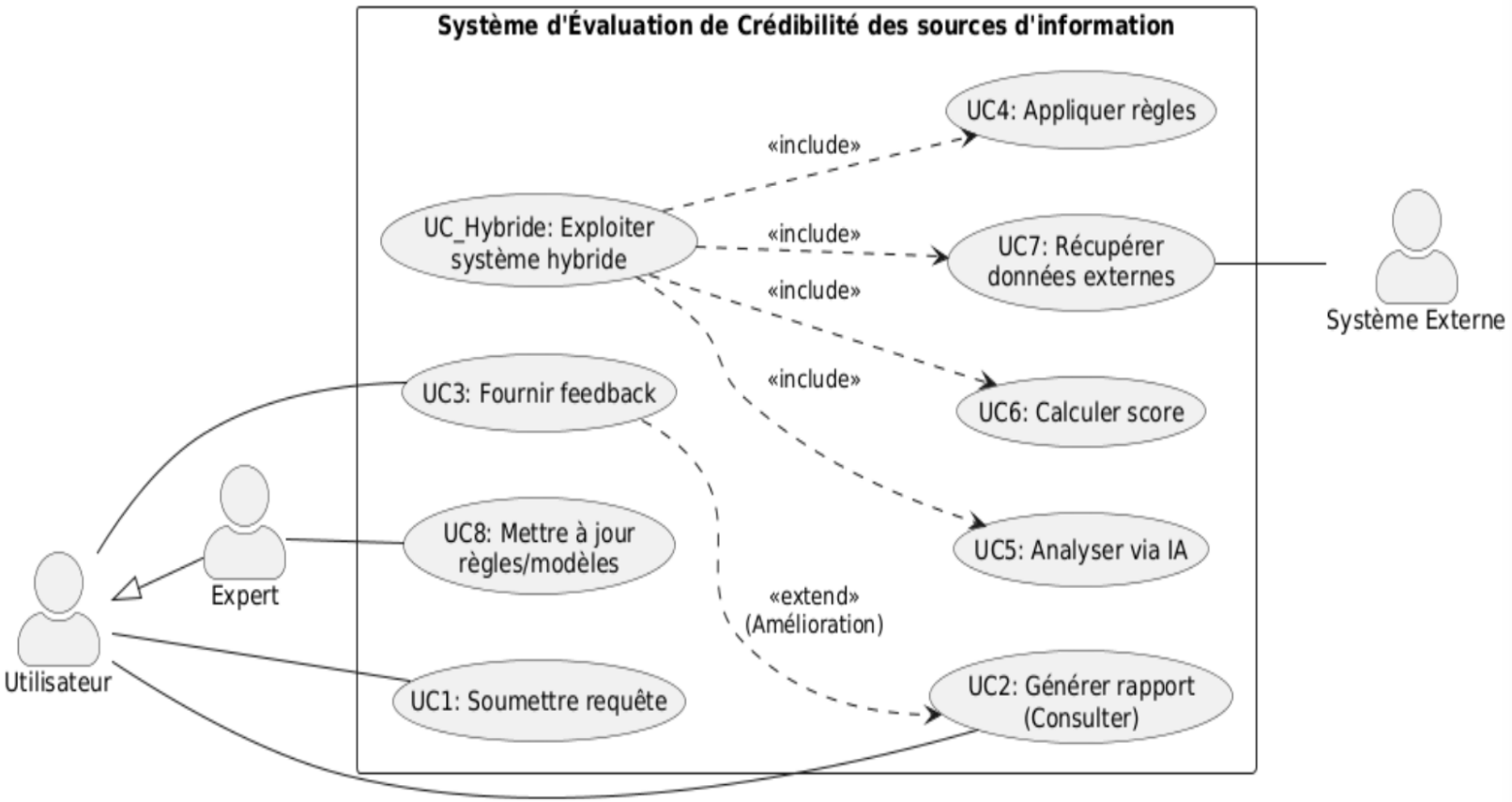

Figure – Diagramme des cas d'utilisation. * Note : UC = Use Case (Cas d'Utilisation).

Explication des Cas d'Utilisation Clés (1/2)

UC1 : Soumettre une requête de vérification
But : Permettre à l'Utilisateur de demander une évaluation.
Acteurs : Utilisateur (primaire), Système, Système Externe (secondaires).
Déroulement :

1. Utilisateur saisit/colle info (texte, URL).
2. Utilisateur soumet.
3. Système reçoit et déclenche le traitement hybride.
4. Système présente le rapport (via UC2).

UC2 : Générer un rapport de crédibilité
But : Présenter les résultats de manière claire.
Acteurs : Système (primaire), Utilisateur (secondaire).
Déroulement :

1. Système collecte résultats (score, détails, sources...).
2. Système formate le rapport (structuré, lisible, explications).
3. Système présente à l'Utilisateur.

Explication des Cas d'Utilisation Clés (2/2)

UC3 : Récupérer des données externes
But : Collecter informations pertinentes (contexte, vérifications).
Acteurs : Sys.(primaire), Sys.Ext. (API Moteur Recher., LLM, «Faits»...).
Déroulement :

1. Système identifie données nécessaires.
2. Système envoie requêtes aux API externes Google (6, 7).
3. Systèmes Externes retournent données.
4. Système prétraite et stocke temporairement.

Points Critiques : Gestion erreurs API, timeouts, limites de taux.

UC8 : Mettre à jour les règles/modèles
But : Permettre à l'Expert d'améliorer/adapter le système.
Acteurs : Expert (primaire), Système (secondaire).
Déroulement :

1. Expert accède à l'interface admin.
2. Expert modifie/ajoute règles ou MAJ/réentraîne modèles IA.
3. Expert teste les modifications.
4. Expert déploie.

Importance :pour maintenabilité face aux nouvelles menaces Nabozny et al

Explication du Cas d'Utilisation Interne : UC4

UC4 : Appliquer les règles
But : Appliquer les règles logiques prédéfinies par l'Expert pour une première évaluation rapide.
Acteur(s) Primaire(s) : Système (via UC_Hybride).
Acteur(s) Secondaire(s) : (Aucun acteur externe direct).
Relation : Inclus (‘«include»‘) dans UC_Hybride.
Description :

- Le système exécute les ‘RegleVerification‘ configurées sur les ‘DonneesBrutes‘ récupérées (issues de UC7).
- Produit des ‘ResultatRegle‘ (ex : score partiel basé sur la source, détection de langage suspect).
- Ces résultats sont utilisés par UC6 pour le calcul du score final.

Explication du Cas d'Utilisation Interne : UC5

UC5 : Analyser via IA
But : Exécuter des analyses sémantiques et contextuelles via des modèles IA/NLP.
Acteur(s) Primaire(s) : Système (via UC_Hybride).
Acteur(s) Secondaire(s) : (Aucun acteur externe direct).
Relation : Inclus (‘«include»‘) dans UC_Hybride.
Description :

- Le système utilise les ‘ModeleIA‘ configurés (sentiment, cohérence, biais, NER...) sur les ‘DonneesBrutes‘.
- Produit un ‘ResultatNLP‘ contenant les analyses détaillées.
- Ces résultats sont utilisés par UC6 pour affiner le score de crédibilité.

Explication du Cas d'Utilisation Interne : UC6

UC6 : Calculer le score
But : Combiner les résultats des analyses pour générer un score de crédibilité global et des métriques détaillées.
Acteur(s) Primaire(s) : Système (via UC_Hybride).
Acteur(s) Secondaire(s) : (Aucun acteur externe direct).
Relation : Inclus (‘«include»‘) dans UC_Hybride.
Description :

- Le système agrège les ‘ResultatRegle‘ (de UC4), le ‘ResultatNLP‘ (de UC5) et potentiellement l'analyse des ‘InfoSource‘ (issues de UC7).
- Applique un algorithme de pondération ou de combinaison (défini par l'Expert via UC8) pour calculer le ‘scoreCredibilite‘.
- Prépare les ‘detailsScore‘ et autres informations pour le ‘RapportEvaluation‘ (généré par UC2).

Explication du Cas d'Utilisation Interne : UC7

UC7 : Récupérer des données externes
But : Collecter les informations pertinentes depuis les sources externes nécessaires à l'analyse.
Acteur(s) Primaire(s) : Système (via UC_Hybride).
Acteur(s) Secondaire(s) : Système Externe (API Moteur Recherche, API LLM, API Fact-checking...).
Relation : Inclus (‘«include»‘) dans UC_Hybride.
Description :

- Le système identifie les données nécessaires (contenu URL, contexte, faits...).
- Interroge les API des ‘SystemeExterne‘ configurés.
- Reçoit et prétraite les ‘DonneesBrutes‘ pour les analyses UC4 et UC5.
- Gère les erreurs, timeouts et limites des API externes.

Diagramme de Classes

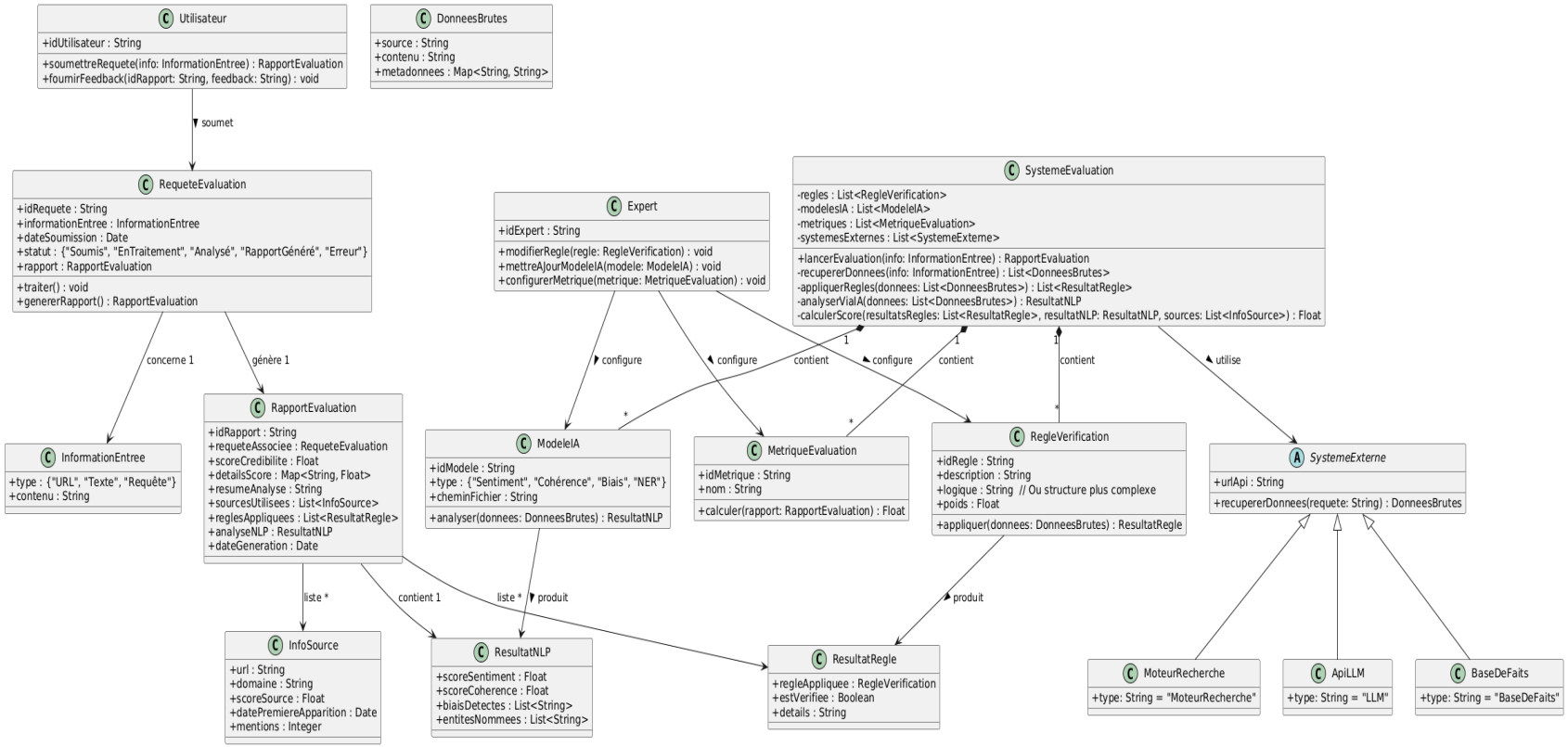

Figure – Structure statique du système.

Points Clés : **Orchestration centrale (‘SystemeEvaluation‘), gestion requêtes (‘RequeteEvaluation‘), extensibilité sources externes (‘SystemeExterne‘), approche hybride explicite (‘RegleVerification‘, ‘ModeleIA‘), rapport riche (‘RapportEvaluation‘), configuration par l’Expert. Fowler (5); Larman (9).**

Synthèse du Diagramme de Classes

Table – Résumé des Classes Principales et Rôles.

Classe Principale	Rôle / Responsabilité	Relations Clés
‘SystemeEvaluation‘	Orchestre l’évaluation hybride	Contient ‘RegleVerification‘, ‘ModeleIA‘, ‘SystemeExterne‘. Utilise ‘RequeteEvaluation‘, ‘RapportEvaluation‘.
‘RequeteEvaluation‘	Représente la demande de l’utilisateur	Concerne ‘InformationEntree‘, Génère ‘RapportEvaluation‘.
‘RapportEvaluation‘	Contient les résultats détaillés de l’évaluation	Contient ‘ResultatRegle‘, ‘ResultatNLP‘, ‘InfoSource‘. Associé à ‘RequeteEvaluation‘.
‘RegleVerification‘	Applique une règle logique spécifique	Produit ‘ResultatRegle‘. Configurée par ‘Expert‘.
‘ModeleIA‘	Exécute une analyse IA/NLP	Produit ‘ResultatNLP‘. Configuré par ‘Expert‘.
‘SystemeExterne‘	Accès aux données/API externes (Classe abstraite)	Utilisé par ‘SystemeEvaluation‘. Spécialisé en ‘MoteurRecherche‘, ‘ApiLLM‘, etc.

Diagramme de Séquence : Scénario 1 (Vérification URL)

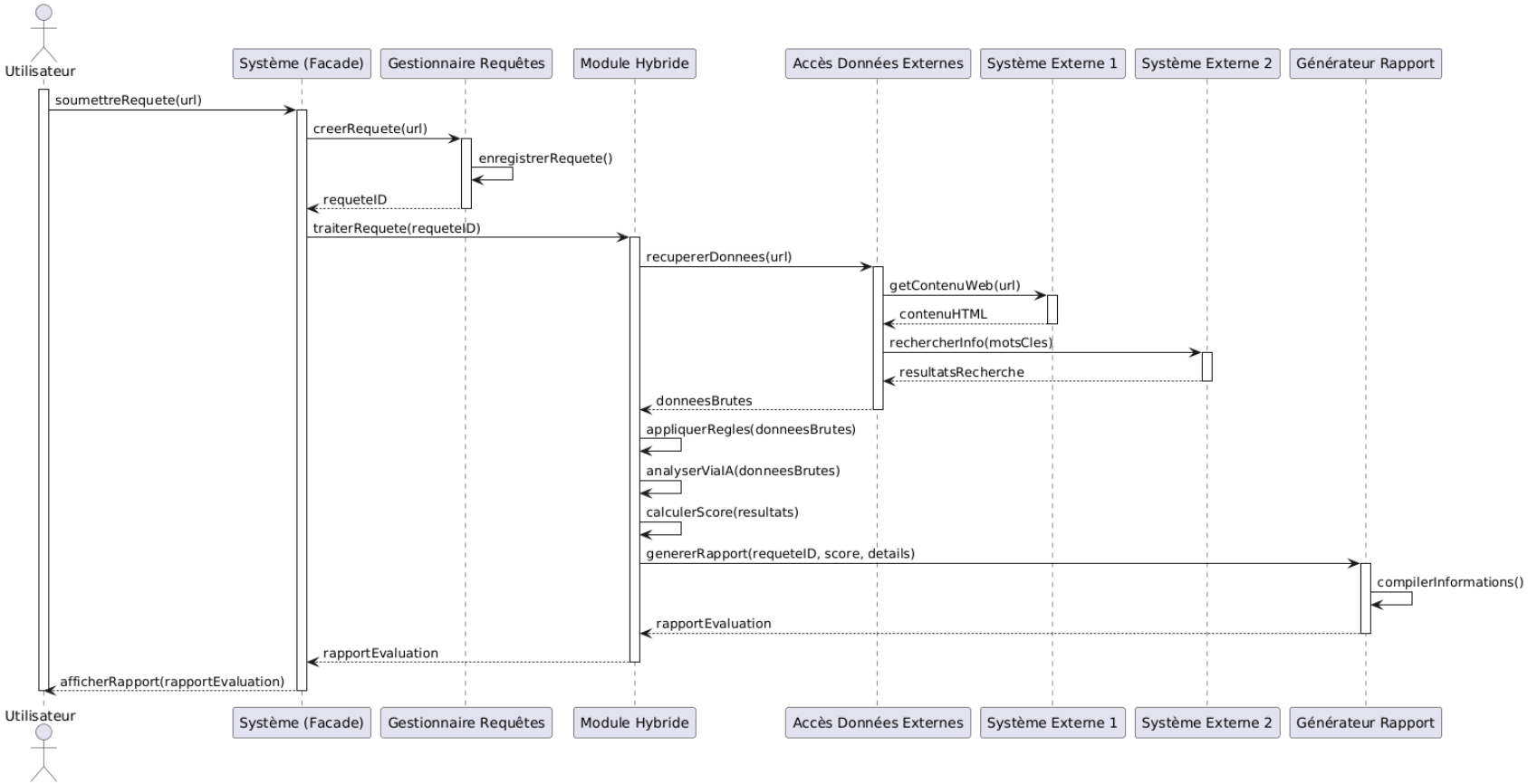

Figure – Interaction pour une évaluation simple par l'utilisateur.

Illustre l'orchestration, l'appel à l'approche hybride, la récupération de données externes ('SE1', 'SE2') Earman (0), et la génération du rapport. La gestion des erreurs API est cruciale.

Diagramme de Séquence : Scénario 2 (MàJ Règle Expert)

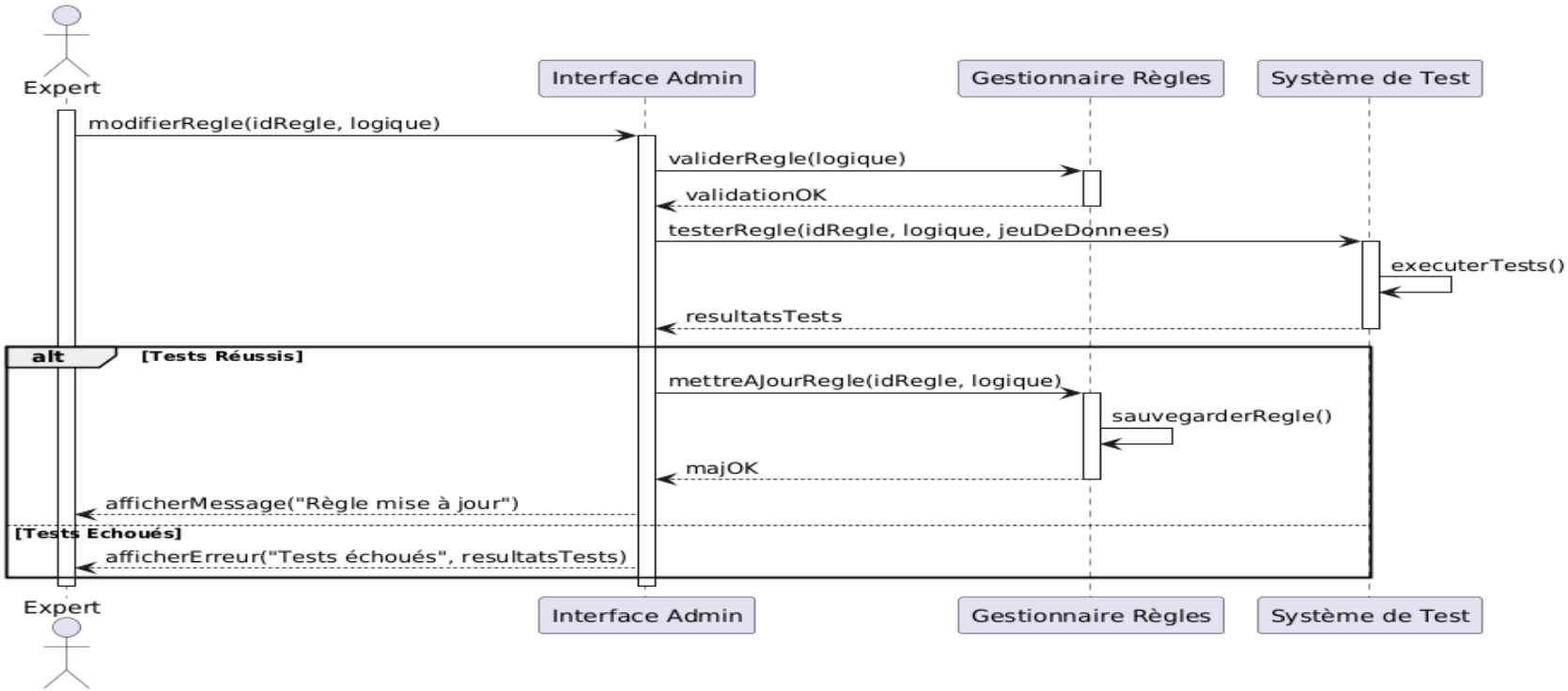

Figure – Interaction pour la mise à jour d'une règle par l'expert (UC8).

Montre l'accès via UI admin, la validation, le test optionnel mais recommandé, et la sauvegarde. Souligne l'importance d'un cycle de validation.

Synthèse des Diagrammes de Séquence

Table – Résumé des Scénarios d'Interaction.

cccc

Scénario	Objectif Principal	Participants Clés	Interaction Clé
Scénario 1 (Vérif. URL)	Traiter une requête utilisateur de A à Z	Utilisateur, Façade, Module Hybride, Accès Données, Syst. Externes, Gén. Rapport	Orchestration du flux : requête -> récup. données -> analyse hybride -> rapport.
Scénario 2 (MàJ Règle)	Mettre à jour une règle par l'Expert (UC8)	Expert, Interface Admin, Gest. Règles, Syst. Test	Processus contrôlé : modif -> validation -> test (avec 'alt') -> sauvegarde.

Diagramme d'États/Transitions (Cycle de vie d'une Requête)

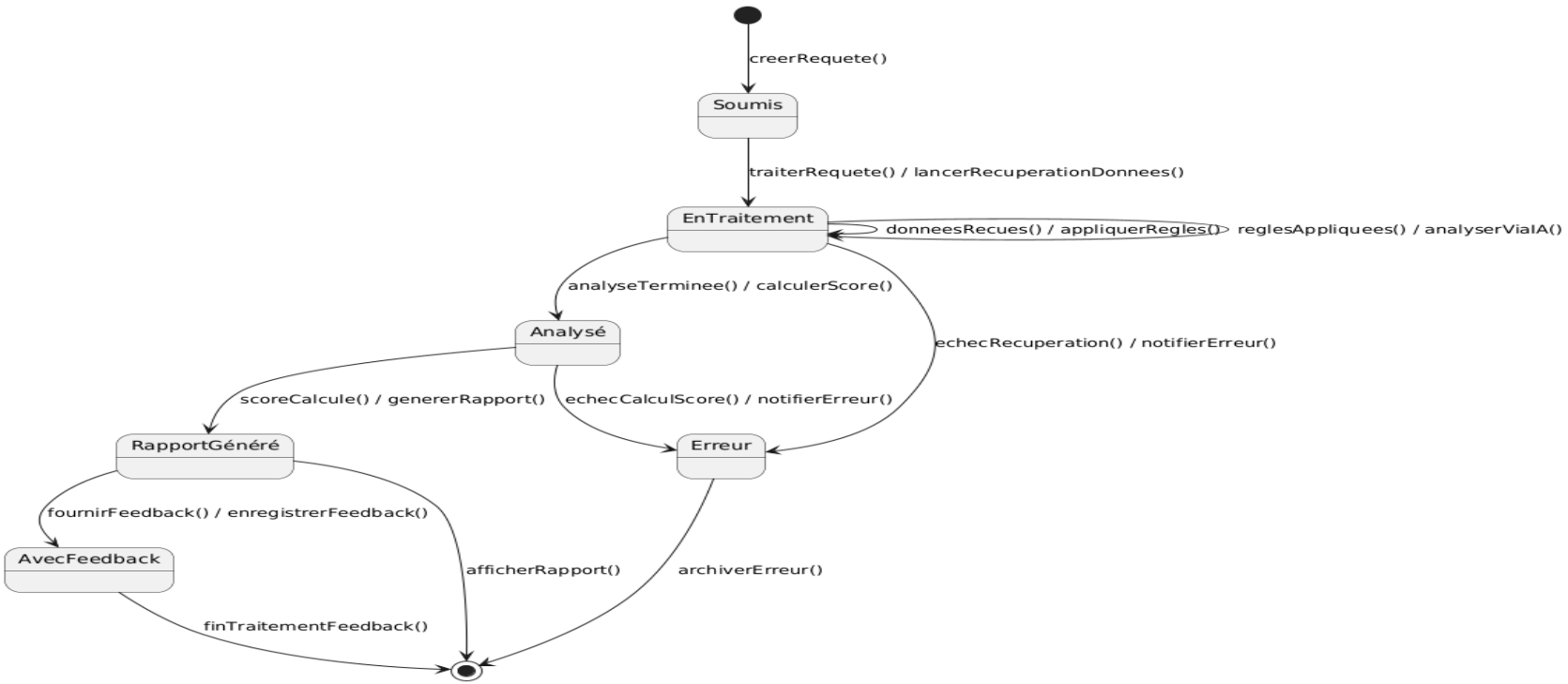

Figure – États possibles d'une 'RequeteEvaluation' Larman (9).

États : **Soumis** → **EnTraitement** → **Analysé** → **RapportGénéré** → **[AvecFeedback]**. **État final alternatif** : **Erreur**.

Synthèse du Diagramme d'États/Transitions ('RequeteEvaluation')

Table – Résumé du Cycle de Vie d'une Requête.

État	Description	Transitions Sortantes Principales (Événement / Action)
'Soumis'	Requête créée, en attente de traitement.	'traiterRequete()' / 'lancerRecuperationDonnees()' → 'EnTraitement'
'EnTraitement'	Récupération données, analyses règles/IA en cours.	'analyseTerminee()' / 'calculerScore()' → 'Analysé' 'echec...()' / 'notifierErreur()' → 'Erreur'
'Analysé'	Analyses terminées, score calculé/en calcul.	'scoreCalcule()' / 'genererRapport()' → 'RapportGénéré' 'echecCalculScore()' / 'notifierErreur()' → 'Erreur'
'RapportGénéré'	Rapport prêt pour consultation.	'fournirFeedback()' / 'enregistrerFeedback()' → 'AvecFeedback'
'AvecFeedback'	Feedback utilisateur reçu.	'finTraitementFeedback()' → (Fin)
'Erreur'	Échec irrécupérable du traitement.	'archiverErreur()' → (Fin)

Table : Indicateurs de Crédibilité et Analyses

Table – Indicateurs Clés et Techniques d'Analyse Associées.

Indicateur de Crédibilité	Techniques d'Analyse Associées
Exactitude / Facticité **Pescuma et al. (15)**	- Vérification via API Fact-Checking (ClaimReview) (author?) (Schema.org Community); Google (7) - NER + Interrogation bases de connaissances / Moteurs recherche Oshikawa et al. (14) - Analyse présence/qualité citations Kedzie et al. (8)
Autorité / Réputation Source **Pescuma et al. (15)**	- Consultation bases réputation (via règles, ex : MediaBias) W3C Credibility Community Group (20) - Analyse historique source (règles : ancienneté domaine...) - NER pour identifier experts/institutions cités Oshikawa et al. (14)
Objectivité / Biais **Pescuma et al. (15)**	- Modèles Détection Biais (Politique, Genre...) Shah et al. (18); Menzner and Leidner (11); Baly et al. (2) - Analyse Sentiment Contextualisée (ton manipulateur) Ahmed et al. (1) - Analyse diversité points de vue (via infos externes)
Présentation / Qualité **Pescuma et al. (15)**	- Analyse Cohérence (Locale/Globale) Willems (21); Barzilay and Lapata (3) - Règles détection langage sensationnaliste Kedzie et al. (8) - Analyse stylométrique Ahmed et al. (1) - Détection texte généré par IA Willems (21); Kedzie et al. (8)
Actualité **Pescuma et al. (15)**	- Règles cohérence dates - Signalement infos obsolètes
Persuasion / Fallacies **Pescuma et al. (15); Kedzie et al. (8)**	- Règles détection patterns (émotion excessive...) Kedzie et al. (8) - Analyse Sentiment (manipulation émotionnelle) Ahmed et al. (1)

Table : Exemple Performance Détection Biais

Table – Exemple Illustratif de Performance (F1-Score) de Modèles Transformers pour la Détection de Biais Multiples (Basé sur Shah et al. (18)).

Type de Biais	BERT	RoBERTa	ALBERT	DistilBERT	XLNet
Politique	0.89	0.87	0.85	0.84	0.86
Genre	0.82	0.80	0.75	0.73	0.76
Entité	0.85	0.84	0.81	0.80	0.82
Racial	0.65	0.62	0.55	0.38	0.51
Religieux	0.78	0.77	0.72	0.70	0.74
Régional	0.70	0.68	0.63	0.59	0.65
Sensationnalisme	0.80	0.79	0.74	0.71	0.75
Moyen (Macro)	0.78	0.77	0.72	0.68	0.73

Note : Valeurs illustratives. Performances réelles dépendent du dataset, fine-tuning, et gestion déséquilibre des classes (ex : Racial).

Références I

[1] Ahmed, T., Traore, I., and Saad, S. (2024). AI-Driven Hybrid Model for Fake News Detection : Integrating NLP, Sentiment Analysis, and Source Credibility to Combat Misinformation. *ResearchGate (Preprint)*. Accessed April 2025, DOI might be available later.

[2] Baly, R., Da San Martino, G., Glass, J., and Nakov, P. (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)*, pages 6785–6791.

[3] Barzilay, R. and Lapata, M. (2008). Modeling local coherence : An entity-based approach. *Computational Linguistics*, 34(1) :1–34.

[4] Chen, K. and Shu, K. (2023). Combating misinformation in the age of llms : Opportunities and challenges.

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

[6] Google. Custom search json api. Google Developers Documentation. Accessed April 2025.

[7] Google. Fact check tools api. Google Developers Documentation. Accessed April 2025.

[8] Kedzie, C., McKeown, K., and Diaz, F. (2018). Content-driven detection of false news. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1771–1783, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

[10] Loth, A., Kappes, M., and Pahl, M.-O. (2024). Blessing or curse ? a survey on the impact of generative ai on fake news. v2, 27 Dec 2024.

[11] Menzner, P. and Leidner, J. L. (2024). Biasscanner : Detecting biased statements in news articles.

[12] Metzger, M. J., Flanagin, A. J., and Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3) :413–439.

Références II

[13] Nabozny, A., Marcinkiewicz, M., Nielek, R., and Wierzbicki, A. (2021). Active annotation in evaluating the credibility of web-based medical information : Guidelines for creating training data sets for machine learning. *JMIR Medical Informatics*, 9(11) :e26065.

[14] Oshikawa, R., Qian, J., and Wang, W. Y. (2020). A survey on natural language processing for fake news detection. *arXiv preprint arXiv :1811.00770*. v3.

[15] Pescuma, V. N., Osborne, F., Recupero, D. R., and Motta, E. (2025). Source credibility assessment : A comprehensive survey. In *International Journal of Interactive Multimedia and Artificial Intelligence*, volume TODO, page TODO.

[16] Rieh, S. Y. (2010). Credibility assessment of online information in context. *Information Research*, 15(3). paper 445.

[Schema.org Community] Schema.org Community. Claimreview - schema.org type. Schema.org Documentation. Accessed April 2025.

[18] Shah, B. S., Shah, D. S., and Attar, V. (2025). Decoding news bias : Multi bias detection in news articles. v1, 5 Jan 2025.

[19] Thibault, C., Tian, J.-J., Peloquin-Skulski, G., Curtis, T. L., Zhou, J., Laflamme, F., Guan, Y., Rabbany, R., Godbout, J.-F., and Pelrine, K. (2025). A guide to misinformation detection data and evaluation. v2, 19 Mar 2025.

[20] W3C Credibility Community Group (2018). Credibility signals (draft community group report). W3C Community Group Draft Report.

[21] Willems, R. (2025). Modeling thematic coherence : An interpretable approach for analyzing fake and llm-generated news. Master's thesis, Utrecht University, Utrecht, The Netherlands. MSc AI Thesis.

[22] Zhou, X. and Zafarani, R. (2020). A survey of fake news : Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5) :1–40.

Discussion et Questions

Merci de votre attention !

Questions ?

Modélisation d'un Système d'Évaluation de la Crédibilité de l'Information

Projet 1 de modélisation, DIC-9251

Dominique S. Loyer

UQAM
loyer.dominique@courrier.uqam.ca

28 avril 2025 (Présentation)

Prototype

Interface

Figure – L'interface

Le système

```
--- Vérification pour : This post is verified and credible. Avoid false information.... ---
Texte nettoyé (extrait) : this post is verified and credible. avoid false information....
[Simulation] Recherche de données externes pour : this post is verified and credible. avoid false i
n...
Données externes (simulées) : {'fact_checks': [], 'source_reputation': 'Unknown', 'domain_age_days': N
one, 'related_articles': [{'title': 'Related Story A', 'url': 'http://example.com/a'}, {'title': 'Rela
ted Story B', 'url': 'http://example.com/b'}]}
Résultats règles : {'linguistic_markers': {'sensationalism': 0, 'certainty': 2, 'doubt': 1}, 'source_a
nalysis': {'reputation': 'Unknown', 'domain_age_days': None}, 'timeliness_flags': [], 'fact_checking':
[]}
Résultats NLP (Sentiment): {'label': 'POSITIVE', 'score': 0.998832643032074}
Résultats NLP (Bias): {'score': 0.5240875482559204, 'label': 'Low Bias Detected (Simulated)'}
Résultats NLP (NER count): 0
Score global calculé : 0.37
```

Figure – Exemple 1 avec du texte

Le système 2

```
--- Vérification pour : http://hoax-site.org/the-truth... ---
[Simulation] Récupération du contenu de : http://hoax-site.org/the-truth
Texte nettoyé (extrait) : shocking conspiracy revealed! experts are wrong. this is a hoax!...
[Simulation] Recherche de données externes pour : http://hoax-site.org/the-truth...
Données externes (simulées) : {'fact_checks': [{'claim': 'Conspiracy theory', 'rating': 'False'}], 'so
urce_reputation': 'Low', 'domain_age_days': 90, 'related_articles': [{'title': 'Related Story A', 'ur
l': 'http://example.com/a'}, {'title': 'Related Story B', 'url': 'http://example.com/b'}]}
Résultats règles : {'linguistic_markers': {'sensationalism': 3, 'certainty': 0, 'doubt': 1}, 'source_a
nalysis': {'reputation': 'Low', 'domain_age_days': 90}, 'timeliness_flags': ['Source domain is relativ
ely new.'], 'fact_checking': [{'claim': 'Conspiracy theory', 'rating': 'False'}]}
Résultats NLP (Sentiment): {'label': 'NEGATIVE', 'score': 0.9988172650337219}
Résultats NLP (Bias): {'score': 0.5322145223617554, 'label': 'Low Bias Detected (Simulated)'}
Résultats NLP (NER count): 0
Score global calculé : 0.05
```

Figure – Exemple 2 avec une URL connue pour un Hoax

Le système 3

```
--- Vérification pour : http://verified-news.com/article123... ---
[Simulation] Récupération du contenu de : http://verified-news.com/article123
Texte nettoyé (extrait) : this official report is verified and credible. all facts checked....
[Simulation] Recherche de données externes pour : http://verified-news.com/article123...
Données externes (simulées) : {'fact_checks': [{'claim': 'Official report facts', 'rating': 'True'}],
'source_reputation': 'High', 'domain_age_days': 1500, 'related_articles': [{'title': 'Related Story
A', 'url': 'http://example.com/a'}, {'title': 'Related Story B', 'url': 'http://example.com/b'}]}
Résultats règles : {'linguistic_markers': {'sensationalism': 0, 'certainty': 3, 'doubt': 0}, 'source_a
nalysis': {'reputation': 'High', 'domain_age_days': 1500}, 'timeliness_flags': [], 'fact_checking':
[{'claim': 'Official report facts', 'rating': 'True'}]}
Résultats NLP (Sentiment): {'label': 'POSITIVE', 'score': 0.999338686466217}
Résultats NLP (Bias): {'score': 0.5220904350280762, 'label': 'Low Bias Detected (Simulated)'}
Résultats NLP (NER count): 0
Score global calculé : 0.88
```

Figure – Exemple 3 (sources connue)

Le système à base de règles de prédicats

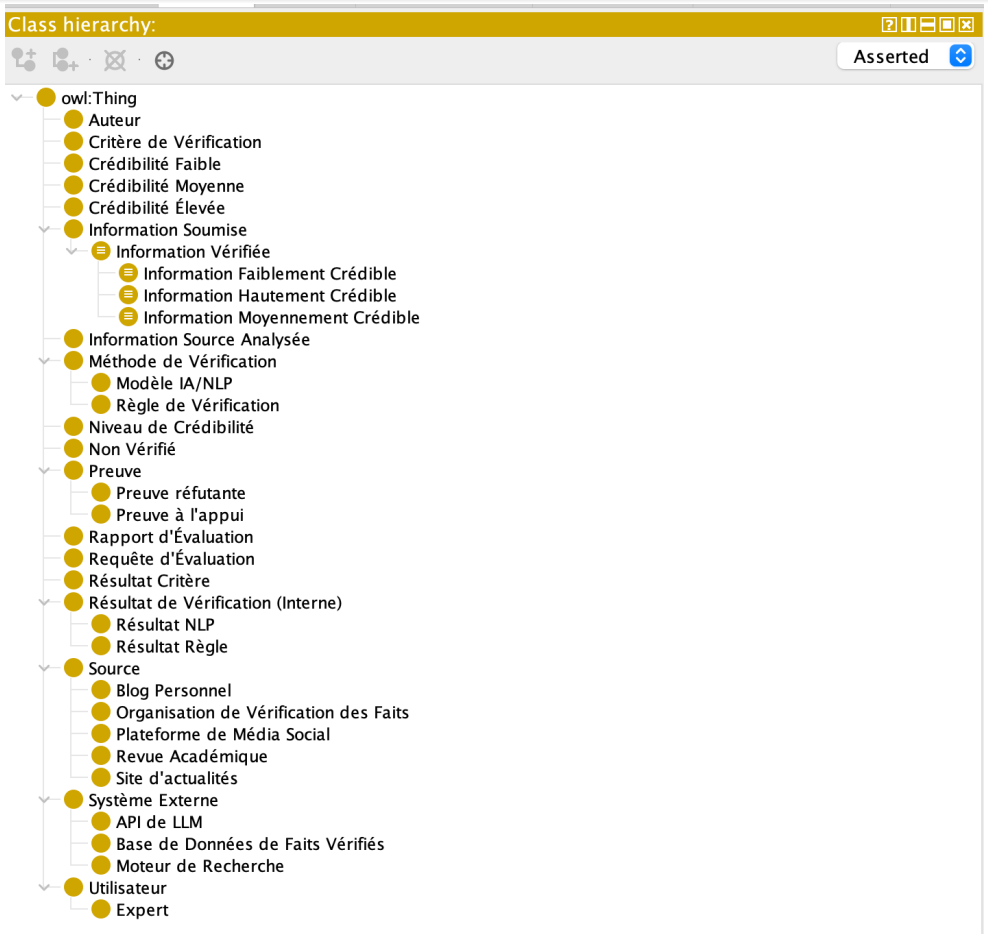

Figure – Les classes dans le système à base de règle de prédicat