Mots-clés : Benchmarking, *Plateformes de veille, text mining, analyses textuelles, intelligence artificielle, Aleph Search Clear, Alerti, AMI EI, Ask'n'Read, CIKISI, Coalition.ai, Cogito Intelligence Platform (CIP), CorTexT, Curebot, Digimind Intelligence, Geotrend, KB Suite, KeyWatch V6.4, Knowledge360, Meltwater, Mytwip, Northern Light SinglePoint, Seemantik, Sindup, SiVeille,*

PLATEFORMES DE VEILLE : ÉTUDE COMPARÉE DES MODULES DE TRAITEMENT AUTOMATIQUE DES TEXTES ET DES IMAGES

***Text mining*, analyse sémantique et intelligence artificielle font désormais partie de l'offre des plateformes de veille. Si la fonction première de ces outils n'est pas d'analyser les informations à la place du veilleur, celui-ci est pour autant en droit d'attendre d'eux une aide à l'exploitation des textes (voire des images), pour pouvoir se concentrer sur la production d'analyses à valeur ajoutée. Dans cet article, nous recensons ces fonctionnalités, de plus en plus précieuses.**

Mathieu ANDRO

Docteur en sciences de l'information et de la communication (thèse sur le crowdsourcing soutenue à Paris 8 en 2016), Mathieu Andro est actuellement animateur du réseau de veille des Services du Premier ministre. Auparavant, il a travaillé pour les bibliothèques du Muséum national d'Histoire naturelle, dirigé celle de l'École Nationale Vétérinaire de Toulouse, conduit les projets de numérisation de la Bibliothèque Sainte-Geneviève, puis développé des services de text mining à l'Institut national de la recherche agronomique avant de devenir chef d'une division spécialisée dans la veille à la Cour des comptes. Il est l'auteur de plus de 50 publications sur les bibliothèques numériques, le crowdsourcing, le text mining, la veille et l'open access.
CV détaillé : **https://bibliotheque-numerique.fr**

⊕ **mathieu.andro@pm.gouv.fr**

Corinne DUPIN

Professionnelle de l'information depuis plus d'une vingtaine d'années, Corinne Dupin a exercé le métier de courtier en information, knowledge manager, responsable de la veille dans des environnements essentiellement conseil et marketing. Consultante au sein de l'agence Help Management pendant quelques années, elle effectue aujourd'hui des missions d'accompagnement de projet, conseil expert et formation au sein du cabinet Ourouk *http://www.ourouk.fr*. Si son domaine d'intervention privilégié est la veille, elle étend de plus en plus le champ de ses activités au knowledge management et, de façon plus globale, à l'accompagnement du changement induit par le numérique. De profil littéraire, elle alimente un compte twitter dédié aux bouleversements opérés par le numérique sur l'économie du contenu (*http://twitter.com/corinnedupin*).

Elle est l'auteur d'un Guide pratique de la veille publié en 2014 aux éditions Klog et d'un manuel sur l'Autoformation : l'apprentissage buissonnier paru aux éditions du Cercle de la librairie en 2017.

corinne.dupin@ourouk.fr

Les outils de veille mettent depuis des années l'accent sur le *text mining*, l'analyse sémantique et l'utilisation de l'intelligence artificielle, comme le signale la revue Netsources[1].

Le *text mining* est un domaine de test et de mise au point de l'intelligence artificielle depuis longtemps. Le recours au traitement automatique de textes peut permettre d'objectiver les analyses. Une note de synthèse réalisée par un humain est guidée par le choix subjectif des points de vue des quelques articles retenus. Le *text mining* permet de produire des analyses textuelles sur plusieurs dizaines de milliers d'articles et de donner une représentation de tel ou tel concept proportionnelle à son importance réelle et objective au sein de la littérature. Ces technologies sont également beaucoup utilisées par les scientifiques afin de produire des méta-analyses ou des revues (*reviews*) sur des sujets dont le volume de textes dépasse depuis longtemps les capacités de lecture des humains.

[1] *https://trends.google.fr/trends/explore?date=all&q=%22Brandwatch%22,%22talkwalker%22,%22Meltwater%22,%22Newswhip%22,%22digimind%22*

En complément des 5 familles de plateformes de veille présentées dans l'article Taxonomie, certains outils comme Cogito et CorTexT se sont spécialisés dans l'analyse de gros volumes de données textuelles (possiblement, donc, des corpus de veille) afin de proposer des états de l'art de la littérature. Grâce à des vocabulaires, thésaurus et ontologies, et/ou au moyen de l'intelligence artificielle, ils permettent d'extraire des entités nommées des textes, ainsi structurés et transformés en bases de données. A partir des index obtenus de noms de concepts, lieux géographiques, organisations ou personnes, ces outils permettent de produire des datavisualisations : cartographies de réseaux de concepts, d'organisations…

Geotrend fonctionne sur ce modèle, à la différence près que ce dernier outil est également capable de crawler des contenus, directement depuis Google par exemple. La stratégie de Geotrend est d'ailleurs désormais de se positionner comme un complément intégré à une plateforme de veille via une option d'analyse de corpus.

Certaines plateformes de veille généralistes comme Qwam CI (Ask'n'Read) proposent également nativement ce type de fonctionnalités.

L'IA, au moyen d'algorithmes et du *machine learning*, pratique la fonction « ressemble à » sur les objets textuels et désormais sur des objets non textuels (images, visages…).

En progression constante sur le traitement de texte par ailleurs, l'intelligence artificielle est désormais en mesure de produire des résumés et des traductions automatiques. Si elle apparaît de plus en plus au-devant de la scène, c'est en raison du saut de performance réalisé depuis peu, compte-tenu de la puissance des machines et de l'augmentation de leur vitesse de traitement.

De nombreuses fonctionnalités d'aide à l'exploitation et à l'analyse des textes et des images

Au nombre des fonctionnalités d'aide à l'exploitation et à l'analyse des textes et des images permises aujourd'hui par les solutions de veille figurent ainsi :

- **La détection des entités nommées** (concepts, noms d'organisations, noms de personnes, lieux géographiques…) : elle est désormais très largement proposée par les éditeurs, avec quelques spécificités comme la détection d'événements (fusion, rachat, embauche…) chez Digimind Intelligence par exemple ;

- **La traduction automatique en français de nombreuses langues étrangères** (dont le chinois, le russe, l'arabe…) : elle est proposée par 2/3 des plateformes de notre panel. Il s'agit pour autant d'aller vérifier dans le détail si la plateforme visée comprend bien l'éventail des langues utiles à sa collecte de contenus, ainsi que les modalités d'obtention de la traduction automatique (soumise le plus souvent à contractualisation). Les plateformes offrant la couverture linguistique la plus vaste sont AMI EI, Cikisi, Digimind Intelligence et Sindup ;
- **La clusterisation ou pré-catégorisation** : peu répandue, elle permet d'obtenir un premier aperçu des principaux sujets/concepts contenus dans un corpus donné ;
- **La production automatique de résumés** : plus rare encore, elle est proposée par AMI EI, Cikisi, Digimind, Northern Light SinglePoint ;
- **La comparaison de contenus** : elle permet, selon les cas, la détection de doublons (fréquente), la détection de similarités ou la détection de plagiats. Ces deux dernières fonctionnalités sont plus inégalement réparties selon les plateformes (un tiers d'entre elles seulement en sont dotées) ;
- **La mise en relation de contenus** : rare, elle permet de repérer des corrélations entre des entreprises, des individus ou des produits ; certains éditeurs comme Aleph Networks, Geotrend et Seemantik permettent même de qualifier la nature des corrélations en question : concurrence, conflit, collaboration, financement… ;
- **La détection de *fake news*** : très rarement proposée, elle peut s'opérer à partir de listes de sources paramétrables (il s'agit de soumettre à la machine un corpus de sources considérées comme pourvoyeuses de *fake news*) et/ou via *machine learning* (la machine apprenant à les détecter au fil des sources estampillées « fallacieuses » par les veilleurs). Le plus souvent, cette fonctionnalité est proposée par l'éditeur en partenariat avec une solution de traque de sources de désinformation (cas de Bertin IT avec Storyzy par exemple) ;
- **L'identification automatique d'objets au sein des images et des vidéos** : l'identification de textes, logos, produits, visages[2] au sein d'images ou de vidéos peut avoir son utilité dans le cas de veilles e-réputation (pour repérer par exemple où apparaît le logo de son entreprise) ou concurrentielle (pour identifier les apparitions du logo

[2] Notons que des outils internationaux comme meltwater et talkwalker ont une audience forcément plus large.

de son concurrent). Moins de la moitié des plateformes de notre panel sont en mesure de reconnaître tout (Northern Light SinglePoint) ou partie (Visibrain, Talkwalker, Digimind, Cikisi, Ami EI, Aleph Search Clear, MyTwip…) de ces objets.

Les éléments d'intelligence artificielle intégrés aux plateformes sont ainsi susceptibles d'apporter des compléments appréciables. Ces vertus de l'IA sont à opposer aux effets plus délétères produits par le modèle de langage GPT-3,[3] capable de produire automatiquement des contenus rédigés avec une qualité suffisante pour que l'humain ne soit plus en mesure de distinguer un texte rédigé par une IA d'un texte rédigé par l'un de ses congénères.

Face aux volumes croissants de contenus, l'aide de l'intelligence artificielle pourrait s'avérer précieuse pour cartographier l'ensemble de la littérature sur un sujet ou des corpus de veille dans le but de faire ressortir similarités, corrélations et signaux faibles[4].

Tous les éditeurs de plateformes de veille ne font pas réellement de l'IA mais le concept est à la mode (phénomène d'*IA washing*).

L'intelligence artificielle peut en tout cas se manifester au final à plusieurs étapes de la veille[4] :

- **Sourcing** : pour suggérer des sources à surveiller à partir de l'apprentissage des similarités avec des sources déjà sélectionnées par un humain ;
- **Collecte** : pour reconnaître des contenus (textes, visages, logos, produits…) dans des textes, des images ou des vidéos, transcrire de l'audio en texte (*speech to text*), traduire des textes en langues étrangères, détecter des *fake news* ;
- **Curation** : pour suggérer des contenus à valider à partir de contenus précédemment validés par un humain, résumer et taguer automatiquement des articles, synthétiser sous forme de textes ou diagrammes de volumineux corpus de textes que les humains ne peuvent plus appréhender ;
- **Analyse** : pour analyser des corpus de veille en identifiant automatiquement des concepts, des organisations, des personnes ou des lieux

[3] Simonite, Tom (2020). Did a Person Write This Headline, or a Machine? GPT-3, a new text-generating program from OpenAI, shows how far the field has come—and how far it has to go. *https://www.wired.com/story/ai-text-generator-gpt-3-learning-language-fitfully*

[4] Bourdet, Julien (2020). Visualiser la recherche sur le coronavirus en un coup d'œil. *https://lejournal.cnrs.fr/articles/visualiser-la-recherche-sur-le-coronavirus-en-un-coup-doeil*

et en cartographiant leurs relations ou pour aider à analyser le sentiment (même si la tonalité d'un contenu demeure imparfaitement détectée par les machines…).

Des subtilités fonctionnelles seront, le cas échéant, d'ultimes critères discriminants pour départager les plateformes :

- Possibilité de détecter l'émergence d'entités nommées non connues préalablement ;
- Possibilité de corriger les index d'entités nommées extraites des corpus de veilles ;
- Extraction de la phrase la plus représentative du texte[5] ;
- Production de cartographies de clusters de sous-corpus par similarités sémantiques ;
- Possibilité d'afficher des cartes géographiques à partir des entités nommées géographiques des corpus ;
- Possibilité d'afficher des analyses sous forme de nuages de mots ;
- Etc.

Grâce à l'intelligence artificielle, les veilleurs pourront à terme vraisemblablement externaliser le travail de collecte et de diffusion d'informations brutes pour se concentrer sur la production d'analyses à valeur ajoutée.

À l'avenir, l'intelligence artificielle devrait aussi permettre de reconnaître de mieux en mieux un contenu généré par une autre intelligence artificielle, de détecter une infox par apprentissage, mais aussi de remonter jusqu'à son origine.

Parmi les éditeurs qui revendiquent l'utilisation de l'IA figure notamment Cikisi. Son outil d'intelligence artificielle, Mila, permet par exemple de :

- Mettre en exergue sur la page d'accueil les articles qui font le buzz selon le nombre de likes et de commentaires qu'ils recueillent, les personnalités les plus mentionnées ;
- Proposer des contenus similaires à ceux qui ont été sélectionnés par les veilleurs (un code couleur différent permet d'identifier les suggestions de Mila) ou éviter le type d'articles considérés comme du bruit par les veilleurs ;

[5] L'algorithme prend en compte les termes de la requête ainsi que l'occurrence et le positionnement des termes dans le texte considéré pour être en mesure d'extraire la phrase considérée comme la plus représentative du contenu.

- Éditer des cartes du monde avec des zones de chaleur (*heat maps*) figurant les villes/régions/pays détectés dans les alertes, modulables selon la zone géographique souhaitée…

Les professionnels de l'intelligence économique et de la veille n'ont peut-être pas encore totalement pris la mesure de l'apport potentiel de l'intelligence artificielle et restent globalement passifs, au risque de rater le train de l'IA et de voir leur profession se ringardiser. Ils ne seraient que 15 % à l'avoir expérimentée[6].

[6] Bondu, Jérôme (2020). Sondage sur la perception de l'intelligence artificielle par les professionnels de l'intelligence économique. *https://www.inter-ligere.fr/index.php/fr/outils/1559-sondage-sur-la-perception-de-l-intelligence-artificielle-par-les-professionnels-de-l-intelligence-economique*