Cahier des charges

Travail de Bachelor 2020

Original Journalistic Information Label and Traceability

Etudiant : Nicolas Solioz

Professeure : Nicole Glassey Balet

HES-SO Valais-Wallis – rue de la Plaine 2 – 3960 Sierre
+41 27 606 89 11 – info@hevs.ch – www.hevs.ch

Table des matières

Introduction

Ce travail de Bachelor s'effectue en complément du projet de recherche « Original Journalistic Information Label and Traceability » actuellement en étude à l'institut informatique de gestion. Il représente la suite du travail de Bachelor réalisé en 2019 par Nicolas Piguet « Vizualising the spread of Fake News on social media ».

Contexte

Les avancées technologiques nous donnent accès à une richesse d'information historiquement inégalable. La popularisation des réseaux sociaux comme Twitter et Facebook a contribué à cette explosion du partage d'information. Or, dans cet océan de données se cachent des fausses informations. La facilité accrue de devenir créateur de contenu donne la possibilité à n'importe quel quidam de publier des informations qui ne sont pas avérées. La machinisation de la création de fausses informations ou de « fake news » a été observée notamment lors des élections américaines de 2016. En effet, plusieurs faux profils de républicains ont été créés par des russes sur Twitter afin de propager des fausses informations concernant la course présidentielle américaine (Watts, 2017). Cette stratégie s'est montrée efficace, car si le cerveau humain reçoit plusieurs fois la même information, il l'interprète comme étant véridique (Vandal, 2019). Ainsi dans cette époque où le partage de fausses informations est devenu dangereusement simple, comment peut-on s'assurer que le texte qu'on lit est de confiance ?

La solution proposée par ce travail est la suivante : mettre en place un label certifiant l'originalité et la qualité d'un contenu journalistique. L'idée est d'offrir aux médias un outil qui vérifie si leurs articles répondent à certains critères journalistiques : qualité du contenu, originalité, provenances des sources etc. Si l'article répond à toutes les exigences de l'outil, il obtient une certification cryptée et non duplicable. Comme les livres et leurs ISBN, une fois certifiées les articles auraient leur propre identifiant unique.

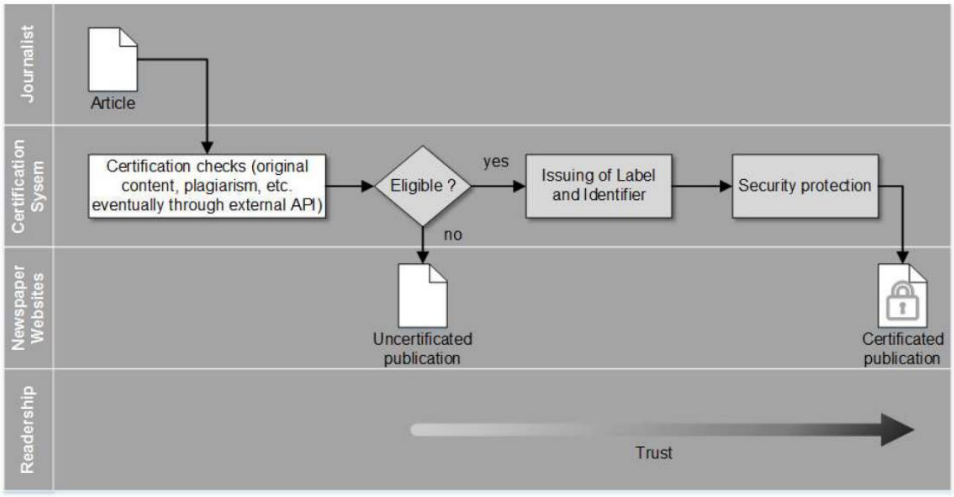

Figure 1 - Processus envisagé de certification et publication labellisée (Institut Informatique Gestion)

L'intérêt pour la mise en place d'un tel label a été démontré dans un rapport sur l'avenir des médias partagé par la Commission fédérale des médias (COFEM, 2020). Dans les « focus groupe » menés par l'institut informatique de gestion, le lectorat a également montré un intérêt pour la labellisation des articles de presse (Institut Informatique de Gestion, 2020). Ainsi, la mise en place d'un tel outil répondrait à un besoin établi.

Problématiques

Critères d'analyse

Afin de déterminer quels articles répondent aux critères d'un contenu journalistique de qualité, il faut établir les conditions d'obtention du label. Lors des focus groupe menés par l'institut informatique de gestion, nous avons relevé que plusieurs éléments contribuent à la perception de la qualité d'un article :

- Présence des sources
- Présence d'un chapô, d'un titre et d'une image
- Respect des droits d'auteur
- Mention d'un organisme étatique
- Originalité du contenu
- Réputation de l'auteur
- Etc…

L'originalité du contenu peut être déterminé avec des outils de détection de plagiat. Or, certains de ces outils ne sont pas gratuits ce qui peut empêcher des journalistes de proactivement chercher la labellisation. De plus, avant de choisir un outil de détection de plagiat, il faut analyser la qualité de cet outil et son objectivité. Nous devons également considérer qu'un texte d'une langue étrangère peut être plagié et traduit. Finalement, nous devons décider du seuil de plagiat que nous autorisons

pour la labellisation. Si un journal mensuel décide de parler d'un événement déjà publié dans un quotidien, il est possible que son score de plagiat soit élevé car certains mots clés se répéteront.

La difficulté de ce travail consistera également à sélectionner d'autres critères d'analyse et leur donner un degré de pondération objectif et facilement quantifiable. Cela sera problématique par exemple pour la réputation d'un auteur. Nous pouvons imaginer lier un score de confiance à un auteur si celui-ci possède déjà plusieurs articles labellisés, mais cela empêcherait les nouveaux auteurs d'obtenir la certification ce qui peut être un frein au déploiement de cet outil.

De plus, la pondération des critères doit être étudiée. Si nous sélectionnons plusieurs critères, il est vital de déterminer lesquels sont essentiels à l'obtention du label. Nous pouvons décider par exemple que l'originalité du contenu est plus importante que la présence d'un chapô dans l'article.

Ces critères doivent être étudiés et décidés en partenariat avec différents partenaires issus du milieu des médias et des universités journalistiques et cela afin de mettre en place un système qui fonctionne pour tous les articles, peu importe la répartition démographique du lectorat.

Sécurité

Le but principal de ce projet est d'offrir un outil permettant d'améliorer la confiance du lectorat. Nous devons donc mettre en place un outil qui lui aussi est digne de confiance. Ainsi, le besoin de sécuriser chaque étape de la labellisation est nécessaire afin d'empêcher une utilisation frauduleuse de la certification. Nous pouvons déterminer quelques risques potentiels :

- Risque d'accès à la base de données : En ayant accès à la base de données contenant les clés de certification des articles, n'importe quelle personne pourrait émettre des labels. Cela éroderait la confiance du lectorat.
- Risque de réutilisation de l'image du label : Pour montrer au lectorat qu'un article est labellisé, nous allons utiliser une image. Or, cette image peut être facilement téléchargée est réutilisée dans un article non-certifiée. Afin de lutter contre ce risque, il est important de lier une image à une clé de chiffrement. Nous pouvons par exemple générer une image avec une clé unique à la fin du processus de chiffrement et offrir au lectorat la possibilité d'introduire l'URL de l'article pour déterminer si ce dernier est bien certifié.
- Risque de génération d'une clé d'identification : Accompagner l'image du label avec une clé d'identification nous exposera à un nouveau risque. En effet, si on connaît la longueur de la clé de chiffrement et son contenu, il est aisé de générer une nouvelle clé même si celle-ci ne se retrouvera pas dans la base de données.
- Risque de modification forcée de la réputation : Si nous décidons d'utiliser la réputation d'un auteur comme critère de labellisation, nous nous exposons à un risque de compétitivité malhonnête. Car, il est envisageable que des tiers exploitent notre algorithme afin de faire chuter le score de réputation d'un journal tier, l'empêchant ainsi

d'obtenir la certification. Nous devons donc faire attention à la mise place de nos critères d'analyse.

- Risque de modification d'un article : Il est fréquent que les médias éditent leurs articles déjà publiés. Il est important de traquer ces modifications afin de déterminer si l'article mis à jour mérite toujours la labellisation. Nous pouvons par exemple autoriser les petites modifications, comme pour l'édition de fautes d'orthographes ou de coquilles. Or, l'insertion de nouveaux paragraphes par exemple devrait nécessiter une nouvelle labellisation.
- Risque de vol d'identité : Le fait qu'une société de presse reconnue rédige un article améliore la confiance du lectorat. Ainsi, l'auteur du texte sera un élément nécessaire dans notre outil d'analyse. Nous devons donc offrir une accréditation et une authentification sûres afin d'empêcher des personnes malveillantes de certifier des articles sous le nom d'un tier.

La sécurité de ce projet est un aspect primordial qui devra être au cœur de la réflexion afin de développer un outil de qualité.

Base de données

L'outil de labellisation sera développé sous la forme d'une plateforme web. Afin de stocker les informations des articles certifiés, les clés d'identification et les métadonnées nécessaires, nous devons mettre en place une base de données. La mise en place de celle-ci présente plusieurs questions :

- Contenu de la base de données : Devons-nous sauvegarder la totalité du texte de l'article ? Cela peut s'avérer complexe pour des travaux de recherche comportant plusieurs pages.
- Accès multiples : Comment gérer une demande de labellisation simultanée de 2 articles identiques ?
- Sauvegarde : Nous devons garantir la sauvegarde régulière de la base de données afin d'empêcher la perte de données.
- Nouveaux médias : Un critère important d'un article est son auteur. Comme cité précédemment dans ce cahier des charges, nous devons permettre aux auteurs et aux sociétés de presse de s'identifier afin de leur permettre de certifier leurs articles. Nous devons également étudier la possibilité d'offrir à des utilisateurs non-inscrits de labelliser leur texte, où alors développer une procédure de validation d'inscription afin d'insérer de nouveaux médias dans notre base de données.

En plus des points énumérés ci-dessus, nous devons également choisir un outil de base de données nous permettant de travailler efficacement sur un outil web, comme par exemple MySQL. Le choix de base de données sera déterminé en fonction du langage de programmation de l'interface web.

Interface web

La nature du métier de journaliste est stressante. Les délais sont courts et les nouvelles méthodes de consommation font que le lectorat a besoin de l'information le plus rapidement possible. Ainsi, ajouter une étape supplémentaire au travail des journalistes n'est pas chose aisée. Afin d'encourager les journalistes à utiliser le système de labellisation, ce dernier doit être attrayant, simple d'utilisation et surtout rapide. L'expérience de l'utilisateur, la simplicité de l'interface et la rapidité de certification seront donc des éléments centraux lors du développement de l'outil. Nous réaliserons cet outil en partenariat avec des personnes du métier afin de l'adapter en fonction de leurs besoins.

Travail à effectuer

Les 5 étapes clés à réaliser ont été reprises depuis les données du travail de Bachelor, reçues le 13 février 2020.

Etape 1

« Identifier les facteurs de différenciation entre du contenu original et non-original »

et

« Analyser les indicateurs et les règles professionnelles journalistiques pour définir les critères de la vérification de l'originalité du contenu »

La première étape regroupe deux étapes proposées par les données du travail de Bachelor. Nous estimons que la différenciation du contenu original et non-original correspond à un critère de vérification d'originalité du contenu. Ainsi, nous allons étudier tous les critères en partenariat avec la presse et différentes universités afin de pouvoir mesurer efficacement la qualité d'un article. Nous allons également décider du degré de pondération de ces différents éléments.

Etape 2

« Dresser un état de l'art relatif aux outils de détection du contenu original en Français et en Allemand »

et

« Analyser les solutions existantes et élaborer une proposition d'un outil d'évaluation du contenu (éventuellement basé sur des outils ou APIs externes) »

Cette étape consiste à étudier les différents outils d'analyse de contenu proposés actuellement en ligne. Afin de choisir les outils que nous allons implémenter, nous devons décider des critères de qualité de ces derniers. Ensuite, nous allons les tester avec des échantillons similaires d'articles afin de pouvoir effectuer une comparaison adéquate. Lors de la discussion avec la professeure du 18 février 2020, nous avons décidé que la gestion de la langue allemande est secondaire. Elle sera incorporée en fonction du temps restant pour la réalisation de ce travail.

Nous allons également préparer des mockups détaillés afin de montrer comment l'interface web fonctionnera. Ceux-ci permettront de comprendre les outils APIs qui seront utilisés, les critères qui seront analysées et le format du label qui sera délivré. Ces mockups seront réalisés en partenariat avec des personnes du métier de la presse.

Etape 3

« Développer la solution dans un démonstrateur Web »

Lors de la 3ème étape, nous allons développer l'interface web ainsi que la base de données permettant de stocker les métadonnées des articles labellisés. Un « Proof of Concept » (POC) étant nécessaire pour démontrer aux partenaires médias l'intérêt de l'outil, cette étape sera celle qui demandera le plus de temps de travail. Un outil de qualité et simple d'utilisation est une carte de visite importante qui contribuera à la pérennité du projet au sein de l'institut d'informatique de gestion.

Le développement sera effectué en incorporant une philosophie « Agile » : celui du travail par itération basé sur les délivrables. Cette manière de travailler nous permettra de rapidement présenter un outil fonctionnel aux médias intéressés.

Tout l'environnement de développement sera mis en place afin de permettre à un développeur tier de facilement reprendre le code (UnitTest, commentaires, documentation etc…).

Etape 4

« Tester les fonctionnalités basées sur les données (articles) ouverts et évaluer les résultats avec l'intégration de l'approche du contrôle de la qualité »

L'étape finale du projet consistera en l'analyse des résultats et de l'efficacité de l'outil mis en place. Cela nous permettra de détecter d'éventuelles failles dans l'interface web afin de proposer des pistes d'amélioration. Cette étape sera également l'opportunité de finaliser toute la documentation du travail afin de livrer le plus de ressources possibles et faciliter la reprise et la modification de l'outil.

Planification

Sur la base des données de travail de Bachelor et compte tenu de la durée fixée dans ce contexte académique, ce travail s'effectuera lors de 11 sprints d'environ 2 semaines. Les objectifs de chaque sprint sont détaillés dans l'échéancier ci-dessous. Il faut relever que ces objectifs sont susceptibles d'évoluer en fonction de l'avancée du projet. Le travail étant estimé à 360 heures, chaque sprint dure environ 33 heures.

Compte tenu du contexte individuel du projet, certains artefacts de la méthodologie Agile ne seront pas appliqués, notamment le « Daily Meeting ». En revanche, la professeure Madame Glassey agira en rôle de « Product Owner » et participera chaque 2 semaines à une « Sprint Review » ce qui nous permettra d'évaluer le travail déjà réalisé. Nous n'allons pas immédiatement donner un délivrable puisque, comme indiqué dans le chapitre « Travail à effectuer », les premières étapes du projet ne contiendront aucune ligne de code.

La planification de chaque sprint, ainsi que son objectif, se situeront dans un fichier Excel consultable dans le repository « GitLab » du projet.

Echéancier

- Sprint 0 — Création du cahier des charges — 18.02.2020 – 08.03.2020
- Sprint 1 — Etape 1 – État de l'art (label) — 09.03.2020 – 25.03.2020
- Sprint 2 — Etape 2 – État de l'art (plagiat) — 26.03.2020 – 08.04.2020
- Sprint 3 — Etape 2 – Test des outils — 09.04.2020 – 22.04.2020
- Sprint 4 — Etape 3 - Développement — 23.04.2020 – 06.05.2020
- Sprint 5 — Etape 3 - Développement — 07.05.2020 – 20.05.2020
- Sprint 6 — Etape 3 - Développement — 21.05.2020 – 03.06.2020
- Sprint 7 — Etape 3 - Développement — 04.06.2020 – 17.06.2020
- Sprint 8 — Etape 3 - Développement — 18.06.2020 – 01.07.2020
- Sprint 9 — Etape 4 – Tester — 02.07.2020 – 15.07.2020
- Sprint 10 — Documentation, mise en forme — 16.07.2020 – 29.07.2020
- FIN — Remise travail — 30.07.2020

Rendus

Rapport

Un rapport complet détaillant les phases d'étude et de développement de la plateforme sera livré le 31 juillet 2020 à la professeure Madame Glassey. Ce dernier permettra à un tier de comprendre l'intégralité du projet afin d'assurer la pérennité de celui-ci.

Le contexte, format et contenu de ce rapport répondra aux demandes fixées par la Haute École de Gestion de Sierre.

Plateforme web

Nous nous engageons à délivrer un outil fonctionnel et déployé sur une plateforme web. Afin de démocratiser l'utilisation de ce label, il est important d'avoir un outil facilement accessible par les utilisateurs potentiels. L'utilisation du web nous permet de mettre à disposition notre outil de labellisation à grande échelle.

Avant de déployer la plateforme à d'autres utilisateurs, nous allons la mettre à disposition à une sélection réduite de journalistes. Ces derniers auront la tâche de tester la plateforme et de certifier leurs articles. Ainsi, lors du déploiement de masse, la plateforme sera déjà testée et possèdera des articles pouvant servir d'exemple. Ces tests seront élaborés dans le cadre de notre « Proof Of Concept ».

Documentation

Par souci de clarté, une documentation complète détaillant les critères d'obtention du label sera disponible sur la plateforme web. Ainsi, les journalistes et autres utilisateurs de l'outil sauront exactement les qualifications nécessaires à l'obtention du label.

Conclusion

Il est difficile de prédire avec exactitude le déroulement d'un projet de cette envergure. La durée de chaque sprint a été déterminée de manière approximative, en prenant en compte la durée limite de 360 heures fixée par le contexte académique du travail de Bachelor. Puisqu'il est important de pouvoir délivrer un POC à la fin du travail, l'étape de développement est considérée comme la phase la plus importante et donc celle nécessitant le plus d'heures de travail. Nous nous permettrons néanmoins d'augmenter ou réduire la durée de certaines tâches en fonction de l'avancée du projet.

Sources

Travaux cités

COFEM, C. f. (2020, Janvier 20). *Services de streaming et plateformes: défis pour les médias et le public en Suisse*. Récupéré sur emek.admin.ch: https://www.emek.admin.ch/fr/actuel/apercu/

Institut Informatique de Gestion. (2020). *Traceable Original Journalistic Content (TOJC), Meeting 13.02.2020*. Sierre: Institut Informatique de Gestion.

Vandal, G. (2019, Juin 17). *À l'ère de la désinformation*. Récupéré sur lesoleil.com: https://www.lesoleil.com/chroniques/gilles-vandal/a-lere-de-la-desinformation-030aa45f522702239f835ac1db1c4132

Watts, C. (2017, Avril 3). *How Russian Twitter Bots Pumped Out Fake News During The 2016 Election*. Récupéré sur npr.org: https://www.npr.org/sections/alltechconsidered/2017/04/03/522503844/how-russian-twitter-bots-pumped-out-fake-news-during-the-2016-election?t=1582530019453