Les défis du traitement automatique du langage pour l'analyse des réseaux sociaux

Atefeh Farzindar[*] [**] — **Mathieu Roche**[***] [****]

[*] *NLP Technologies Inc., Montréal (Québec), Canada*
www.nlptechnologies.ca
[**] *Université de Montréal, Montréal (Québec), Canada*
[***] *UMR TETIS (AgroParisTech, Cirad, Irstea), Montpellier, France*
[****] *LIRMM (CNRS, Université de Montpellier), Montpellier, France*

RÉSUMÉ. *Dans le contexte de l'analyse du contenu des réseaux sociaux, le volume sans précédent des données textuelles, leur variété ainsi que le réseau d'interaction des utilisateurs représentent de nouvelles opportunités pour la compréhension du comportement social. L'étude des messages échangés constitue un nouveau défi pour le Traitement Automatique des Langues (TAL). Dans ce cadre, cet article introductif au numéro spécial discute des différentes approches et domaines d'application pour traiter les réseaux sociaux. Les défis du TAL pour analyser ces nouveaux modes de communication sont enfin discutés.*

ABSTRACT. *In the context of the analysis of social media content, the unprecedented volume of textual data, variety and the user interaction network are the new opportunities for understanding social behavior. Study of the exchanged messages is a new challenge for natural language processing (NLP). In this context, this introduction to the special issue discusses different approaches and applications address to the social networks. Challenges for NLP and analyze of these new modes of communication are discussed.*

MOTS-CLÉS : *TAL, Réseaux sociaux, Analyse sémantique*

KEYWORDS: *NLP, Social networks, Semantic analysis.*

1. Introduction

Les réseaux sociaux, structures dynamiques formées d'individus et/ou d'organisations, ont toujours joué un rôle majeur dans nos sociétés. Ils se sont développés et diversifiés avec le Web 2.0. qui ouvre la possibilité aux utilisateurs de créer et de partager du contenu par l'intermédiaire de multiples plates-formes (blogs, micro-blogs, wikis, sites de partage, etc.). Dans ce contexte, le volume sans précédent des données textuelles, leur variété ainsi que le réseau d'interaction des utilisateurs représentent de nouvelles opportunités pour la compréhension du comportement social. L'étude des messages échangés constitue un nouveau défi pour le Traitement Automatique des Langues (TAL). Dans ce cadre, il est intéressant de discuter la robustesse des méthodes de TAL (analyseurs morpho-syntaxiques, systèmes d'extraction de la terminologie et d'entités nommées, etc.) sur ces données.

Ces modes de communication sont de puissants outils collectifs où s'invente et s'expérimente le langage. De nouveaux sens sont alors associés à certains mots et la création de mots ou de nouvelles structures syntaxiques se généralisent (par exemple en mixant différentes langues). Ainsi, la création, la dissémination et le traitement de ce vocabulaire original sont discutés et, plus globalement, ce numéro spécial permet de mettre en exergue une nouvelle manière de communiquer. Certaines méta-données (par exemple, les hashtags) et les descripteurs linguistiques issus des textes constituent un socle solide pour l'analyse des réseaux sociaux. Ils permettent de mettre en avant différentes communautés socio-économiques, politiques, géographiques, etc. Par ailleurs, les descripteurs linguistiques sous forme de mots ou relations syntagmatiques permettent d'analyser avec précision les sentiments et opinions contenues dans les messages. Par exemple, les spécificités lexicales, graphiques voire syntaxiques (émoticônes, abréviations, répétition de caractères, etc.) dans les données textuelles recèlent des informations précieuses pour la détection d'opinion ou l'analyse de sentiment (détection fine des émotions, identification de l'ironie, etc.).

Enfin, les travaux actuels s'intéressent aux nouvelles problématiques qu'engendre le développement des réseaux sociaux. Par exemple, les systèmes de surveillance des réseaux sociaux doivent être capables de détecter d'éventuels usurpateurs ou d'étudier la propagation d'information.

Dans le cadre de ce numéro spécial, de nouvelles approches sont présentées dans le but d'analyser ces données textuelles massives, souvent bruitées et hétérogènes issues des réseaux sociaux. Nous avons alors sélectionné l'article "Code-Mixing in Social Media Text : The Last Language Identification Frontier ?" de *Amitava Das* et *Björn Gambäck* parmi sept articles soumis ainsi que l'article invité "Détection d'évènements à partir de Twitter de *Houssem Eddine Dridi* et *Guy Lapalme*. Les travaux présentés dans ces articles présentent différentes problématiques du traitement automatique des réseaux sociaux. Pour illustrer une telle diversité, les sections suivantes décrivent des travaux d'analyse des réseaux sociaux (section 2) ainsi que des applications associées (section 3). Enfin, les défis liés au traitement automatique du langage face aux réseaux sociaux sont détaillés (section 4).

2. L'analyse des réseaux sociaux

Au cœur de la structure des réseaux sociaux se trouvent des acteurs (personnes ou organismes), reliés entre eux par un ensemble de relations binaires (par exemple, relations, liens ou interactions). Dans ce contexte, le but est de modéliser la structure d'un groupe social, en vue de déterminer l'influence qu'elle exerce sur d'autres variables, et d'assurer le suivi de son évolution. L'analyse sémantique des médias sociaux (ASMS) se définit comme l'art de comprendre comment on recourt aux réseaux sociaux pour générer du renseignement stratégique, opérationnel ou tactique. L'ASMS favorise la création d'outils et d'algorithmes visant à surveiller, à saisir et à analyser les données volumineuses extraites des médias sociaux dans le but d'anticiper les comportements.

Un rapport publié par eMarketer (New Media Trend Watch, 2013) estimait qu'une personne sur quatre à l'échelle mondiale était susceptible d'utiliser les médias sociaux en 2013. Les statistiques sur les médias sociaux pour l'année 2012 révèlent que Facebook a dépassé la barre des 800 millions d'utilisateurs actifs, dont 200 millions de nouveaux adhérents au cours d'une seule année. La plateforme Twitter, quant à elle, compte maintenant 100 millions d'utilisateurs et LinkedIn, plus de 64 millions, en Amérique du Nord seulement (Digital Buzz, 2012). Récemment, des ateliers tels que "L'analyse sémantique des médias sociaux" de EACL 2012 (Farzindar et Inkpen, 2012), "L'analyse linguistique dans les médias sociaux", l'atelier de la NACL-HLT 2013 (Farzindar *et al.*, 2013) et l'atelier de EACL 2014 (Farzindar *et al.*, 2014) ont témoigné de l'intérêt grandissant à l'égard de l'impact des médias sociaux sur la vie quotidienne des gens, tant sur le plan personnel que professionnel.

À l'origine, les données extraites des médias sociaux sont issues de sources ouvertes obtenues à partir de blogues, de microblogues, de forums de discussion, de clavardages, de jeux en ligne, d'annotations, de classements, de commentaires et de FAQ générées par des utilisateurs. Ces données possèdent de nombreuses propriétés. D'abord, la publication des conversations est menée en temps réel. La géolocalisation de conversations sur un sujet commun est donc importante en raison des émotions, des néologismes ou des rumeurs qu'elles véhiculent. Ce type de textes, rédigés par des auteurs différents dans une variété de langues et de styles, n'adopte aucune structure précise et se présente sous une multitude de formats. Par ailleurs, les erreurs typographiques et l'argot propres au clavardage sont maintenant courants sur les réseaux sociaux, notamment sur Facebook et Twitter.

L'analyse et la veille de ce riche contenu sans cesse renouvelé donnent accès à une information précieuse que les médias traditionnels ne peuvent fournir (Melville *et al.*, 2009). L'analyse sémantique des médias sociaux a ouvert la voie à l'analyse de données volumineuses, discipline émergente inspirée de l'analyse des réseaux sociaux, de l'apprentissage automatique, de l'exploration de données, de la recherche documentaire, de la traduction automatique (Gotti *et al.*, 2014), du résumé automatique (Farzindar, 2014) et du TAL plus globalement.

Les défis actuels de l'ASMS résident dans la conception de méthodes et d'algorithmes robustes de TAL afin d'analyser un important volume de données, souvent hétérogènes, dans le cadre d'applications qui sont détaillées dans la section suivante.

3. Les applications d'analyse des médias sociaux

3.1. *Généralités*

Le traitement automatique de données extraites des médias sociaux doit arriver à déterminer les méthodes les plus appropriées pour l'extraction d'information, la classification automatique, l'indexation de données pour la recherche documentaire ou la traduction automatique par exemple. On sait que le seul volume des données et la vitesse à laquelle de nouveaux contenus sont créés suffisent à rendre irréalisable toute tentative de veille ou d'analyse manuelle significative.

La veille des médias sociaux est l'une des plus importantes applications de l'ASMS. Comme dans sa définition traditionnelle, la veille médiatique consiste en l'activité de surveillance et de suivi des médias, le contenu en ligne et les médias de diffusion, notamment dans un but politique, commercial ou scientifique. L'importante quantité d'information accessible dans les médias sociaux représente une manne de renseignements. Ces derniers ajoutent une dimension absente des médias traditionnels en nous informant sur les opinions et sentiments des auteurs. Or, les variations sur le plan du contenu et de la taille des documents sources qui en sont extraits, par exemple la combinaison de blogues et les séries de gazouillis, rendent leur analyse difficile.

Les médias sociaux soulèvent l'important problème de la recherche d'événements en temps réel et de la nécessité de les détecter (Farzindar et Khreich, 2013). L'objectif de la recherche documentaire dynamique et de la recherche d'événements en temps réel est de mettre en place des stratégies de recherche efficaces à partir de différentes fonctionnalités qui tiennent compte de multiples dimensions, y compris les liens spatiaux et temporels. Dans un tel cas, certaines méthodes de TAL, par exemple, la recherche documentaire et le résumé automatique de données sociales sous forme de documents de diverses sources, deviennent essentielles à la recherche d'événements et à la détection d'information pertinente.

L'analyse sémantique des conversations portant sur un événement qui se déroule la même journée ou de la même semaine sur les réseaux sociaux, ou d'un ensemble de discussions sur un sujet apparenté se bute aux difficultés soulevées par les aspects multilingues du TAL. Cette spécificité est accrue avec les réseaux sociaux qui peuvent mêler différentes langues dans un même contenu.

3.2. *Domaines d'application*

L'important volume d'information accessible sur les réseaux sociaux peut être mis à profit dans certains secteurs d'activités, notamment le secteur industriel et ceux de la sécurité publique et des soins de santé. Ici sont présentés quelques intégrations innovantes dans le domaine de la veille des médias sociaux ainsi que des scénarios types d'applications utilisées pour la communication entre les décideurs et les utilisateurs visant à être au fait des situations et d'en assurer la coordination. Les outils de TAL permettent également d'interpréter des données en temps réel, ou presque, favorisant ainsi la prise de décision aux plans stratégiques et opérationnels.

• **Secteur industriel.** L'intérêt pour la surveillance de données extraites des médias sociaux est considérable dans le secteur industriel. En effet, ces données sont susceptibles d'aider en optimisant de manière importante l'efficacité de la veille stratégique. L'intégration de telles données aux systèmes de veille stratégique déjà en place permet aux entreprises d'atteindre différents objectifs, notamment concernant la stratégie de marque et la notoriété, la gestion des clients actuels et potentiels et l'amélioration du service à la clientèle. Le marketing en ligne, la recommandation de produits et la gestion de la réputation ne sont que quelques exemples d'applications concrètes de l'ASMS.

• **Défense et sécurité nationale.** Ce secteur s'intéresse beaucoup à l'étude de ce type de sources d'information et de résumés pour comprendre différentes situations, procéder à l'analyse des sentiments d'un groupe de personnes partageant des intérêts communs et s'assurer d'être à l'affût de menaces potentielles dans leur domaine d'intervention. Certaines méthodes d'extraction d'information à partir du Web 2.0 y sont présentées pour établir des liens entre différentes données dénotant des entités et analyser les caractéristiques et le dynamisme des réseaux au sein desquels évoluent des organismes et des discussions. Dans ce contexte, les agrégats de comportements sociaux offrent de précieux renseignements en matière de sécurité nationale.

• **Soins de santé.** Au fil du temps, les médias sociaux se sont intégrés aux soins de santé. Ce secteur y recourt pour favoriser l'implication citoyenne et améliorer ses relations avec la clientèle. L'utilisation de Twitter comme plateforme de discussion sur des sujets tels que les maladies, les traitements, les médicaments ou les recommandations à l'intention des fournisseurs et des bénéficiaires (patients, familles et aidants) illustre bien la pertinence des médias sociaux dans ce domaine. On a donné au phénomène le nom de « santé sociale ».

• **Politique.** La veille des médias sociaux permet d'assurer le suivi des mentions faites par différents citoyens d'un pays ainsi que de l'opinion internationale à l'égard d'un parti politique. Le nombre d'abonnés que compte un parti est essentiel au déroulement de sa campagne électorale. L'extraction d'opinions et le suivi des déclarations publiées sur les forums de discussion permettent à un parti politique de mieux saisir la teneur de certains événements, lui donnant ainsi l'occasion de s'ajuster pour améliorer sa position.

3.3. *L'analyse de sentiments*

Dans ces différents domaines d'application, de nombreuses tâches se concentrent sur l'analyse des sentiments dans les différents domaines mentionnés précédemment, par exemple, pour le traitement de données politiques (Bakliwal *et al.*, 2013), médicales (Bringay *et al.*, 2014), etc. En effet, alors que les textes visent à offrir une information objective, neutre et factuelle, les médias sociaux, quant à eux, sont beaucoup plus porteurs de sentiments voire d'émotion (Neviarouskaya *et al.*, 2011). L'information subjective joue donc un rôle essentiel dans l'analyse sémantique des textes issus des réseaux sociaux. En règle générale, l'identification de sentiments repose sur deux familles d'approches. La première est fondée sur des informations statistiques liées au nombre de descripteurs linguistiques positifs et négatifs apparaissent dans chaque texte (Turney, 2002). Dans le cadre de ces approches, il est alors pertinent de s'appuyer sur des ressources existantes telles que SentiWordNet (Esuli et Sebastiani, 2006). Chaque caractéristique de cette ressource est associée à trois scores numériques décrivant l'intensité des descripteurs linguistiques selon trois critères : objectif, positif et négatif. Par ailleurs, des approches classiques fondées sur l'apprentissage supervisé peuvent également proposer des résultats tout à fait satisfaisants pour l'analyse de sentiments (Pang *et al.*, 2002).

4. Les défis liés au traitement du contenu des réseaux sociaux

L'information diffusée dans les médias sociaux, notamment dans les forums de discussion, les blogues et les gazouillis, est hautement dynamique. En conséquence, l'application des méthodes habituelles de TAL dans ce contexte ne se fait pas sans difficulté en raison du bruit, de l'orthographe inhabituelle et des fonctions limitées de classification automatique. L'importance des médias sociaux émane du fait que chaque utilisateur est désormais un auteur potentiel et que le langage se rapproche davantage de sa réalité que d'une quelconque norme linguistique (Zhou et Hovy, 2006), (Beverungen et Kalita, 2011). Les blogues, les gazouillis et les mises à jour de statuts sont rédigés de manière informelle, sur le ton de la conversation, et ressemblent plus à un "état d'âme" qu'au travail réfléchi et révisé avec soin habituellement attendu d'un média papier. Ce caractère informel crée des difficultés à tous les plans du TAL.

De prime abord, les outils habituels de TAL, conçus pour les données traditionnelles, se butent à l'emploi irrégulier, voire l'omission, de la ponctuation et des majuscules, ce qui tend à compliquer la tâche de détection des limites d'une phrase - parfois même pour le lecteur. Par ailleurs, l'utilisation de binettes, l'orthographe incorrecte ou inhabituelle et la multiplication d'abréviations populaires compliquent les tâches telles que la segmentation et l'étiquetage morphosyntaxique. Une adaptation des outils traditionnels est nécessaire pour prendre en compte les nouvelles variations comme la répétition des lettres (par exemple, "suuuuuper"). Un autre obstacle à toute forme d'analyse syntaxique est la grammaticalité, ou plutôt son absence fréquente dans les médias sociaux. En effet, les phrases fragmentées sont devenues la norme à

l'instar des phrases complètes et le choix entre différents homophones semble arbitraire (par exemple, *c'est*, *ces*, *ses*).

Les médias sociaux génèrent aussi beaucoup plus de bruit que les médias dits traditionnels. En effet, les réseaux sociaux comportent un nombre considérable de pourriels, de publicités et d'une importante quantité de contenus non sollicités, non pertinents ou dérangeants. En outre, une grande partie du contenu qualifié d'authentique et de légitime ne répond pas mieux aux besoins d'information, et est donc jugé non pertinent, comme l'illustrait bien l'étude de (André *et al.*, 2012), visant à mesurer la valeur que les utilisateurs accordaient à différents gazouillis. Des quarante mille évaluations de gazouillis recueillies, 36 % recevaient la mention "vaut la peine d'être lu" et 25 %, "ne vaut pas la peine d'être lu". Les gazouillis qui attestent seulement de la présence d'un utilisateur sur la plateforme (par exemple, "Alloooo Twitter !") se sont vus attribuer la plus faible valeur. Cela souligne l'importance du prétraitement, visant à filtrer les pourriels et autres contenus non pertinents, et de la création de modèles de gestion du bruit efficaces, en vue du traitement du langage dans les médias sociaux.

Certaines caractéristiques des médias sociaux se prêtent difficilement aux approches de TAL. En effet, les particularités d'un média et l'utilisation qui en est faite sont déterminantes dans le succès d'une approche de résumés automatiques. Par exemple, les gazouillis, avec leur limite de 140 caractères, sont plus pauvres sur le plan contextuel que les documents traditionnels. Aussi, la redondance est problématique dans une suite de gazouillis, en partie en raison de la fonction de partage. D'ailleurs, les expériences de (Sharifi *et al.*, 2010) avec les techniques d'exploration de données visant à générer des résumés automatiques de sujets à la mode sur Twitter les ont amenés à conclure à l'important problème posé par la redondance de l'information.

L'un des défis les plus importants en détection d'événements d'intérêt de diverses sources sur Twitter est la distinction entre l'information triviale et polluée et les événements concrets d'intérêt. La dispersion des données, l'absence de contexte et la diversité du vocabulaire rendent les techniques traditionnelles d'analyse textuelle difficilement applicables aux gazouillis (Metzler *et al.*, 2007). En outre, différents événements n'atteindront pas la même popularité chez les utilisateurs et peuvent grandement varier sur le plan du contenu, de la période couverte, de la structure inhérente, des relations causales, du nombre de messages générés et du nombre de participants (Nallapati *et al.*, 2004).

Les digressions sont fréquentes dans les médias sociaux, tant à cause de la nature des textes, plus proche de la conversation, qu'en raison de sa diffusion en flux. Cela ouvre la porte à l'exploration de nouvelles dimensions dans lesquelles diverses sources d'information et d'applications doivent être évaluées et exploitées. Contrairement aux textes traditionnels, considérés statiques et achevés, l'information diffusée dans les médias sociaux, notamment les forums de discussion, les blogues et les gazouillis, est hautement dynamique et caractérisée par l'interaction entre différents participants. Si elle complexifie d'autant plus le recours aux approches traditionnelles de résumés automatiques, elle offre en revanche l'occasion d'utiliser de nouveaux contextes pour enrichir les résumés, et permet même de créer de nouveaux procédés de résumés auto-

matiques. Par exemple, (Hu *et al.*, 2007) suggèrent de procéder au résumé automatique d'une publication tirée d'un blogue en extrayant des phrases représentatives à partir d'information recueillie à partir de commentaires d'utilisateurs. (Chua et Asur, 2012), quant à eux, se servent de la corrélation temporelle de gazouillis pour extraire ceux susceptibles d'être pertinents pour le résumé automatique. Aussi, (Lin *et al.*, 2009) abordent le résumé automatique non à partir de publications ou de messages, mais bien de l'ensemble du réseau social par l'extraction d'utilisateurs, d'actions et de concepts temporairement représentatifs sur Flickr.

Enfin, les méthodes classiques de TAL utilisées dans le contexte des médias sociaux se butent à l'orthographe inhabituelle, au bruit, aux fautes et aux limites de ses fonctionnalités. Certaines techniques de TAL, dont la normalisation, l'expansion morphologique, la sélection améliorée de caractéristiques et la réduction du bruit ont été proposées pour améliorer les performances de classification automatique de nouvelles extraites de Twitter (Beverungen et Kalita, 2011). Le repérage des noms et des variations de langue dans une phrase requiert que la reconnaissance des entités nommées et les techniques de détection de la langue soient précises et rapides.

5. Conclusion

L'informatique sociale est un nouveau domaine axé sur la modélisation, l'analyse et la surveillance des comportements sociaux observés sur des plateformes et médias variés dans le but de concevoir des applications intelligentes. Les médias sociaux se définissent par le recours à des outils électroniques et à l'Internet dans le but de partager et d'échanger efficacement de l'information et des expériences (Moturu, 2009). Différentes plateformes, telles que les réseaux sociaux, les forums de discussion et les blogues et microblogues, ont récemment connu des améliorations qui favorisent tant la formation de communautés virtuelles que la connectivité et la collaboration entre les utilisateurs. Alors que les médias traditionnels - tels que journaux, télévision et radio - se caractérisent par un mode de communication unidirectionnel de l'entreprise jusqu'au consommateur, les médias sociaux, eux, proposent différentes plateformes où l'interaction dans les deux sens est possible. Pour cette raison, ils sont maintenant la source primaire d'information au moment de réaliser une veille stratégique. C'est ainsi que plus récemment, les recherches ont misé sur l'analyse du langage dans les médias sociaux pour comprendre les comportements sociaux et concevoir des systèmes socioadaptés. L'objectif est d'analyser le langage en fonction des impacts possibles dans des domaines comme la linguistique informatique, la sociolinguistique et la psycholinguistique.

Remerciements

Nous remercions les auteurs pour la qualité des contribution, les relecteurs pour l'évaluation des articles soumis et Jean-Luc Minel pour son soutien et ses conseils avisés tout au long du processus.

6. Bibliographie

André P., Bernstein M., Luther K., « Who Gives a Tweet ? : Evaluating Microblog Content Value », p. 471-474, 2012.

Bakliwal A., Foster J., van der Puil J., O'Brien R., Tounsi L., Hughes M., « Sentiment Analysis of Political Tweets : Towards an Accurate Classifier », *Proceedings of the Workshop on Language Analysis in Social Media*, Association for Computational Linguistics, Atlanta, Georgia, p. 49-58, June, 2013.

Beverungen G., Kalita J., « Evaluating Methods for Summarizing Twitter Posts », *Proceedings of the 5th AAAI ICWSM*, 2011.

Bringay S., Kergosien E., Pompidor P., Poncelet P., « Identifying the Targets of the Emotions Expressed in Health Forums », *Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II*, p. 85-97, 2014.

Chua F. C. T., Asur S., Automatic Summarization of Events from Social Media, Technical report, HP Labs, 2012.

Esuli A., Sebastiani F., « SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining », *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC ?06*, p. 417-422, 2006.

Farzindar A., *Social Network Integration in Document Summarization*, IGI-Global, chapter 6, p. 139-162, 2014.

Farzindar A., Gamon M., Inkpen D., Nagarajan M., Danescu-Niculescu-Mizil C. (eds), *Proceedings of the Workshop on Language Analysis in Social Media*, Association for Computational Linguistics, Atlanta, Georgia, June, 2013.

Farzindar A., Inkpen D. (eds), *Proceedings of the Workshop on Semantic Analysis in Social Media*, Association for Computational Linguistics, Avignon, France, April, 2012.

Farzindar A., Inkpen D., Gamon M., Nagarajan M. (eds), *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, Association for Computational Linguistics, Gothenburg, Sweden, April, 2014.

Farzindar A., Khreich W., « A Survey of Techniques for Event Detection in Twitter », *Computational Intelligence*, 2013.

Gotti F., Langlais P., Farzindar A., « Hashtag Occurrences, Layout and Translation : A Corpus-driven Analysis of Tweets Published by the Canadian Government », *in* N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, may, 2014.

Hu M., Sun A., Lim E.-P., « Comments-oriented Blog Summarization by Sentence Extraction », ACM, p. 901-904, 2007.

Lin H., Bilmes J., Xie S., « Graph-based Submodular Selection for Extractive Summarization », *The eleventh biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 2009)*, IEEE, p. 381-386, 2009.

Melville P., Sindhwani V., Lawrence R. D., « Social Media Analytics : Channeling the Power of the Blogosphere for Marketing Insight », 2009.

Metzler D., Dumais S., Meek C., « Similarity Measures for Short Segments of Text », *Advances in Information Retrieval*, vol. 4425 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 16-27, 2007.

Moturu S., Quantifying the Trustworthiness of User-Generated Social Media Content, PhD thesis, Arizona State University, 2009.

Nallapati R., Feng A., Peng F., Allan J., « Event Threading Within News Topics », *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, ACM, New York, NY, USA, p. 446-453, 2004.

Neviarouskaya A., Prendinger H., Ishizuka M., « Affect Analysis Model : Novel Rule-based Approach to Affect Sensing from Text », *Nat. Lang. Eng.*, vol. 17, no 1, p. 95-135, January, 2011.

Pang B., Lee L., Vaithyanathan S., « Thumbs Up ? : Sentiment Classification Using Machine Learning Techniques », *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 79-86, 2002.

Sharifi B., Hutton M.-A., Kalita J. K., « Experiments in Microblog Summarization », *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, IEEE, p. 49-56, 2010.

Turney P. D., « Thumbs Up or Thumbs Down ? : Semantic Orientation Applied to Unsupervised Classification of Reviews », *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 417-424, 2002.

Zhou L., Hovy E. H., « On the Summarization of Dynamically Introduced Information : Online Discussions and Blogs. », *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*, p. 237, 2006.