Prédiction des *buzz* sur Twitter

Mohamed Morchid et Georges Linarès

LIA,
339, chemin des Meinajaries Agroparc – B.P. 1228
F-84911 Avignon cedex 9
mohamed.morchid@univ-avignon.fr, georges.linares@univ-avignon.fr

RÉSUMÉ. *La prédiction des buzz sur Internet est une tâche difficile notamment parce que le phénomène est dépendant de paramètres très divers, liés au contenu du message lui-même mais aussi au contexte de sa diffusion et à la dynamique de propagation de l'information sur la toile. Ces difficultés se trouvent augmentées par la dimension du Web et la dispersion et la fragmentation des informations qui s'y trouvent. Twitter est un espace d'expérimentation plus contraint et délimité que le Web dans sa globalité; dans cet article, nous présentons une méthode de prédiction des buzz appliquée à la prédiction des pics de ré-émissions (retweets) des messages postés sur Twitter. La méthode proposée repose essentiellement sur trois types d'indicateurs dont nous pensons qu'ils participent à la probabilité de re-diffusion d'un tweet : la popularité, la saillance thématique et l'expressivité. Ces descripteurs sont utilisés comme variables d'entrée d'un réseau de neurones dont le rôle est de prédire le dépassement d'un seuil de ré-émission du message. Les tests, conduits sur un ensemble d'environ 30000 messages, montrent l'efficacité de l'approche proposée : le système détecte plus de* 72% *des messages re-diffusés au moins* 60 *fois.*

ABSTRACT. *The prediction of bursty events on the internet is a challenging task. Difficulties are due to the diversity of information sources, the size of the internet, dynamics of popularity, user behaviors... On the other hand, Twitter is a structured and limited space. In this paper, we present a bursty event prediction method applied to the Twitter plateform. The proposed method uses tweet contents to predict the retweet rate. Prediction system extracts 3 types of features, related respectively to popularity, saliance and expressivity. These descriptors constitute the input features of a multilayer perceptron that predicts the retweet rate. Our experiments are conducted on a test corpus composed by about 30,000 tweets. On this test set, the proposed system detects more than 72% of the tweets that have been forwarded at least 60 times.*

MOTS-CLÉS : *Buzz, audience, twitter, modèles thématiques, Allocation Latente de Dirichlet*

KEYWORDS: *Buzz, bursty events detection, twitter, topic models, Latent Dirichlet Allocation*

1. Introduction

Les informations diffusées sur les plateformes de microblogging sont de portées très variables mais l'audience potentielle et la rapidité du support favorisent la médiatisation "explosive" de certaines d'entre elles. Cet article traite de la prédiction *a priori* de ces "explosions d'activité médiatique" (Froissar, 2007) que l'on nommera *buzz*.

La prédiction des *buzz* est une tâche difficile notamment parce que le phénomène est dépendant de paramètres très divers, liés à l'événement, à ses conséquences éventuelles, à la sensibilité du public... mais aussi aux aspects dynamiques de la médiatisation : les canaux par lesquels l'information circule, les relais, la tendance du bruit médiatique à s'auto-alimenter... Ces difficultés se trouvent augmentées par la dimension du Web et la dispersion et la fragmentation des informations qui s'y trouvent. Plusieurs études ont portées sur les modèles de diffusion de l'information (Bass, 2004, Goldenberg *et al.,* 2001). (Bass, 2004, Goldenberg *et al.,* 2001, Kempe *et al.,* 2003) étudie l'impact du "bouche à oreille" et du processus de "publicité viral" pour la diffusion de l'information. D'autres études ont porté sur la dispersion de l'information en utilisant des modèles basés sur le seuillage (Kempe *et al.,* 2003). De ce point de vue, Twitter est un espace d'expérimentation plus facile à traiter que le Web dans sa globalité ; dans cet article, nous présentons une méthode de prédiction des *buzz* appliquée à la prédiction des pics de ré-émissions (*retweets*) des messages postés sur Twitter. Certaines techniques de prédiction de messages spécifiquement sur Twiter ont fait l'objet d'étude (Romero *et al.,* 2011b, Yang *et al.,* 2010a). Concrètement, nous considérerons qu'un message aura fait du *buzz* si son nombre de *retweets* dépasse un seuil fixé *a priori*. Un *tweet* est un message très court (composé d'au plus 140 caractères), diffusé à un ensemble de personnes, appelées abonnés ou *followers*, qui suivent l'activité d'un individu sur la plateforme. Le *retweet* est le mécanisme de ré-émission permettant à un utilisateur de renvoyer un message vers ses propres abonnés. Twitter a suscité de nombreux travaux récents qui le considèrent soit comme un objet d'étude (Yang *et al.,* 2010b) soit comme un terrain expérimental structuré et délimité, permettant d'étudier des phénomènes liés à l'Internet. Cette dernière approche a donné lieu à des applications très diverses : pour la prédiction des catastrophes naturelles (Vieweg *et al.,* 2010), l'aide à la pédagogie (Grosseck *et al.,* 2008), l'analyse politique (Tumasjan *et al.,* 2010, Golbeck *et al.,* 2010), le marketing (Wright, 2009). Le *retweet* a lui-même fait l'objet de plusieurs études récentes. Certaines considèrent les relations entre les utilisateurs de Twitter et leurs abonnés pour mesurer l'influence des réseaux sociaux dans le processus de *retweet* (Hong *et al.,* 2011, Kwak *et al.,* 2010, Suh *et al.,* 2010, Cha *et al.,* 2010, Peng *et al.,* 2011, Romero *et al.,* 2011a). Trois types de critères contribuant à la probabilité de ré-émission sont généralement étudiés : le contenu et le contexte dans lequel ce contenu est produit et diffusé (Hong *et al.,* 2011, Kwak *et al.,* 2010, Suh *et al.,* 2010, Peng *et al.,* 2011), la popularité des usagers (Cha *et al.,* 2010, Romero *et al.,* 2011a) et les relations entre utilisateurs (Peng *et al.,* 2011). Dans (Naveed *et al.,* 2011), les auteurs utilisent un modèle régressif pour pondérer des facteurs de *retweet*, tels que la proximité et la nature d'émoticônes négatifs ou positifs ou d'URLs, dont les auteurs montrent que la présence est corrélée au taux de rediffusion.

Dans cet article, nous nous concentrons sur la prédiction des *buzz* par l'analyse du contenu textuel des messages. Nous proposons d'extraire de ces contenus trois types d'indicateurs dont nous pensons qu'ils participent à la probabilité de re-diffusion d'un message : la popularité des termes du *tweet*, la saillance thématique et l'expressivité. Pour chacun d'eux, nous proposons des caractérisations qui sont utilisées comme variables d'entrée d'un prédicteur neuronal.

Cette méthode est décrite dans la section 2 ; les évaluations sont présentées et commentées dans les sections 3 et 4 puis nous concluons et présentons les perspectives de cette première étude.

2. Méthode de prédiction des *buzz*

2.1. *Architecture générale du prédicteur*

Le système de prédiction proposé enchaîne trois étapes principales. La première est l'extraction des mots clés contenus dans un message. Cette phase de sélection des mots clés repose sur un modèle thématique estimé par l'analyse latente de Dirichlet (LDA, *Latent Dirichlet Analysis*). Ensuite, les descripteurs de popularité, de saillance et d'expressivité sont extraits du message lui-même et de sa représentation dans l'espace thématique. Enfin, un réseau de neurones est entraîné sur un sous-ensemble du corpus pour détecter les rediffusions massives d'un message à partir des descripteurs issus de la seconde étape. La figure 1 illustre le schéma global du système proposé, qui exécute les opérations suivantes :

1) filtrage de l'ensemble des mots W^t du *tweet* t par une liste d'arrêt composée des mots les plus fréquents du français.

2) sélection des thèmes dominants S^z du message. Cette sélection repose sur un modèle thématique T_{spc} de vocabulaire V^t, estimé *a priori* par LDA sur un corpus D de très grande taille. Ici, chaque thème est implicitement associé à une classe LDA.

3) extraction d'un sous-ensemble S^w des mots clés du *tweet* sachant W^t pour chacune des méthodes.

4) à partir de S^w, calcul des indicateurs d'expressivité et de popularité.

5) extraction des indicateurs de saillance sémantique à partir de S^w et du modèle thématique T_{spc}. L'ensemble des descripteurs constitue un vecteur de paramètres $S(t)$.

6) le prédicteur $NN(S(t))$ fournit une estimation de la probabilité que le nombre de *retweets* $r(t)$ dépasse un seuil α. Ce prédicteur est un réseau de neurones qui a été entrainé à simuler la fonction $f(S^w) = 1$ si $r(t) > \alpha$, 0 sinon.

2.2. *Extraction des mots clés*

Le langage utilisé dans Twitter est parfois assez atypique et contraint par la limite de taille des messages, inférieure à 140 caractères. L'utilisation des méthodes classiques d'extraction de mots-clefs peuvent être perturbées par ces particularités,

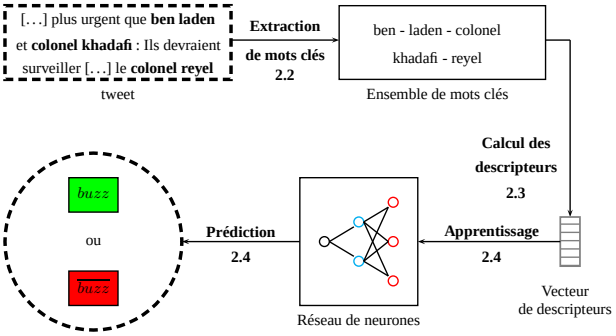

Figure 1. *Architecture globale du système de prédiction du buzz.*

notamment lorsqu'elles reposent sur l'analyse de grands corpus comme c'est le cas pour les méthodes de calcul de fréquences relatives, telles que TF-IDF-RP (**T***erm* **F***requency* - **I***nverse* **D***ocument* **F***requency* - **R***elative* **P***osition*). De façon à augmenter la robustesse de la méthode et à compenser la taille réduite des messages, nous proposons de passer par l'intermédiaire d'un espace thématique. Les mots clés sont déduits de la position du message dans cet espace thématique. Ce processus peut être vu comme une forme d'expansion du *tweet* : la projection s'appuie sur le contenu du message pour élargir l'éventail des mots clés candidats. Cette approche est comparée à celle, plus classique, basée sur TF-IDF-RP.

2.2.1. *Extraction de mots clés par TF-IDF-RP*

Cette méthode permet une extraction simple des n mots les plus représentatifs du document à partir de la fréquence du mot (TF) au sein du *tweet* t, cette dernière étant le plus souvent inférieure ou égale à 1, et de la fréquence inverse du mot au sein d'un corpus (IDF). Dans nos expériences, l'IDF est estimé sur 100000 articles issus, en parts égales, de Wikipedia et de l'AFP. Le système extrait les 50 mots ayant obtenu un TF-IDF-RP le plus élevé :

$$tf_w = \frac{n_w(t)}{|t|}, \quad idf_w = log\frac{|D|}{|\{d : w \in d\}|}, \quad rp_w = \frac{|t|}{fp(w)} \qquad [1]$$

où n_w est le nombre d'occurrences du mot w dans le *tweet* et $|D|$ le nombre de documents du corpus. $|\{d : w \in d\}|$ représente le nombre de documents du corpus où le mots w est apparu. $fp(w)$ est la position de la première occurrence du mot dans le message. $|t|$ est le nombre de mots contenus dans le *tweet*. Cette valeur étant identique pour chacun des mots du message, il est raisonnable de prendre $rp = \frac{1}{fp(w)}$.

2.2.2. *Extraction de mots clés par LDA*

LDA est un modèle qui considère un document comme un mélange probabiliste de thèmes latents, que l'on caractérise par une distribution probabiliste des mots qui leur sont associés. À l'issue de l'analyse LDA, nous obtenons n classes avec, pour chacune

d'elle, l'ensemble des mots caractéristiques de la classe accompagnés de leur probabilité d'émission. Dans nos expériences, LDA est appliqué à un corpus constitué de Wikipédia (1 Go) et des brèves de l'AFP (4.5 Go), soit un total d'environ 1 milliard de mots. Un espace de 5000 thèmes est estimé, ce nombre ayant été fixé empiriquement. Pour chacune des classes LDA, les 50 mots de poids maximum sont sélectionnés.
L'ensemble des données traitées (corpus LDA et messages) est lemmatisé et filtré par une liste d'arrêt composée de 600 mots outils du français. Un vecteur caractéristique du message est ensuite constitué, contenant, pour chaque mot du lexique, sa fréquence (le plus souvent égale à un, étant donné la compacité des messages). Ce vecteur est ensuite projeté dans l'espace de thèmes, la mesure de similarité *tweet*/thème utilisée étant le cosinus. Les mots clés sont obtenus en cherchant l'intersection des principaux m (m=2) thèmes du document. Cette intersection est composée des n mots w de score $S(w)$ maximal :

$$S(w) = \sum_{k=1}^{m} cos(t, z_k).P(w|z_k) \qquad [2]$$

où $P(w|z_k)$ représente la probabilité d'apparition du mot w sachant le thème z_k et $cos(t, z_k)$ représentant la similarité cosinus entre le thème z_k et le message.

2.3. *Indicateurs d'audience potentielle*

Bien que les raisons pour lesquelles une information fait du *buzz* soient complexes, certains éléments semblent clairement contribuer à l'audience, ou être caractéristiques de l'audience potentielle de l'événement évoqué. Par exemple, certains thèmes sont populaires et peuvent susciter des réactions passionnées. D'autre part, la saillance de l'événement, son coté inhabituel voir insolite éveille l'intérêt. Enfin, si la nature de l'événement est difficile à capturer par la seule analyse du message, la façon dont il est évoqué peut donner une indication de la vivacité des réactions qu'il peut susciter.
Les trois indicateurs que nous calculons relèvent de ces 3 aspects : le premier représente la popularité "récente" des mots en fonction d'une analyse statistique des flux RSS. Le second repose sur la probabilité d'association des thèmes dominants du *tweet* ; c'est une mesure de saillance qui utilise les associations thématiques improbables comme un facteur favorisant l'audience. Le dernier indicateur évalue l'expressivité des mots du *tweet* à partir d'un *lexique de sensibilité* préalablement annoté. Ces indicateurs sont calculés à partir des mots clefs extrait selon l'une des méthodes.

2.3.1. *Popularité*

Cet indicateur dynamique permet de connaître la fréquence d'un mot caractéristique du message dans les média. Les flux RSS de quatre grands quotidiens nationaux (Le Monde, Libération , Le Figaro et L'Equipe) sont traités dans la période précédant l'émission du *tweet* t. La *popularité* de chaque mot w de du vocabulaire de l'espace thématique V^z au sein des flux RSS est déterminée par :

$$p(w) = \frac{n(w)}{\underset{w_{rss} \in V^z}{\operatorname{argmax}} (n(w_{rss}))} \quad , \quad p(t) = \underset{w \in t}{\operatorname{argmax}} \ (p(w)) \qquad [3]$$

où $n(\cdot)$ est le nombre d'apparitions maximum d'un mot dans le flux RSS. Un classement des mots rencontrés du plus "populaire" ($p(w) = 1$) au moins "populaire" ($p(w) \approx 0$) est ainsi constitué. Concrètement, ces statistiques ont été estimées sur la période de diffusion du message sur Twitter.

2.3.2. *Saillance thématique*

Ici, nous considérons qu'un sujet "ordinaire" (i.e. fréquent) a moins de chance d'être diffusé qu'un autre, plus atypique. Nous allons donc chercher à quantifier la "singularité" du sujet évoqué dans le message. De façon plus générale, on utilise ici l'idée assez intuitive que la saillance d'un document favorise sa popularité. L'approche la plus naturelle consisterait à estimer empiriquement, sur un grand corpus, la probabilité de co-occurrence des thèmes. Malheureusement, la complexité globale du modèle à 5000 classes nous oblige à mettre en œuvre une approche heuristique, plus compatible avec les ressources dont on dispose. Dans un premier temps, nous limitons à deux le nombre de thèmes caractéristiques du *tweet*, l'objectif étant de déterminer la probabilité $P(z_i, z_j|d)$ que ces deux thèmes proches de S^w soient associés dans un même document. Plutôt que de chercher à estimer directement cette probabilité dans l'espace des classes (5000x5000), nous utilisons le fait que les classes elles-mêmes sont issues d'une analyse des co-occurrences de mots. Nous considérons alors que la probabilité de co-occurrence de deux thèmes dépend de l'intersection des classes de mots qui les caractérisent. Par ailleurs, le nombre de classes et la taille du corpus utilisé laissent supposer que la seule co-occurrence des mots ne permet pas d'évaluer toutes les associations thématiques théoriquement possibles. Pour pallier ce problème, nous construisons un graphe dont les sommets sont les thèmes et les arcs sont évalués par la mesure de divergence de Kullback-Lieber ψ normalisée entre les deux thèmes :

$$\psi(z_i, z_j) = \frac{1}{2}\left(\sum_{w \in z_i} p^i log\frac{p^i}{p^j} + \sum_{w \in z_j} p^j log\frac{p^j}{p^i}\right) \quad [4]$$

où p^i (p^j) est la probabilité qu'un mot w apparaisse dans le thème z_i (z_j). Globalement, la distance $\psi^{\triangleright}$ entre deux thèmes est alors déterminée en cherchant le plus court chemin les reliant. Cette valeur sera déterminée par la distance de Dijkstra ϕ normalisée par le nombre d'arcs ($|a(i,j)|$) et γ est alors l'indice de saillance entre les thèmes proches du tweet :

$$\psi^{\triangleright}(z_i, z_j) = \frac{1}{|a(i,j)|}\phi(z_i, z_j)\ ,\ \ \gamma(t) = \underset{(z_i,z_j)\in(S^z)^2}{\text{argmax}}\ (\psi^{\triangleright}(z_i, z_j))\ . \quad [5]$$

2.3.3. *Expressivité*

Nous mesurons l'expressivité du *tweet* au moyen du lexique de sensibilité issu de Twitter, constitué par (Paroubek, 2010). Ce lexique contient 976 mots en anglais traduit issus de ANEW (*Affective Norms of English Words*) (Bradley *et al.*, 1999). Les auteurs proposent de compter, pour chacun des mots, le nombre de fois où il apparaît près d'une émoticône positif **:)** ou négatif **:(**. Cette étude permet d'obtenir la valence

du mot dans les *tweets*, calculée en tenant compte de la probabilité que le mot soit rencontré dans un contexte positif $P(M^+|w)$) :

$$valence(w) = 8.P(M^+|w) + 1 \ . \qquad [6]$$

La valeur de valence varie entre 1 et 9, correspondants à des contextes allant de très négatif (valeur à 1) à très positif (valeur à 9). Un mot peut être considéré comme expressif s'il est rencontré dans un contexte positif ou négatif. Nous introduisons une nouvelle mesure δ qui traduit cette propriété quantifiant le niveau d'expressivité du message. Cette valeur est comprise entre 0 (quelconque) et 1 (sensible) :

$$\delta(w) = 2 \times |P(M^+|w) - \frac{1}{2}| \ , \ \delta(t) = \underset{w \in S^w}{\operatorname{argmax}} \ (\delta(w)) \ . \qquad [7]$$

2.4. *Modèle prédictif*

La prédiction est réalisée par un réseau de neurones qui évalue la probabilité de *buzz* à partir des 3 indicateurs (p, γ, δ) décrits précédemment. Il s'agit d'un perceptron multi-couches à une couche cachée, entraîné avec l'algorithme de rétro-propagation du gradient. Dans nos expériences, les indicateurs ont été testés individuellement, puis en étant combinés. Pour chacune de ces 4 configurations, un réseau de neurones $NN(S(t))$ a été entraîné. Tous les réseaux mono-indicateurs ont des couches cachées de 2 cellules, avec des couches d'entrées et de sorties limitées à une seule cellule. Dans tous les cas, les cellules d'entrées reçoivent les indicateurs, et la cellule de sortie produit une valeur entre 0 et 1 indiquant la "probabilité" de *buzz* du message considéré. L'entraînement de ces réseaux dépend du comportement attendu et notamment de la règle qu'on adopte pour interpréter un nombre de *retweets* comme un *buzz*. Plutôt que de fixer arbitrairement un nombre de ré-émissions minimal, nous avons, dans nos expériences, évalué diverses configurations, correspondant à des seuils de *retweets* α allant de 10 à 100, qui est la valeur plafond enregistrée par Twitter. Avec un seuil à 10, nous considérons qu'un *tweet* ayant été ré- émis plus de 10 fois a fait du *buzz*. Un prédicteur neuromimétique spécifique est appris pour chaque seuil testé.

3. Expériences

Le corpus est composé de $315\ 354$ (cf. table 1) messages en français obtenus en utilisant l'API de Twitter [1]. 90% sont consacrés à l'apprentissage et 10% pour le test. Nous avons évalué chacune des propositions présentées dans ce travail. En premier lieu, l'expansion des *tweets* par LDA pour l'extraction de mots clés est comparée à la méthode classique d'extraction de mots clés par TF-IDF-RP. Ces premiers tests sont réalisés en fonction du seuil du *buzz* α, dont on peut penser qu'il impacte significativement le comportement du système. La valeur de α (nombre de *retweets* minimal pour faire du *buzz*) varie de $0 \leq \alpha \leq 90$ par pas de 10. Les performances sont évaluées en terme de F-mesure (cf. courbe 2). Le système est évalué avec deux méthodes :

1. https ://dev.twitter.com/

#retweets	≥ 0	$\geq$10	$\geq$20	$\geq$30	$\geq$40	$\geq$50	$\geq$60	$\geq$70	$\geq$80	$\geq$90	$\geq$100
#tweets	315 354	6624	4770	3962	3501	3183	2927	2730	2575	2439	2300
Proportion (%)	100	2.1	1.51	1.25	1.11	1.01	0.93	0.86	0.81	0.77	0.73

Tableau 1. *Répartition des tweets en fonction du nombre de retweets.*

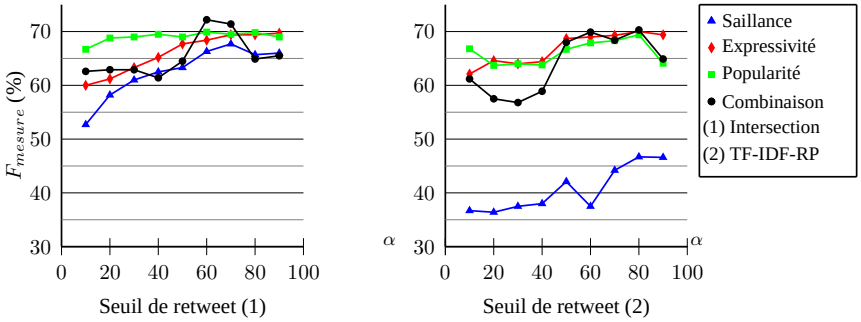

Figure 2. *Evolution de la F_{mesure} en fonction de α pour les méthodes (1) et (2)*

Intersection de thèmes et TF-IDF-RP. Les performances individuelles de chaque indicateur sont évaluées, puis celles des trois indicateurs combinés. La table 2 montre quelques exemples de *tweets* avec leurs dates de diffusion et l'intervalle de *retweets*. L'expérience consiste à évaluer l'évolution de la F-mesure en fonction de ce seuil. Nous pouvons nous attendre à une détection plus facile des exemples "extrêmes" qui sont soit très peu, soit très largement *retweetés*.

tweet	**date de diffusion**	**plage de retweets**
Oussama Ben Laden abattu par les forces spéciales américaines. . .(Radio Japon international)	02/05/2011	[0; 10[
Incroyable humanité qui fait de la sortie d'un smartphone blanc un événement planétaire	28/05/2011	[0; 10[
La une de Libé de demain #DSK K.O. http ://ow.ly/4VVhH"	16/05/2011	[10; 20[
Et même si c'est bien un complot, et que c'est prouvé d'ici les élections, le mal est fait, non ? C'est foutu, non ? Il r . . ."	17/05/2011	[10; 20[
J'ai un plan anti-sécheresse : la vaseline.	16/05/2011	[20; 30[
Trafiquer la carte de Dora pour qu'elle se perde."	16/05/2011	[20; 30[
Cette chanson sur ton MP3 qui tu saute toujours mais que tu n'efface jamais."	16/05/2011	[30; 40[
Journée internationale contre l'homophobie. Soyons fier de qui nous sommes !"	17/05/2011	[30; 40[
Toutes les filles ont une copine un peu p. . . . Si tu n'en as pas, c'est que c'est toi."	16/05/2011	[40; 50[
Faire flipper les 6emes disant que la porte du CDI mène a la chambre des secrets. #. . ."	13/05/2011	[40; 50[
Le texte de la plainte déposée contre Dominique Strauss-Kahn, traduit en fran00e7ais http ://ow.ly"	16/05/2011	[50; 60[
Twitteur, c le cite ou kan tu fé une seul fote tu tan pren plain la geulle. #. . . ;)"	13/05/2011	[50; 60[
Tu es un con. . . . Un en. . .Je te déteste. Tu me fais mal. T'es un nul, un gros raté. Je t'aime."	12/05/2011	[60; 70[
Merde, quel con...sur ma déclaration d'impôt, j'ai oublié de rajouter Bernard Tapie comme personne à charge.	13/05/2011	[60; 70[
Dimanche, le jour du seigneur... Et des devoirs.	15/05/2011	[70; 80[
Hervé Ghesquière, Stéphane Taponier et leurs trois accompagnateurs afghans sont retenus en otage depuis 500 jours.	13/05/2011	[70; 80[
Maman , je peux avoir un sac Longchamps - Tout ce que t'auras c'est un sac Auchan .	05/05/2011	[80; 90[
Dans une conversation : "Je suis souvent sur twitter .." "ah j'ai déjà crée un compte mais j'ai rien compris à ce truc.	15/05/2011	[80; 90[
Expliquer pourquoi je t'aime, c'est un peu comme expliquer le goût de l'eau ; Impossible."	12/05/2011	[90; 100]
Retenir sa respiration quand LE prof qui pue de la gueule vient te parler.	15/05/2011	[90; 100]

Tableau 2. *Exemples de tweets (mai 2011) avec date de diffusion et plage de retweets.*

4. Résultats

La première remarque est que les résultats obtenus en utilisant l'extraction de mot-clefs par LDA sont meilleurs que ceux issus d'un simple TF-IDF-RP, ce qui

confirme l'idée qui avait motivé cette approche : le passage par cette représentation intermédiaire améliore la robustesse du système à la langue "bruitée" de Twitter. Nos courbes montrent que l'utilisation de TF-IDF-RP est très mauvaise pour l'indicateur de saillance et que, plus généralement, elle a tendance à produire des résultats instables pour tous les indicateurs individuels. Avec TF-IDF-RP, la combinaison n'apporte rien par rapport à l'expressivité (environ 30% des tweets contiennent au moins un mot "sensible"). Les résultats obtenus avec les mots clés thématiques sont bien plus consistants : les trois indicateurs individuels obtiennent des résultats assez proches. Chacun des indicateurs pris séparément est supposé capturer un aspect particulier du potentiel d'audience. Les résultats de la combinaison semblent montrer cette complémentarité, même si les résultats individuels sont quantitativement proches.
De plus, les résultats obtenus avec l'expressivité le placent comme le meilleur indicateur individuel, et, plus surprenant, il dépasse la combinaison sur des seuils de *retweets* très faibles, ce qui indique une certaine capacité à détecter le *retweet* (même faible) plutôt que le *buzz*, qui est une situation d'explosion. Globalement, le fait que l'expressivité du message et le taux de ré-émission soient corrélés est un résultat assez intuitif qui se trouve confirmé ici. On voit aussi que ce n'est pas un critère suffisant pour la prédiction des *buzz*, la combinaison améliorant sensiblement les résultats dans la zone de forts taux de *retweets*.

5. Conclusions et perspectives

Dans cet article, nous avons proposé une méthode de prédiction du buzz, méthode évaluée dans le contexte de la plateforme Twitter. Ces indicateurs sont évalués séparément, puis combinés. Nos résultats montrent leur complémentarité : le meilleur système obtient une F-mesure de l'ordre de 72% pour un seuil de *retweets* $\alpha = 60$. Bien que ces performances puissent sembler de bon niveau, il est clair que chacun des indicateurs que nous avons proposé pourrait être affiné. Par ailleurs, le *buzz* est un phénomène dynamique dont la prédiction pourrait s'appuyer sur des modèles qui intègrent non seulement les contenus (ce qui a été réalisé dans ce travail) mais aussi la façon dont l'information se propage. Intégrer les aspects dynamiques et/ou structurels du mécanisme de diffusion pourrait améliorer considérablement la qualité de la prédiction. C'est cette voie que nous explorerons dans le futur.

6. Remerciements

Nous tenons à remercier les différents *reviewers* pour leurs commentaires sur une précédente version du papier. Ce travail a été réalisé dans le cadre du projet SuMACC de l'Agence National de Recherche (ANR) en vertu du contrat ANR-10-CORD-007.

7. Bibliographie

Bass F., « Comments on "A New Product Growth for Model Consumer Durables The Bass Model" », *Management science*, vol. 50, n° 12 supplement, p. 1833-1840, 2004.

Bradley M., Lang P., « Affective norms for English words (ANEW) : Instruction manual and affective ratings », *Univ. of Florida : The Center for Research in Psychophysiology*, 1999.

Cha M., Haddadi H., Benevenuto F., Gummadi K., « Measuring user influence in Twitter : The million follower fallacy », *International AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2010.

Froissar P., « Buzz, bouffées d'audience et rumeur sur Internet », 2007.

Golbeck J., Grimes J., Rogers A., « Twitter use by the US Congress », *Journal of the American Society for Information Science and Technology*, vol. 61, n° 8, p. 1612-1621, 2010.

Goldenberg J., Libai B., Muller E., « Using complex systems analysis to advance marketing theory development : Modeling heterogeneity effects on new product growth through stochastic cellular automata », *Academy of Marketing Science Review*, vol. 9, p. 1-18, 2001.

Grosseck G., Holotescu C., « Can we use Twitter for educational activities », *International Conf. on e-Learning and software for education, Bucharest, Romania*, 2008.

Hong L., Dan O., Davison B., « Predicting popular messages in twitter », *ACM International Conf. on companion on World wide web*, p. 57-58, 2011.

Kempe D., Kleinberg J., Tardos É., « Maximizing the spread of influence through a social network », *ACM SIGKDD International Conf. on Knowledge discovery and data mining*, p. 137-146, 2003.

Kwak H., Lee C., Park H., Moon S., « What is Twitter, a social network or a news media ? », *ACM International Conf. on World Wide Web*, p. 591-600, 2010.

Naveed N., Gottron T., Kunegis J., Alhadi A., « Bad news travel fast : A content-based analysis of interestingness on twitter », 2011.

Paroubek A., « Construction d'un lexique affectif pour le français à partir de Twitter », 2010.

Peng H., Zhu J., Piao D., Yan R., Zhang Y., « Retweet modeling using conditional random fields », *IEEE International Conf. on Data Mining Workshops (ICDMW)*, p. 336-343, 2011.

Romero D., Galuba W., Asur S., Huberman B., « Influence and passivity in social media », *Machine Learning and Knowledge Discovery in Databases*, vol. , p. 18-33, 2011a.

Romero D., Meeder B., Kleinberg J., « Differences in the mechanics of information diffusion across topics : idioms, political hashtags, and complex contagion on twitter », *ACM International Conf. on World Wide Web*, p. 695-704, 2011b.

Suh B., Hong L., Pirolli P., Chi E., « Want to be retweeted ? Large scale analytics on factors impacting retweet in Twitter network », *Social Computing (SocialCom)*, p. 177-184, 2010.

Tumasjan A., Sprenger T., Sandner P., Welpe I., « Predicting elections with twitter : What 140 characters reveal about political sentiment », *International AAAI Conf. on Weblogs and Social Media*, p. 178-185, 2010.

Vieweg S., Hughes A., Starbird K., Palen L., « Microblogging during two natural hazards events : what twitter may contribute to situational awareness », *ACM International Conf. on Human factors in computing systems*, p. 1079-1088, 2010.

Wright A., « Mining the Web for feelings, not facts », *New York Times*, 2009.

Yang J., Counts S., « Predicting the speed, scale, and range of information diffusion in twitter », *Proc. ICWSM*, 2010a.

Yang Z., Guo J., Cai K., Tang J., Li J., Zhang L., Su Z., « Understanding retweeting behaviors in social networks », *ACM International Conf. on Information and knowledge management*, p. 1633-1636, 2010b.