# Radarly : écouter et analyser le web conversationnel en temps réel

Jade Copet Christine de Carvalho Virginie Mouilleron Benoit Tabutiaux Hugo Zanghi
Linkfluence, 5 rue Choron, 75009 Paris, France
prenom.nom@linkfluence.com

## RESUME

De par le contexte conversationnel digital, l'outil *Radarly* a été conçu pour permettre de traiter de grands volumes de données hétérogènes en temps réel, de générer de nouveaux indicateurs et de les visualiser sur une interface cohérente et confortable afin d'en tirer des analyses et études pertinentes. Ce document expose les techniques et processus utilisés pour extraire et traiter toutes ces données.

## ABSTRACT

**Real time listening and analysis of the social web using *Radarly*.**
From a digital conversational context, *Radarly* has been designed to process massive heterogeneous data in real time and to allow their efficient visualization in order to draw analysis and pertinent studies. This paper presents the techniques and processes used to extract our clients data streams.

MOTS-CLES : Veille stratégique, Social Media Intelligence
KEYWORDS: Web listening, Social Media Intelligence, Social Media analytics

## 1 Hétérogénéité et volume des données : Un défi pour la veille

La démultiplication des supports de communication conjuguée à une quantité de données textuelles disponibles en croissance constante rend la capacité à accéder à l'information sous-jacente cruciale. C'est pourquoi être capable d'analyser des données à partir des contenus les plus conversationnels pour transformer les masses de données traitées en enseignements opérationnels et stratégiques nécessaires à la prise de décision est l'un des aspects clefs de ce qu'on appelle la Social Media Intelligence.
Développé depuis 2012 par *Linkfluence*, l'outil *Radarly*, plateforme intégrée de veille et d'engagement présenté ici dans sa version 2, permet de répondre à ces besoins en permettant la collecte et l'analyse de contenus web provenant de 300 millions de sources dans plus de 70 langues différentes en temps réel afin de répondre à l'exigence de ce marché cible où chaque minute compte.
Les données sont récupérées par l'intermédiaire de crawlers, d'API publiques et auprès de fournisseurs de données en temps réel (110 millions de contenus captés par jour) et se répartissent par source de la façon suivante : contenus de site web, médias, blogs, forum, sites d'avis et enfin réseaux sociaux tel que Twitter, Facebook, Instagram, Youtube, Dailymotion, Google+ ou SinaWeibo. On retrouve par conséquent des formats très variés (articles, publications courtes, messages, vidéos, images) et dont la qualité du contenu varie fortement : texte plus ou moins édité

(article, billet de blog, tweet), utilisation d'abréviations, fautes d'orthographe, etc. posant ainsi de véritables verrous techniques lors de la fusion et du traitement des données.

# 2 Approche et traitement des données

**Organisation et visualisation des données**

Ecouter le web requiert de trouver un équilibre entre quantité et qualité. Le préalable au paramétrage d'un projet de veille dans l'outil *Radarly* est ainsi de délimiter le sujet à étudier. Cela passe par l'agrégation "naïve" d'un volume conséquent de données en rapport avec la thématique puis par une exploration qualitative des données collectées par la création de différents filtres en fonction des angles thématiques à étudier. Ceux-ci fonctionnent de concert par un système d'héritage qui permet de requêter et de tagguer les données de manière très précise.

La syntaxe de ces filtres utilise les opérateurs classiques du traitement des langues naturelles (opérateurs booléens, opérateurs de proximité) auxquels s'ajoutent d'autres variables propres aux médias sociaux (recherches sur différents champs du corps du texte, hashtags, utilisateurs, smileys, reconnaissance de logos, etc.). Les données issues de ces combinaisons de filtres sont ensuite affichées dans un ou plusieurs tableaux de bord dynamiques (dashboards) visuellement structurés.

**Traitement des données multilingues et extraction d'informations**

Sur la base de ces données brutes et implicites, de nouveaux traitements sont opérés, par calcul ou par agrégation, de manière à construire de nouveaux indicateurs et de nouvelles connaissances. En raison des volumes et de l'exigence du temps réel, les analyses effectuées reposent sur des méthodes statistiques et des algorithmes d'apprentissage automatique issues du traitement automatique du langage réadaptées par les équipes R&D à ces usages particuliers. On retrouve ainsi :

**La détection de la langue,** première étape essentielle dans une plate-forme multilingue où les modèles utilisés sont souvent spécifiques à la langue de la publication. Différents algorithmes sont utilisés selon la longueur du contenu. L'approche par Infinity-grams (Nakatoni, 2012) est ainsi utilisée pour la détection de langue sur textes courts.

**L'extraction de termes** par la détection d'entités nommées au moyen de CRFs (Lafferty et al, 2001), l'extraction de groupes nominaux significatifs à partir de POS-tagging avec un modèle d'entropie maximale (Ratnaparkhi, 1996) et d'un système de règles grammaticales ou encore l'extraction spécifique des hashtags, mentions, et Emojis contenus dans les publications.

Les termes les plus saillants sont présentés sous forme de nuage de mots dans l'interface, agrégés et triés selon leur nombre d'occurrences, ou bien leur score tf-idf ou Okapi. L'interface offre aussi la possibilité d'étiqueter les publications de manière semi-automatique.

**Clustering :** Des clusters de publications sont créés en temps réel à partir d'un algorithme de similarité textuelle.

**Analyse de sentiments :** Une tonalité est attribuée à chaque publication via un ensemble d'algorithmes de classification s'appuyant sur des techniques d'analyse syntaxique et sémantique des contenus. Cette tonalité reflète le sentiment global exprimé dans le contenu et s'étend sur quatre nuances : positif, neutre, négatif et mitigé. Des méthodes de tonalisation semi-automatiques sont disponibles dans l'interface pour modifier a posteriori la tonalité.

**Deep profiling :** Les profils des influenceurs sont enrichis par des algorithmes de détection de genre et d'âge, de situation géographique ainsi que par l'extraction de données déclaratives telles que la profession.

# 3 Utilisation

*Radarly* est utilisé quotidiennement en production par plusieurs milliers d'utilisateurs réunis autour de plus de 500 projets clients. Il est également au cœur de projets de recherche académique tel que ALGOPOL (ANR-12-CORD-0018) et CODDDE (ANR-13-CORD-0017) ainsi que de projets en recherche industrielle (Projets Datascale & CuratedMedia).

# Références

LAFFERTY J., MCCALLUM .A, PEREIRA F. (2001) "Conditional random fields: Probabilistic models for segmenting and labelling sequence data",. June 2001

NAKATONI S. (2012) "Short Text Language Identification with Infinity-Gram", in Proc. of The association for NLP in Japan (NLP 2012)

RATNAPARKHI A. (1996) "A maximum entropy model for part-of-speech tagging". In : *Proceedings of the conference on empirical methods in natural language processing*. 1996. p. 133-142.