Analyse et visualisation d'opinions dans un cadre de veille sur le Web

Mohamed Dermouche*,** Leila Khouas** Sabine Loudcher* Julien Velcin* Eric Fourboul**

*Université de Lyon (ERIC LYON 2),
5 av. P. Mendès-France 69676 Bron Cedex, France
{mohamed.dermouche, sabine.loudcher, julien.velcin}@univ-lyon2.fr

**AMI Software R&D,
1475 av. A. Einstein 34000 Montpellier, France
{lkh, efo}@amisw.com

Résumé. L'analyse d'opinions est une tâche qui consiste en l'identification et la classification de textes subjectifs. Dans ce travail, nous nous intéressons au problème d'analyse d'opinions dans un contexte de veille sur le Web. Nous proposons une approche pour visualiser les résultats d'analyse d'opinions, basée sur l'utilisation de termes clés. Nous décrivons également la plateforme de veille sur le Web AMIEI, au sein de laquelle notre approche a été implémentée. La démonstration consistera en une expérimentation de la plateforme de veille AMIEI et du module d'analyse d'opinions sur un corpus de tweets politiques.

1 Introduction

L'analyse d'opinions est une tâche de fouille de textes qui consiste en l'identification et la classification des textes subjectifs en plusieurs catégories d'opinions (polarités). Dans la dernière décennie, beaucoup de travaux se sont penchés sur cette problématique, en prenant le problème sous différents angles (principalement statistique et/ou linguistique). Cependant, la question de visualisation n'a pas bénéficié de cet intérêt. La plupart des travaux proposent une visualisation basique (e.g., graphiques en secteurs), ce qui est clairement insuffisant dans un contexte de *big data* où l'utilisateur a d'autant plus besoin d'explorer les données dans l'ensemble, mais aussi dans le détail.

Dans ce travail, nous nous situons dans un contexte de veille sur le Web et nous nous intéressons au problème d'analyse d'opinions dans un contexte de veille. Ainsi, nous proposons une méthode de visualisation d'opinions basée sur l'utilisation de termes clés afin de restituer le maximum d'information à l'utilisateur. Notre méthode est implémentée au sein de la plateforme de veille AMIEI [1].

La section suivante présente la problématique de recherche que nous traitons. La section 3 présente le processus général de veille avec la plateforme AMIEI. La section 4 présente notre approche pour l'analyse d'opinions et la visualisation des résultats. Enfin, la section 5 présente un exemple d'application sur un corpus de tweets politiques.

1. AMI Enterprise Intelligence.

2 Contexte et Problématique

L'analyse d'opinions est un domaine de recherche qui se concentre sur l'identification et la classification des opinions dans les données textuelles. Beaucoup de travaux se sont intéressés à l'une ou l'autre de ces problématiques mais la plupart se sont intéressés à la classification d'opinions, i.e., l'association d'un texte à une catégorie d'opinions (e.g., opinion positive vs. négative).

La problématique a été majoritairement approchée sous un angle statistique et/ou linguistique. D'un point de vue statistique, le texte est représenté sur l'espace de descripteurs (e.g., termes) afin qu'il puisse être traité par les outils d'apprentissage statistique, e.g., Pak et Paroubek (2010); Pang et al. (2002). Ces méthodes sont connues pour leur généricité (bon rappel). De l'autre part, les méthodes de linguistique, également appelées méthodes à base de règles, ont été largement déployées pour l'analyse d'opinions, e.g., Kennedy et Inkpen (2006); Wilson et al. (2005). Ces méthodes sont connues pour leur spécificité (bonne précision). Enfin, d'autres travaux ont tenté de mixer la généricité de la statistique et la spécificité de la linguistique afin de proposer des méthodes à la fois robustes et précises (méthodes hybrides), e.g., Dermouche et al. (2013); Kamps et al. (2004); Turney et Littman (2003).

Dans ce travail, nous nous intéressons au problème de visualisation de l'opinion. En effet, la problématique de visualisation n'a pas été suffisamment étudiée dans ce domaine en se contentant de visualiser les proportions de chaque polarité d'opinion sur un graphique. Cette méthode est clairement insuffisante dans le cas où l'on veut savoir davantage sur ses données. Par exemple, dans le domaine industriel, il serait intéressant d'identifier les idées redondantes et les concepts qui sont présents dans une catégorie d'opinion et pas dans une autre. Une telle visualisation a des applications directes dans plusieurs domaines, e.g., la veille stratégique et économique, la CRM, la e-réputation, etc.

3 La Plateforme de Veille AMIEI

La plateforme AMIEI est une solution logicielle destinée à répondre à l'ensemble du cycle de veille des entreprises dans des contextes divers tels que l'intelligence économique, la veille technologique, ou l'analyse comportementale et l'e-réputation. La plateforme AMIEI consiste en une suite de modules indépendants qui permettent de mettre en œuvre les quatre principales phases d'un processus de veille (voir Figure 1), à savoir, acquisition de l'information, capitalisation et traitement, analyse de l'information et enfin partage et diffusion.

3.1 Acquisition de l'information

Cette phase permet l'acquisition de l'information selon plusieurs modes :
- Un moteur de recherche pour faire des recherches ponctuelles pouvant être capitalisées.
- Un automate de collecte pour des opérations récurrentes à des fins de capitalisation.

3.2 Capitalisation et traitement

La plateforme AMIEI est construite autour d'une base de données qui permet de capitaliser sous une forme organisée et maîtrisée les documents qui ont été collectés. Cette base de

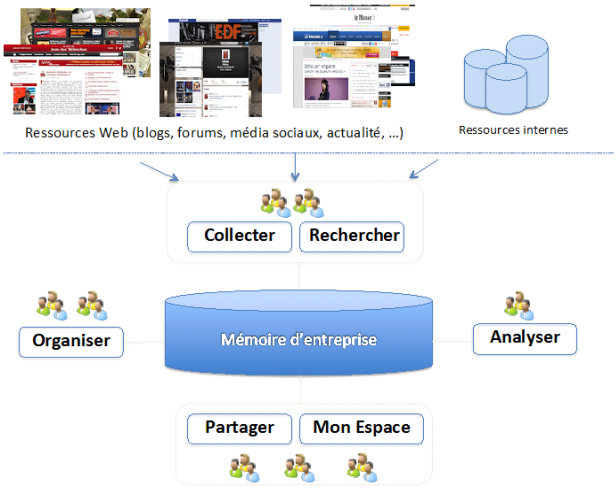

FIG. 1 – *Processus général de veille au sein de la plateforme AMIEI.*

données, appelée Mémoire d'entreprise constitue un capital important de données. Au fur et à mesure de l'exploitation du logiciel, elle favorisera la corrélation d'informations entre elles, permettra de retrouver des connaissances enregistrées depuis plusieurs mois ou années, pour devenir un véritable lieu de partage de connaissances et d'informations. Cette Mémoire d'entreprise est structurée sous forme d'un plan de classement, qui reçoit l'ensemble des résultats de la phase d'acquisition, et sur laquelle s'appuient les fonctions Analyser, Partager et Mon Espace.

3.3 Analyse de l'information

L'objectif de cette phase est l'exploration des données textuelles acquises, en vue d'en extraire une information utile et pertinente. Divers outils d'analyse statistique, de fouille de textes et de visualisation sont proposés. Ils permettent de fournir une cartographie des données collectées selon différents critères (temps, sources, etc.) et permettent la mise en évidence de tendances, la détection d'informations de rupture (signaux faibles), la recherche d'informations corrélées ou l'extraction automatique d'entités nommées (personne, lieu, organisation, concepts généraux, etc.).

3.4 Partage de l'information

Le partage et la diffusion des informations acquises et validées, ainsi que les résultats de l'analyse se font à travers un portail de consultation, permettant la recherche et le partage des informations organisées par thématique avec une gestion des droits d'accès à partir de profils prédéfinis. Le partage peut également se faire via "Mon espace" ; un module permettant de personnaliser, pour chaque utilisateur, son accès à la plateforme AMI EI.

4 Analyse d'Opinions avec AMIEI

Pour l'analyse d'opinions, la plateforme AMIEI offre la fonctionnalités suivantes :
- Indicateurs classiques de l'opinion globale dans un corpus de documents (distribution des documents sur les classes de polarité).
- Visualisation des termes clés pour chaque classe de polarités. Un terme clé doit être fréquent et discriminant vis-à-vis de la classe d'opinion qu'il caractérise.
- Evolution des termes clés à travers le temps.

4.1 Méthode et implémentation

Pour l'analyse d'opinion, nous avons choisi d'utiliser la méthode hybride décrite par Dermouche et al. (2013) et ce pour les raisons suivantes :
- La méthode est multilingue : il suffit d'avoir à disposition un corpus annoté dans la langue souhaité.
- La méthode est incrémentale : de nouvelles données peuvent facilement être intégrées en apprentissage sans avoir à reconstruire le modèle à partir de zéro.
- Simplicité et faible complexité algorithmique : basée sur la méthode statistique *Naive Bayes*, cette méthode est très simple d'implémentation, en plus d'être de faible complexité algorithmique (par rapport à d'autres méthodes statistiques comme SVM).

En se basant sur les résultats de cette méthode hybride, nous calculons un score de confiance pour chaque document classifié. Le score est compris entre 0 et 1 et est calculé, pour un document d, de la manière suivante :
- Les probabilités $p(c_i|d)$ sont triées telles que : $p(c_m|d) > p(c_n|d) > ... > p(c_p|d)$, où c_i sont les classes d'opinion (polarités).
- Le score de confiance Confiance$(d) = p(c_m|d) - p(c_n|d)$. Il représente l'écart entre la classe la plus probable et la deuxième classe la plus probable. Plus cet écart est important, plus le modèle est "sûr" d'affecter le document d à la classe d'opinion m.

Le modèle permet aussi d'"expliquer" la classification en déterminant, pour chaque document d, l'ensemble des termes qui ont conduit à ce résultat. Ceci est réalisé de la manière suivante :
- Soit c la classe d'opinion du document d (classe la plus probable).
- Evaluer chaque terme w_i du document d selon un critère de spécificité (pouvoir discriminatif du terme au regard de la classe d'opinions). Ici, nous choisissons comme critère le gain informationnel (IG). Ensuite, trier les termes w_i du document selon ce critère : IG$(w_m|c) >$ IG$(w_n|c) > ... >$ IG$(w_p|c)$.
- Les K premiers termes sont ceux qui "expliquent" le mieux cette classification.

Nous précisons que les termes discriminants de deux classes différentes sont deux ensembles disjoints. En effet, un terme ne peut être responsable d'affecter un texte qu'à une seule classe.

4.2 Visualisation

La visualisation est une étape clé dans le processus d'analyse d'opinions, d'autant plus dans un contexte de *big data*. En effet, l'information utile est encore plus enfouie et difficile à retrouver, ce qui nécessite des techniques de visualisation efficaces et adaptées à ce contexte particulier. Dans AMIEI, nous proposons de visualiser l'opinion contenue dans un corpus de

FIG. 2 – *Visualisation en nuage de termes (extrait).*

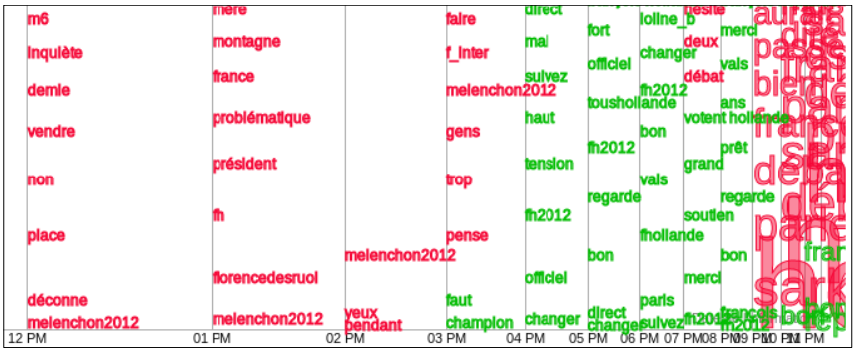

FIG. 3 – *Visualisation en fisheye (extrait).*

documents par un "nuage de termes" construit à partir de l'ensemble des termes discriminants responsables de la classification (cf. section 4.1.). Chaque terme discriminant est ainsi représenté par une taille proportionnelle à sa fréquence dans le corpus des textes. Nous proposons également une visualisation temporelle du nuage de terms en utilisant la technique de *fisheye*.

5 Etude de Cas

Nous réalisons une expérimentation sur un corpus composé de 50000 tweets issus d'une collecte massive réalisée par la plateforme de veille AMIEI dans la soirée du 02 Mai 2012 avec le tag "#ledebat" (400000 tweets collectés). Ces tweets sont relatif au débat télévisé du second tour de l'élection présidentielle française de 2012 ayant opposé F. Hollande et N. Sarkozy. Nous appliquons, comme prétraitement, la suppression de mots outils et de numériques.

Les Figures 2 et 3 représentent la visualisation du résultat d'analyse du corpus Politique. Pour une meilleure lisibilité, seulement une sélection de termes fréquents est représentée ici. La polarité des termes est représentés par une couleur (vert pour le positif et rouge pour le négatif). Ces résultats nous ont permis de cerner les termes et les concepts les plus importants dans chaque catégorie d'opinion. A partir de cette visualisation, nous pouvons tirer plusieurs enseignements dont voici quelques uns :

– Le concept de "changement" dans toutes ses variantes (slogan phare de la campagne du candidat F. Hollande) est largement repris par les internautes, et ce de manière positive.

- "melenchon2012" (compte officiel de compagne du candidat J.-L. Mélenchon) est souvent associé à des opinions négatives. En effet, ce compte est particulièrement critique envers les deux candidats (notamment N. Sarkozy) et n'a cessé de twitter des critiques acerbes tout au long du débat.
- "DSK" est marqué négativement. En effet, beaucoup d'utilisateurs (notamment des soutiens de F. Hollande) ont commenté négativement la référence de N. Sarkozy à l'affaire DSK durant le débat, et l'ont interprété comme un manque d'arguments sur d'autres sujets plus sérieux.
- Le concept "Sarko" est marqué négativement. En effet, cette abréviation du nom du candidat N. Sarkozy est surtout utilisée par ses détracteurs et non par ses soutiens.

Références

Dermouche, M., L. Khouas, J. Velcin, et S. Loudcher (2013). AMI & ERIC : How to Learn with Naive Bayes and Prior Knowledge : an Application to Sentiment Analysis. In *Proceedings of : 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, Volume 2, Atlanta, GA, USA, pp. 364–368. ACL.

Kamps, J., M. Marx, R. J. Mokken, et M. de Rijke (2004). Using wordnet to measure semantic orientations of adjectives. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, PT, pp. 1115–1118.

Kennedy, A. et D. Inkpen (2006). Sentiment classification of movie reviews using contexual valence shifters. *Computational Intelligence 22*(2), 110–125.

Pak, A. et P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, pp. 1320–1326. ELRA.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP'02)*, Philadelphia, PA, USA, pp. 79–86. ACL.

Turney, P. D. et M. L. Littman (2003). Measuring praise and criticism : Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS) 21*(4), 315–346.

Wilson, T., J. Wiebe, et P. Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, Vancouver, Canada, pp. 347–354. ACL.

Summary

Opinion mining and sentiment analysis focus on the extraction and classification of subjective text. In this paper, we propose a method for opinion visualization. Our method is based on the use of keywords from each sentiment category. We also describe the AMIEI market-intelligence platform within which our method has been implemented.