Université de Montréal

Analyse des données de microblogs

par
Houssem Eddine Dridi

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Rapport pour la partie orale
de l'examen pré-doctoral

juin, 2012

Université de Montréal
Faculté des études supérieures

Cet examen pré-doctoral intitulé:

Analyse des données de microblogs

présenté par:

Houssem Eddine Dridi

a été évalué par un jury composé des personnes suivantes:

Jian-Yun Nie,	président-rapporteur
Guy Lapalme,	directeur de recherche
Philippe Langlais,	membre du jury

Examen accepté le: . . . . . .

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

La révolution de l'information permet aux utilisateurs d'exprimer leurs sentiments et leurs opinions. Ce comportement conduit à une accumulation d'une énorme quantité d'informations. L'objectif ce travail est de développer un système qui analyse ces informations pour déterminer l'opinion publique sur différents sujets et de prédire les tendances et les préoccupations des utilisateurs.

Ces informations jouent un rôle important dans la prise de décision pour plusieurs personnes et organisations. Par exemple, il est important pour un gouvernement de connaître les opinions et les préoccupations des citoyens.

Pour obtenir des informations pertinentes qui reflètent la situation actuelle, nous devons recueillir l'information la plus récente qui contient des avis hétérogènes. Pour cette raison, il est utile d'analyser le contenu des outils de réseaux social (comme Facebook[1] ou Twitter[2]) qui ont acquis une grande popularité dans le monde. Étant donné le nombre d'utilisateurs et de leur niveau élevé d'utilisation, ces outils peuvent être utilisés comme un reflet de l'humeur du public, l'opinion publique par rapport aux événements actuels.

La majorité des travaux, qui s'intéressent à l'analyse de contenu dans le web, utilisent les outils des réseaux sociaux qui offrent le service de microblogage, les données trouvées dans ces outils étant par défaut publiques.

Les *microblogs* permettent aux utilisateurs de publier des messages courts de 140 à 200 caractères qui peuvent inclure des hyperliens et divers types de multimédias (images, vidéos, audio). Le style d'écriture utilisé dans les microblogs, est souvent non standard. Les utilisateurs commettent souvent des erreurs comme les fautes d'orthographe et de grammaire, des abréviations ou des mots étirés.

Notre thèse est menée dans le contexte d'une collaboration entre notre laboratoire (*RALI* [3]) et l'entreprise *MediaBadger*[4]. *MediaBadger* est la seule entreprise de son genre au Canada. Elle a développé une technologie exclusive qui lui permet de recueillir et

[1] `www.faceboo.com`
[2] `www.twitter.com`
[3] `rali.iro.umontreal.ca`
[4] `http://www.mediabadger.com/`

d'analyser une vaste quantité d'informations provenant de sources en ligne, y compris les médias sociaux, bases de données publiques et le web. Leur système est limité aux informations en anglais. Notre objectif est de développer un système automatique pour l'analyse des opinions et de détecter les préoccupations des utilisateurs et les tendances à partir des informations publiées sur des microblogs. Dans notre travail, nous avons l'intention d'analyser des textes également en français et en arabe.

Nous avons collecté des tweets portant sur la Tunisie que nous avons analysés.

Puisque notre objectif consiste à analyser les opinions trouvées sur le web, nous montrons dans le premier chapitre les principaux types de recherche afin de résoudre cette tâche pour des textes assez longs.

Dans le deuxième chapitre, nous présentons la différence entre les blogs traditionnels et les *microblogs*, la plate-forme Twitter, les principaux travaux qui s'intéressent à des données provenant des *microblogs*.

Le troisième chapitre présente notre corpus avec quelques statistiques. Nous terminerons par une présentation de notre plan de travail.

CHAPITRE 1

FOUILLE D'OPINIONS

Avec l'apparition du web 2, nous trouvons plusieurs plates-formes qui permettent aux internautes d'échanger leurs idées et d'exprimer leurs opinions sur un sujet particulier : un produit commercial (ordinateur portable, une voiture...), un service (agence de voyage, fournisseur de services internet...) ou bien un événement (élection, nouvelle loi...). Ce comportement d'internaute entraîne une accumulation d'une énorme quantité d'informations non structurées. L'analyse de ces informations est une tâche indispensable vu l'importance et l'utilité des avis et des opinions pour un consommateur, une entreprise ou un gouvernement.

Avant l'apparition du web, si une personne voulait acheter un produit, utiliser un service, ou décider pour qui voter, elle cherchait les opinions et les avis d'autres personnes pouvant l'aider dans son choix. Avec le web, pour connaître les avis des autres, une des premières choses à faire est de se connecter au web et de consulter les commentaires et les opinions publiées par les internautes sur ce sujet. Ces commentaires sont également intéressants pour une entreprise qui veut connaître la réputation de ses produits et la comparer à celle des produits des concurrents. Dans le domaine de la politique, le plus grand souci des gouvernements et des politiciens est de connaître ce que les citoyens pensent de la politique adoptée par le gouvernement, leurs avis sur les différents partis politiques, ou leurs points de vue par rapport à une décision prise par le gouvernement. Pour découvrir les opinions des autres, il est difficile de lire tous les commentaires qui portent sur un sujet vu la grande quantité trouvée. Cette difficulté a encouragé l'apparition de plusieurs travaux de recherche dont l'objectif est d'analyser les opinions exprimées par des internautes.

1.1 Prétraitement et nettoyage

L'un des principaux risques de prendre de l'information sur le web est le fait qu'elle n'est pas toujours fiable ou qu'elle est écrite d'une manière incompréhensible. Pour améliorer les performances de l'analyse des textes porteurs d'opinions, un prétraitement et nettoyage de ces textes sont toujours recommandés. Les tâches de cette étape peuvent être : élimination des doublons, correction orthographique...

Certains travaux se concentrent sur la détection des *spams* qui visent à promouvoir et favoriser un sujet par rapport aux autres, ou même de nuire à la réputation d'un tel sujet.

Par exemple, dans une période d'élection on peut trouver des gens qui publient d'une façon étrange des avis qui vantent un parti particulier. Ce type d'opinions cherche à tromper les systèmes d'analyse d'opinions et à retourner une information fausse aux internautes.

Dans notre travail, nous supposons que les textes reflètent l'opinion de vraies personnes, ce type de filtrage ayant déjà été appliqué dans une étape préalable.

1.2 Classification des sentiments

Pour avoir une idée concernant un sujet (produit, service,...) la première question qu'on se pose est : *est-il est bon ?*. Pour répondre à cette question à partir des commentaires publiés par les internautes, il est important de les classifier selon le type global de l'opinion exprimée : favorable, défavorable ou neutre.

Plusieurs travaux (Yu et Hatzivassiloglou [2003], Wiebe et al. [2001] et Dave et al. [2003]) ont montré qu'il faut s'assurer que le texte exprime une opinion avant de déterminer son type global.

On parle souvent de polarité de texte dans ces cas : polarité positive correspondant à une opinion favorable et négative à défavorable. Pour certaines études, on utilise d'autres classes telles que *excellent*, *très bien*, *bien*, *moyen* ou *neutre*

L'objectif de la classification des sentiments est de donner rapidement une impression du ton d'un texte. Cette classification est appliquée à chaque phrase ou paragraphe dans le texte. Celle-ci ressemble à la classification des textes qui permet d'annoter les

textes avec des classes (`sport`, `politique`, `éducation`...). Pour chaque classe C, on trouve des termes importants considérés comme des indices pour C (Liu [2007]) . Par exemple, les termes `loi`, `gouvernement`, `président`, `justice`... sont des indicatifs du sujet politique. Dans la classification des sentiments les termes qui indiquent un sentiment (négatif ou positif) sont importants : *like*, *dislike*, *hate*,*good*, *excellent*...

Dans la littérature deux types de méthodes ont été utilisées pour trouver la polarité d'une opinion : méthodes par apprentissage supervisé et par orientation sémantique.

1.2.1 Méthodes par apprentissage supervisé

La première étape de cette approche nécessite une intervention manuelle pour annoter les données d'apprentissage, ces données sont un ensemble d'opinions. Cette étape consiste à lire le commentaire (ou l'opinion) et à affecter une classe (positive, négative ou neutre), cette tâche est généralement effectuée par un ou plusieurs experts.

L'étape suivante consiste à construire un classificateur appris à partir des données d'apprentissage annotées à l'étape précédente. Ce classificateur permet de déterminer la classe (la polarité) d'un nouveau commentaire (ou opinion).

Pour l'apprentissage, il faut sélectionner un ensemble de caractéristiques utilisé dans la classification de texte, permettant de décrire les données d'apprentissage avant de construire le classificateur. Ces caractéristiques peuvent être : ensemble de mots, position de certains mots dans le texte ou des ponctuations.

Chaque document d sera présenté comme : $d = (n_1(d), n_2(d), \ldots, n_m(d))$.
Où $n_i(d)$ est le nombre de fois que la caractéristique c_i apparaît dans le document d.

Afin de construire un classificateur performant, on peut modifier les caractéristiques jusqu'à trouver les caractéristiques qui donnent le meilleur classificateur.

Les éléments indispensables dans cette approche sont : les classes, les données d'apprentissage annotées, le classificateur construit. Cette approche reprend les méthodes d'apprentissage utilisées souvent dans la classification de textes, comme les *réseaux bayésiens naïfs*, *machines à vecteurs de support*...

Pang et al. [2002] ont appliqué cette approche pour des commentaires sur des films.

Les commentaires sont obtenus de l'archive de Internet Movie Database (IMDb)[1].

1.2.2 Méthodes par orientation sémantique

Contrairement à l'approche par apprentissage supervisé, ces méthodes n'utilisent pas un ensemble d'apprentissage. La plupart des études basées sur cette approche considèrent que les adjectifs et les adverbes sont de bons indicateurs pour déterminer la polarité d'une opinion.

Turney [2002] a proposé une méthode permettant de prédire la polarité d'une opinion, celle-ci était déterminée par l'orientation sémantique des phrases qui contiennent un adjectif ou un adverbe. L'orientation sémantique d'une phrase est négative si elle a un caractère défavorable (*very bad*), et positive pour un caractère favorable (*is good*). Cette méthode comporte trois étapes :

1. Attribuer à chaque mot sa catégorie grammaticale (*Nom*, *Verbe*, *Adjectif*,...). Puis sélectionner les syntagmes de taille égale à 2 selon des règles prédéfinies (tableau 1.I).

Tableau 1.I – Règles pour extraire les syntagmes de taille 2 à partir d'une opinion

premier terme	**deuxième terme**	**troisième terme (qui ne doit pas être)**
1. *Adjectif*	*Nom*	Rien
2. *Adverbe*, *Adverbe comparatif* ou *Adverbe superlatif*	*Adjectif*	ne doit pas être Nom
3. *Adjectif*	*Adjectif*	ne doit pas être Nom
4. *Nom*	*Adjectif*	ne doit pas être Nom
5. *Adverbe*, *Adverbe comparatif* ou *Adverbe superlatif*	*Verbe à l'infinitif*, *au passé*, *au passé composé* ou *participe présent*	Rien

2. On calcule une mesure d'orientation sémantique (OS) inspirée de la mesure *PMI (Pointwise Mutual Information)* qui est utilisée pour calculer le degré de similarité

[1] `http://reviews.imdb.com/Reviews/`

entre deux termes.

$$PMI(terme1, terme2) = \log(\frac{P(terme1 \& terme2)}{P(terme1)P(terme2)})$$

P(terme1&terme2) est la probabilité que *terme1* et *terme2* apparaissent ensemble dans un même document. *P(terme2)P(terme2)* est la probabilité que *terme1* et *terme2* apparaissent ensemble, s'ils sont statistiquement indépendants. Le ratio entre *P(terme1&terme2)* et *P(terme1)P(terme2)* mesure le degré de dépendance entre *terme1* et *terme2*.

Pour chaque syntagme *s*, on calcule son *OS*. Les termes *excellent* et *poor* sont utilisés parce qu'ils représentent respectivement des bons indicateurs, des caractères favorable et défavorable.

$$`OS(s) = \log(\frac{hits(s\ NEAR\ \text{“}excellent\text{”})hits(\text{“}poor\text{”})}{hits(s\ NEAR\ \text{“}poor\text{”})hits(\text{“}excellent\text{”})})$$

D'où *hits("poor")* est le nombre de documents qui contiennent *poor*, *hits(s NEAR "excellent")* (*hits(s NEAR "poor")*) contient le nombre de documents où *s* et *excellent* (*poor*) cooccurrent. Turney a utilisé le moteur de recherche *AltaVista*[2] dans ses expériences.

3. La polarité de l'opinion est déterminée par la moyenne des orientations sémantiques des syntagmes sélectionnés (tableau 1.I). Si la valeur obtenue est positive, la polarité est positive sinon elle est négative.

D'autres travaux utilisent des dictionnaires (e.g. *general inquirer*[3]) pour déterminer l'orientation sémantique des termes présents dans l'opinion. Pour ce faire, on compare le nombre de termes à orientation sémantique positive avec le nombre de termes à orientation sémantique négative. Si le nombre de positifs est plus grand, la polarité de l'opinion est considérée positive et vice versa.

[2] www.altavista.com

[3] http://www.wjh.harvard.edu/~inquirer/

1.3 Résumé automatique des opinions

Il est intéressant de connaître la polarité d'une opinion, mais il est souvent souhaitable d'obtenir plus de détails sur les évaluations ou les avis : Pourquoi un internaute donne-t-il un avis favorable ou défavorable sur un sujet ? Une opinion défavorable n'implique pas toujours que l'internaute soit totalement insatisfait. Une description détaillée des opinions est alors indispensable et il faut construire un texte qui résume les avis des internautes pour chaque caractéristique.

Les principaux travaux qui s'intéressent au résumé automatique des opinions poursuivent la démarche suivante pour obtenir une description détaillée pour les opinions qui portent sur un produit P :

1. Identifier les caractéristiques C $\{c_0, c_1, c_2, ...\}$ du produit P (Hu et Liu [2004], Popescu et Etzioni [2005]).

2. Extraire les phrases qui contiennent une opinion sur une caractéristique c_i (Yu et Hatzivassiloglou [2003], Dave et al. [2003], Wiebe et al. [2001]) .

3. Trouver la polarité de chaque phrase extraite (Turney [2002], Pang et al. [2002]).

L'approche de Hu et Liu [2004] produit des résumés pour des commentaires qui portent sur un produit spécifique. Le système énumère les caractéristiques du produit avec le nombre des phrases positives et négatives pour chaque caractéristique. La figure 1.1 montre un exemple de résumé pour des commentaires.

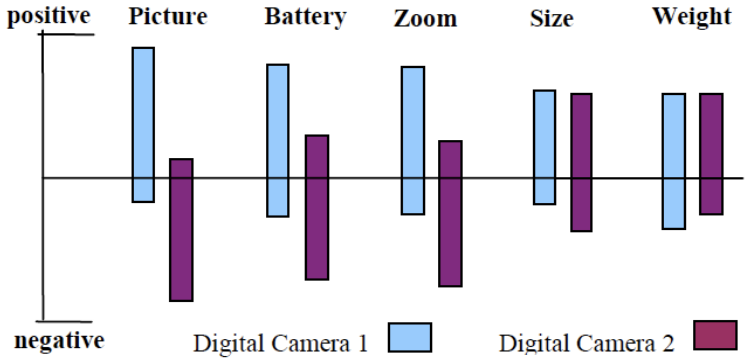
Figure 1.1 – Exemple de résumé (Figure1 Hu et Liu [2004]

*Digital_camera_*1:

Feature: **picture quality**
Positive: 123 <individual review sentences>
Negative: 6 <individual review sentences>
Feature: **size**
Positive: 82 <individual review sentences>
Negative: 10 <individual review sentences>
...

Liu et al. [2005] ont proposé un système qui produit des résumés sous forme graphique pour la comparaison entre des produits compétitifs. D'un coup d'œil, l'utilisateur peut voir clairement les forces et les faiblesses de chaque produit. La figure 1.2 montre une comparaison graphique entre deux appareils photo. Chaque barre dans la figure montre le pourcentage de commentaires (sur une caractéristique d'une caméra) positifs et négatifs.

Figure 1.2 – Comparaison graphique entre deux appareils photo (Figure1 Liu et al. [2005])



La hauteur de la barre représente le nombre de phrases positif (ou négatives) pour une caractéristique c, notée par L^+ (ou L^-), est calculée comme suit :

$$L^+_{i,j} = \frac{N^+_{i,j}}{Max(M^+, M^-)} \qquad L^-_{i,j} = \frac{N^-_{i,j}}{Max(M^+, M^-)}$$

Où $N^+_{i,j}$ ($N^-_{i,j}$) est le nombre de phrases positives (négatives) pour une caractéristique j d'un produit i ; M^+ (M^-) est le nombre maximal des phrases positives (négatives) de toutes les caractéristiques du produit i ; $Max(M^+, M^-)$ est le maximum entre M^+ et M^-.

1.4 Conclusion

Dans ce chapitre, nous avons présenté la revue de littérature pour l'analyse des opinions. Ceci comprend les tâches de résumé de textes contenant d'opinions et la classification des sentiments. Nous avons distingué deux types d'approches pour la classification

des sentiments qui nous intéressent dans notre travail pour déterminer les attitudes des utilisateurs pour différents sujets.

Les travaux présentés dans ce chapitre traitent des textes assez longs, dans le chapitre suivant nous montrerons les principaux travaux de recherche qui s'intéressent aux textes courts.

CHAPITRE 2

ANALYSE DES TEXTES COURTS

Dans le chapitre précédent, nous avons présenté l'analyse de textes d'opinion assez longs. Dans ce chapitre nous présentons la différence entre les blogs et les microblogs, la plate-forme Twitter, le type de données que nous allons utiliser et les principaux travaux de recherche qui s'intéressent aux textes courts.

L'analyse des données de réseaux sociaux est devenue une tendance majeure dans le domaine de traitement de la langue naturelle.

Ainsi, les grandes communautés de TALN ont accordé sa juste part à l'analyse des données de microblogs. Dans les dernières années, les grandes conférences (EACL[1], NAACL[2]) ont créé des workshops pour l'analyse des données dans les réseaux sociaux. Dans le même contexte, beaucoup de conférences ont été créées telle que : International Conference on Web-blogs and Social Media (ICWSM) par AAAI et Advances in Social Network Analysis and Mining (ASONAM) .

2.1 Blogs et Microblogs

Un blog est un site web, sous forme d'un journal, où une ou plusieurs personnes appelées blogueurs publient leurs opinions et leurs analyses sur un événement actuel ou d'autres questions. Le texte publié peut contenir des liens hypertextes et plusieurs types de multimédias (images, vidéos, audio). Les blogs fournissent une forme d'interaction en ligne où les visiteurs peuvent lire le contenu du blog, laisser leurs commentaires, laisser des liens vers des informations supplémentaires, participer à des discussions avec les blogueurs et avec d'autres visiteurs. La discussion dans les blogs se fait d'une manière asynchrone (comme les courriels). Plusieurs plateformes de création des blogs (Blog-

[1] `http://eacl2012.org/home/index.html`
[2] `http://www.naaclhlt2012.org/`

ger[3], Skyrock Blog[4]) permettent une diffusion simple et facile pour les blogueurs et les invités. Plusieurs travaux intéressés par l'analyse d'opinions ont eu recours aux blogs pour collecter les données.

Un article doit être pertinent et bien rédigé pour attirer l'attention des visiteurs. Cela implique qu'un article demande du temps pour le préparer. Pour cette raison, la plupart des internautes aiment mieux être passifs, ils préfèrent lire les articles des autres ou laisser des commentaires, que de créer des blogs et diffuser leurs propres articles. Généralement, les informations publiées dans les blogs ne sont pas en temps réel.

Une nouvelle tendance a émergé est celle du microblogage. C'est un dérivé du blog traditionnel qui permet aux utilisateurs de publier des messages courts (entre 140 et 200 caractères au maximum) sans titre, qui peut contenir également plusieurs types multimédias.

Au début, l'idée était de permettre aux internautes d'indiquer aux autres ce qu'ils sont en train de faire. Cependant, les choses se sont développées vite et les internautes ont profité de ce service pour exprimer leurs opinions sur différents sujets, diffuser des informations et faire des discussions. Contrairement aux blogs la plupart des internautes sont actifs, car ils n'ont plus besoin de rédiger de longs textes.

Le style d'écriture, employé dans les microblogs, est parfois incompréhensible par les non-initiés ou par les gens qui ne font pas partie de la conversation. Les utilisateurs commettent fréquemment des fautes d'orthographe et de grammaire, utilisent des abréviations, étirent des mots, et utilisent d'onomatopées (rire = *ha ha*) et des néographies (qui =*ki*). Plusieurs plateformes offrent le service de microbloggage tels que : *Twitter*[5], *Tumblr*[6], *Jaiku*[7], *Plurk*[8], *Identi.ca*[9].

Au départ, le service de microblogage était utilisé notamment par les jeunes, mais présentement tous les groupes d'âge utilisent ce service. Les *microblogs* (Plates-formes

[3] www.blogger.com/
[4] http://www.skyrock.com/blog/
[5] https://twitter.com/
[6] https://www.tumblr.com/
[7] http://www.jaiku.com/
[8] http://www.plurk.com/
[9] http://www.identi.ca/

pour le microbloagage) sont devenus d'excellents outils pour des entreprises pour faire des publicités sur leurs produits et services ou pour les célébrités pour communiquer avec leurs fans. Les *microblogs* jouent aussi le rôle d'un réseau social, parce que les utilisateurs sont en mesure de faire des relations avec d'autres. On peut trouver deux types de relation dans les microblogs :

- **Asymétriques :** Un utilisateur A suit un utilisateur B sans que B suive A, cela implique que A peut consulter (sur son tableau de bord) les messages de B. Par contre, B ne peut pas consulter ceux de A.
- **Symétrique :** Un utilisateur A suit un utilisateur B et B suit A, cela implique que chacun peut consulter (sur son tableau de bord) les messages de l'autre.

Plusieurs raisons ont encouragé les internautes à s'inscrire à des *microblogs* : facilité d'utilisation, contact avec les amis, information en temps réel ou échange d'idées avec les autres.

Les études les plus récentes, portant sur l'analyse des opinions et des sentiments, ont choisi les *microblogs* comme source des données vu le nombre important des messages publiés par jour. Les messages peuvent contenir beaucoup de sentiments et d'émotions parce que la majorité des messages sont publiés d'une façon spontanée.

2.2 Twitter

2.2.1 Présentation

Twitter est actuellement la plate-forme de microbloggage la plus populaire. Son premier slogan était `Que faites-vous?` néanmoins l'utilisation a pris une autre piste où les utilisateurs échangent des avis et des informations, le slogan devient `Quoi de neuf ?`. Plusieurs célébrités utilisent Twitter, on y trouve même des chefs d'État.

Twitter limite le nombre de caractères utilisés dans un message, appelé tweet, à 140 et qui peut contenir également des liens hypertextes. Les utilisateurs peuvent recevoir et envoyer des messages via le service SMS.

Selon les derniers chiffres [10]

- Twitter a plus que 475 millions utilisateurs inscrits.
- 175 millions de tweets envoyés chaque jour.

Le tableau 2.I illustre les 5 pays qui ont le plus grand nombre d'utilisateurs.

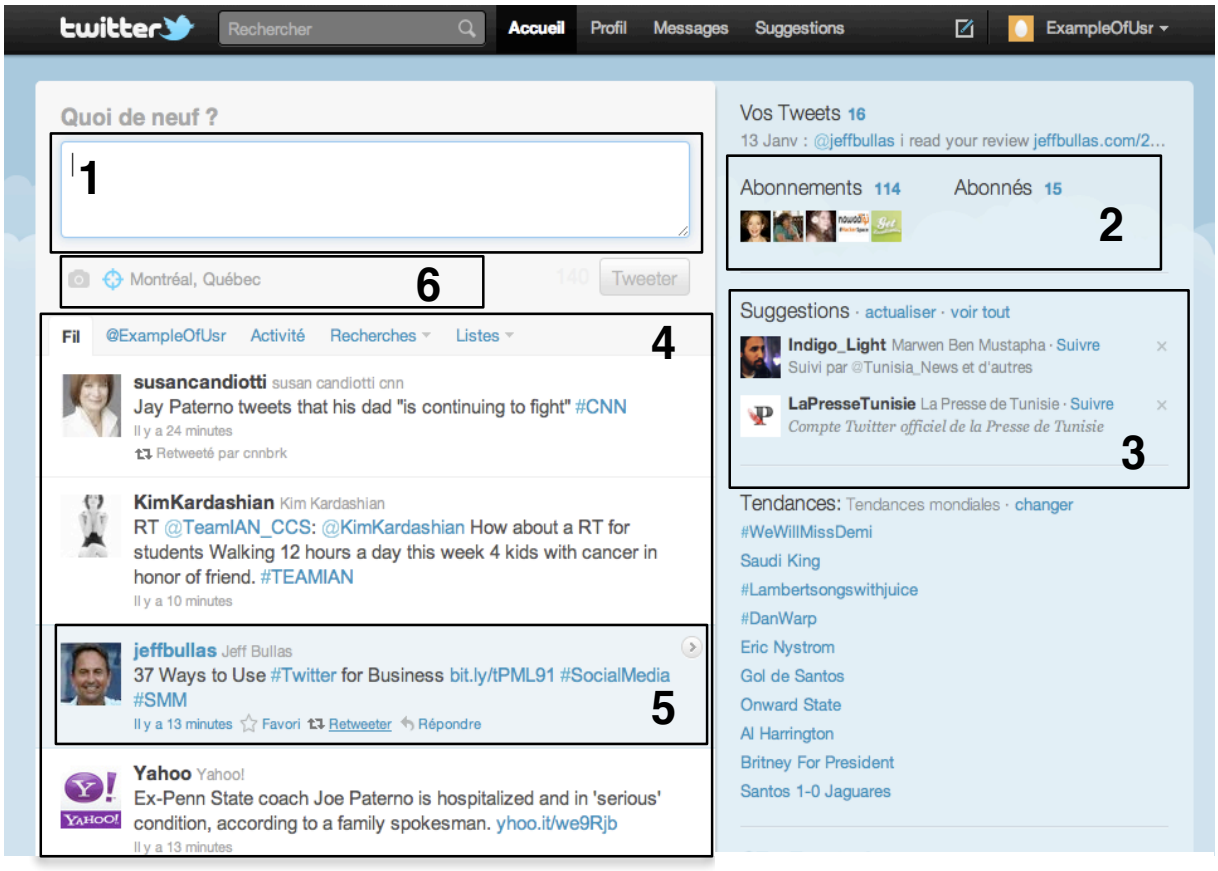
Tableau 2.I – Les 5 pays qui ont le plus grand nombre d'utilisateurs.

Pays	Nombre d'utilisateur (en million)
États-Unis	107.0
Brésil	33.0
Japon	30.0
Royaume-Uni	23.8
Indonésie	19.5

[10] `http://www.jeffbullas.com/2012/04/23/48-significant-social-media-facts-figures-and-statistics-plus-7-infographics/`

2.2.2 Fonctionnalités et caractéristiques

Figure 2.1 – Page d'accueil Twitter avec des zones d'informations : 1- champ de texte où l'utilisateur doit saisir son tweet ; 2- nombre d'abonnements et d'abonnés de `ExampleOfUsr` ; 3- suggestion d'utilisateurs à suivre ; 4- le fil de tweets des abonnements de `ExampleOfUsr` ; 5- un tweet de l'utilisateur `@jeffbullas` ; 6- localisation du tweet.



Les tweets sont, par défaut, publics. Cependant, Twitter offre à ses membres la possibilité de limiter la liste des utilisateurs autorisés à voir leurs tweets. La plupart des relations effectuées entre les utilisateurs sont asymétriques. Selon Kwak et al. [2010], seulement 22% des relations dans Twitter sont symétriques. Les utilisateurs sont en mesure d'envoyer et de recevoir des messages privés. Cette fonctionnalité est autorisée, seulement, si la relation entre l'expéditeur et le récepteur est symétrique.

Une fois connecté, un fil de tweets écrits par les abonnements s'affiche. Les abonnements sont les utilisateurs que l'on a choisis de suivre. La figure 2.1 présente la page d'accueil de l'utilisateur `ExampleOfUsr`. Twitter permet à ses membres de :

- Répondre à un tweet ou de le retweeter (zone d'informations 5). c'est-à-dire renvoyer un tweet sans changer son contenu.
- Géolocaliser les tweets envoyés (zone d'informations 6) : cette option permet de définir l'endroit d'où est envoyé un tweet.

Le tableau 2.II présente un exemple de tweets (qui portent sur les élections tunisiennes en 2011).

Conventions d'écriture :

- Le nom d'un utilisateur est un identifiant qui doit être précédé toujours par `@`. Exemple : `@ExampleOfUsr`.
- Dans un tweet on peut étiqueter les sujets dont on parle. Un sujet est précédé par un hashtag `#`. Pour le tweet :"`j'ai voté, ta7ya tounes #TnElec #Vote`", les sujets sont `TnElec` et `Vote`. En cliquant sur `#TnElec` la liste des tweets qu'ils ont comme sujet `TnElec` s'affiche.
- Une réponse à un tweet de l'utilisateur `X` commence toujours par `@X`.
- Un retweet commence par `RT @Y` où `Y` est le titulaire du tweet.
- Pour mentionner un utilisateur dans un tweet il suffit de taper son nom précédé par `@`.

Tableau 2.II – Exemples de tweets

Tweet	**Explications**
RT @Aida_SAFI: La démocratie c'est que tu accept le résultat du scrutin quel que soit le parti majoritaire #Tnelec #Tunisie	C'est un retweet. L'auteur d'origine est @Aida_SAFI
j'ai voté, ta7ya tounes #TnElec #Vote	L'auteur a informé qu'il a déjà voté. Les mots ta7ya et tounes, dans le deuxième tweet, sont des néographes des mots arabes qui correspondent respectivement à vive et la Tunisie.
#google #tnelec http://t.co/tnNdPZs8	L'auteur a introduit #google et #tnelec comme sujets du tweet et a inséré un lien http://t.co/tnNdPZs8 qui accède à la page google.tn qui a fêté les élections tunisiennes.
@so9rat11 nous sommes ts tunisiens et ns ns devons de respecter la loi q est au dessus de ts! #tnelec #Tunisie	C'est une réponse au tweet de @so9rat11. ns=nous, q=qui, ts =tous.
Quoi qu'il arrive, 1000 mercis à @ISIETN et bravo si Kamel Jandoubi #tnelec #Tunisie	L'auteur a mentionné l'utilisateur @ISIETN, si est un néographe de mot arabe corresponde à monsieur.
تصويرة بن علي رجعت في حلق الوادي http://t.co/2w4Ishy0 excellent !!! #tnelec	Le texte arabe signifie que la photo Le texte arabe signifie que la photo de Ben Ali (président tunisien déchu) a été republiée dans La Goulette (ville tunisienne).

Certains URLs qu'on veut publier sur Twitter sont trop longues et dépassent la taille permise pour un tweet. Pour cela, Twitter utilise un service de réduction d'URL qui rend la page accessible par l'intermédiaire d'une très courte URL. Il existe plusieurs

services de réduction tels que *TinyURL*[11], *bitly*[12] et *t.co* (créé par Twitter et qui est utilisé seulement pour les URLs insérées dans les tweets). Par Exemple l'URL `http://www.iro.umontreal.ca/rubrique.php3?id_rubrique=13` devenue `http://t.co/NrUGjAtx` par le service `t.co`. Un service génère toujours la même URL pour la même entrée.

2.3 Analyse des sentiments

Comme nous l'avons défini dans le chapitre précédent (section 1.2), cette tâche consiste à déterminer la classe (positive, négative, neutre) de textes assez longs. Dans cette section nous présentons les principaux travaux traitant des textes provenant de microblogs.

2.3.1 Go et al. [2009]

Go et al. [2009] ont développé une application qui est disponible en ligne intitulée : *twitter sentiment*[13]. Cette application permet de déterminer les polarités des messages publiés sur Twitter et qui répondent à la requête envoyée par l'utilisateur.

Dans ce travail, les auteurs ont considéré seulement deux classes (positive et négative). Les méthodes utilisées pour déterminer la polarité des messages sont des méthodes basées sur l'apprentissage supervisé : *méthode bayésienne naïve*, *méthode d'entropie maximale*, *les machines à vecteurs de support*. Comme nous l'avons mentionné au chapitre précédent, les méthodes par apprentissage supervisé utilisent un ensemble d'apprentissage annoté a priori. Pour le construire, les auteurs se sont basés sur les émoticônes pour déterminer la polarité d'un message. Les émoticônes, appelés aussi *smilies*, sont utilisés souvent dans les messages envoyés sur les microblogs, ils sont employés pour exprimer des émotions (heureux, triste, nerveux...).

Les auteurs ont classifié le message comme positif s'il contient un émoticône positif `:)`, `:p`, `;)` et comme négatif s'il contient un émoticône négatif `:(`, `:s` . Un tweet

[11] `tinyurl.com`
[12] `bit.ly`
[13] `http://twittersentiment.appspot.com/`

qui contient un émoticône positif et un émoticône négatif à la fois est supprimé de l'ensemble d'apprentissage. Les tweets dupliqués et les retweets sont également supprimés.

Une autre méthode pour déterminer la polarité d'un message est basée sur un ensemble[14] de termes positifs (`quite amazing`, `thks`, `so great`, `highly positive`, `;-)`...) et négatifs (`FTL`, `in a bad way`, `doesn't recommend`, `upset`, `:(...)`.

La polarité est déterminée par la différence entre le nombre de termes positifs et négatifs, présents dans le message, si les termes positifs sont plus nombreux que les termes négatifs la polarité sera positive et vice versa.

Avant la phase de la classification, un prétraitement a été effectué sur les données d'apprentissage dont le but est de réduire l'espace de *features* :

- Remplacer les noms d'utilisateur dans le message par une seule expression : `USERNAME`. Exemple : le message `@Marwen tu es optimiste ! #tnelec` sera remplacé par `USERNAME tu es optimiste ! #tnelec`.

- Remplacer tous le(s) URL(s) par l'expression : URL. Par exemple : `Radio-Canada dément avoir publié "La mère de Stéphen Harper renie son fils: http://is.gd/ioeg2p Original: http://bit.ly/l9a3mI"` sera `Radio-Canada dément avoir publié "La mère de Stéphen Harper renie son fils: URL Original:URL"`.

- Supprimer les lettres répétées trois fois ou plus par deux. Par exemple : `i am very huuuuungry` sera `i am very huungry`

Ce travail considère les tweets comme étant toujours subjectifs i.e. exprimant des sentiments. Un tweet, qui ne contient aucun sentiment, va avoir une polarité positive ou négative. Alors que ce tweet doit être considéré comme objectif et ne peut pas avoir une polarité.

Ce travail propose une bonne idée qui permet d'éviter l'annotation manuelle de l'ensemble d'apprentissage.

[14] `http://twitrratr.com/`

2.3.2 Barbosa et Feng [2010]

Deux étapes ont été utilisées par Barbosa et Feng [2010] pour classifier les tweets :

1. Classification de subjectivité : déterminer si le tweet est subjectif ou non.
2. Classification de polarité : déterminer, parmi les tweets subjectifs, la polarité de chacun.

Les méthodes utilisées afin de réaliser ces tâches, sont basées sur l'apprentissage supervisé. Pour collecter les données d'apprentissage, les auteurs ont eu recours à 3 applications qui analysent les sentiments en utilisant *Twitter* :*Twendz*[15], *TweetFeel*[16], *Twitter Sentiment* [2.3.1]. Ces applications retournent des tweets, qui contiennent le mot clé saisi par l'utilisateur, avec leurs classes. Pour recueillir des tweets génériques (qui portent sur différents sujets), les auteurs ont utilisé le mot clé `of`, un mot très fréquent en anglais qui permet de récupérer beaucoup de tweets.

Avant la classification, les auteurs ont effectué un filtrage :

- Suppression des tweets qui n'ont pas la même classe par les différentes applications.
- Ne conserver qu'un tweet par utilisateur ayant envoyé plusieurs tweets. Les auteurs ont constaté que la plupart de tweets, envoyés par des personnes qui publient fréquemment, sont des publicités ou des informations de recrutement.
- Élimination de l'ensemble d'apprentissage (objectif) des mots fortement subjectifs. Exemple : `awesome` Le but de cette tâche est de diminuer l'importance de mots subjectifs pour classifier les tweets objectifs.

Deux ensembles de *features* ont été exploités pour représenter les tweets :

- La façon dont le tweet est écrit :
  - S'il est un retweet.

[15] `http://twendz.waggeneredstrom.com/`
[16] `http://www.tweetfeel.com/`

  - S'il est une réponse.
  - Les liens : si le tweet contient un lien ou non.
  - S'il contient : des points d'exclamations et/ou d'interrogations, des émoticônes ou des hashtags.

- Meta-information sur les mots composant le tweet :
  - Leur classe grammaticale : adjectif, verbe...
  - Leur degré de subjectivité : si un mot est fortement subjectif ou non.
  - Leur polarité

La polarité et le degré de subjectivité des mots sont annotés à l'aide d'une base lexicale[17].

Les auteurs ont constaté que la polarité des mots est le *feature* le plus important pour déterminer la polarité des tweets subjectifs. Cependant, la polarité de certains mots peut changer selon le contexte du tweet. Pour cela, les auteurs ont vu qu'il n'est pas efficace d'utiliser la polarité fournie par la base lexicale. Ainsi, ils ont attribué une probabilité à la polarité d'un mot présent à la fois dans la base lexicale et dans l'ensemble d'apprentissage. la probabilité d'une polarité d'un mot *m* est déterminée comme suit :

$$polarite_{pos}(m) = \frac{nombre(positive, m)}{nombre(m)}$$

D'où $polarite_{pos}(m)$ est la probabilité que la polarité du *m* est positive, *nombre(positive,m)* est le nombre de fois où *m* est présent dans un tweet positif, *nombre(m)* est le nombre d'occurrences de *m* dans l'ensemble d'apprentissage. Si $polarite_{pos}(m) > 0{,}5$ la polarité sera considérée positive, négative sinon.

Barbosa et Feng [2010] ont proposé une autre méthode, différente à celle de Go et al. [2009] (section 2.3.1), pour la construction de l'ensemble d'apprentissage. Contrairement à l'approche de Go et al. [2009], les auteurs ont distingué entre les tweets objectifs et les tweets subjectifs. Ils ont également montré l'importance d'autres éléments dans la classification de sentiment tels que : les retweets, les hashtags...

[17] `http://www.cs.pitt.edu/mpqa/`

2.3.3 Jiang et al. [2011]

Comme dans le travail précédent, Jiang et al. [2011] ont utilisé deux classifieurs : un classifieur de subjectivité et un classifieur de polarité pour les tweets classifiés comme subjectifs.

Au début, les auteurs ont fait des tâches de prétraitement des données :

- Définir la classe grammaticale des mots
- Racinisation des mots : transformer chaque mot par sa racine (la racine des mots *traveling* et *traveled* est *travel*).
- Normalisation des mots : correction des fautes d'orthographe.

L'un des problèmes rencontrés pour la classification des tweets est qu'ils sont généralement ambigus et courts, ils ne contiennent pas assez d'informations. Il est difficile de déterminer la classe du tweet `First game: Lakers!`, qui contient seulement trois mots. En plus, le tweet `People everywhere love Windows & vista. Bill Gates` qui n'exprime pas aucun sentiment sur `Bill Gates`, mais il peut être classifié comme positifs par les méthodes de Go et al. [2009] et Barbosa et Feng [2010] vu la présence du terme `Love`.

Les auteurs ont vu qu'il faut sélectionner seulement les termes qui portent sur le sujet cible et qu'il ne suffit pas considérer que le tweet à classifier.

Pour résoudre le premier problème les auteurs ont décidé de :

- Sélectionner les syntagmes nominaux incluant le sujet cible. Si le sujet est `Microsoft`, `Microsoft technology` est un syntagme nominal.
- Sélectionner les pronoms et les groupes nominaux qui réfèrent au sujet. Dans le tweet `Oh, Jon Stewart. How I love you so.`, `You` sera sélectionné parce qu'il réfère à `Jon Stewart` (le sujet).
- Sélectionner les noms qui sont fortement associés au sujet, c'est-à-dire qui cooccurrent souvent avec le sujet. Le degré d'association est calculé à l'aide de PMI [1.2.2].

- Sélectionner les *features*, qui ont une relation avec le sujet ou l'un de ses attributs (les termes sélectionnés auparavant), en se basant sur un ensemble de règles. Le tableau 2.III présente deux exemples de règles.

Tableau 2.III – Exemple de règles pour la sélection des features

Sujet	**Règle**	**Exemple**	**Feature**
iPhone	Si le mot *m* est un verbe transitif et le sujet (ou l'un de ses attributs) est son objet alors le *feature m_arg2* est généré	`I love iPhone`	*love_arg2*
John	Si le mot *m* est un adjectif ou un verbe intransitif se trouve seul dans une phrase et le sujet (ou l'un de ses attributs) apparaît dans la phrase précédente alors le *feature m_arg* est généré	`John did that. Great!`	*great_arg*

Les features utilisés par Barbosa et Freng (section 2.3.2) sont aussi employés dans ce travail.

Pour le deuxième point qui consiste à ne pas considérer seulement le tweet à classifier, les auteurs ont envisagé d'étudier les polarités de ses voisins (les tweets qui ont une relation avec lui).

Dans ce contexte trois types de relation entre tweets ont été étudiés :

- Les retweets.
- Les tweets envoyés par le même utilisateur et qui portent sur le sujet.
- Tweets répondant à ou répondus par le tweet à classifier.

Suite à ces relations, les tweets qui portent sur le même sujet peuvent être présentés dans un graphe G. La probabilité qu'un tweet tw appartienne à une classe sera calculée de la façon suivante :

$$p(c|\tau, G) = p(c|\tau) \sum_{N(tw)} p(c|N(tw))p(N(tw))$$

D'où c est la classe (neutre, positive, négative) à laquelle *tw* peut appartenir, τ est le contenu de *tw*. *p(c|τ)* est la probabilité que *tw* appartienne à la classe c sachant son contenu τ. *N(tw)* est l'ensemble de polarités attribuées aux tweets connectés à *tw*.

Initialement, ils ont calculé pour chaque tweet les probabilités d'appartenance aux différentes classes. Par la suite, ils ont appliqué la méthode de relaxation qui permet d'ajuster itérativement les probabilités de chaque tweet à classifier en tenant compte des probabilités de ses voisins. À la fin des itérations, la classe qui a la plus grande valeur de $p(c|\tau, G)$ sera considérée.

Cet article a pris en considération la taille (généralement courte) et l'ambiguïté de texte. Les auteurs ont montré que ces facteurs peuvent influencer sur la classification d'un tweet. Ils ont prouvé que la relation entre les tweets peut résoudre ces problèmes et améliorer l'exactitude de la classification.

2.3.4 Tan et al. [2011]

Les auteurs ont considéré que les utilisateurs qui sont en relation sont plus susceptibles d'avoir le même sentiment. Les relations entre les utilisateurs sont représentées dans un graphe d'où chaque utilisateur est connecté aux utilisateurs qui sont en relation avec lui. Les auteurs ont étudié quatre types de graphe :

- *Directed to-follow graph* : utilisateur u_i suit u_j par contre l'inverse n'est pas nécessaire.
- *Mutual t-follow graph* : utilisateur u_i suit u_j et vice versa.
- *Directed @ graph* : utilisateur u_i a mentionné u_j par contre l'inverse n'est pas nécessaire.
- *Mutual @ graph* : utilisateur u_i a mentionné u_j et vice versa.

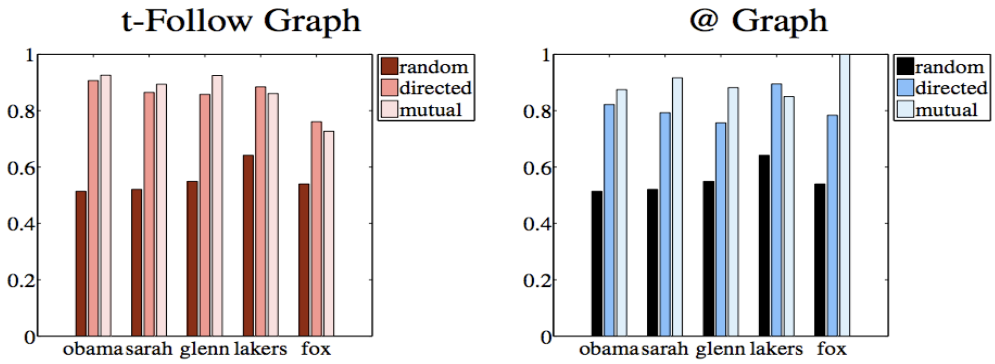
La méthode proposée sert à déterminer la polarité de sentiment de l'utilisateur pour un sujet et non pas la polarité d'un tweet. L'idée est de représenter les utilisateurs qui ont publié des tweets portant sur un sujet `s` dans un graphe. Une portion de ces utilisateurs sont déjà annotés manuellement, le but donc est de prédire les étiquettes des autres utilisateurs.

Dans la collecte des données, les auteurs ont sélectionné les utilisateurs dont les opinions sont claires. Un utilisateur qui cite dans sa description `anti-Obama and America FIRST` ou son nom d'utilisateur est `against_obama`, son sentiment pour le sujet `Obama` est considéré comme négatif.

Les auteurs ont effectué des statistiques sur les données collectées pour estimer le degré d'exactitude de leur hypothèse (qui considère que deux utilisateurs connectés partagent le même sentiment pour un sujet `s`) :

- La probabilité que deux utilisateurs aient la même étiquette sachant qu'ils sont connectés (Figure 2.2).

Figure 2.2 – Probabilité que deux utilisateurs aient le même sentiment sachant le type de leur relation (Figure1 Tan et al. [2011]). L'axe des abscisses : les sujets traités ; l'axe des ordonnées : les probabilités ; random : paires des utilisateurs choisies aléatoirement ; directed : existe au moins un lien deux utilisateurs ; mutual : deux utilisateurs sont reliés dans les deux sens.

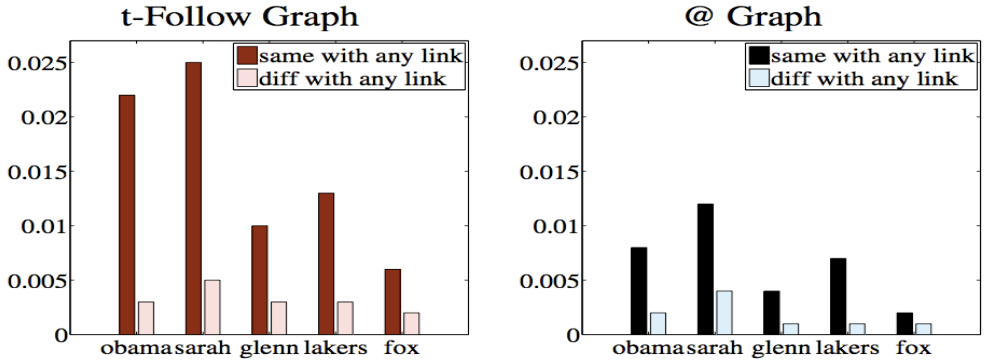


- Probabilité que deux utilisateurs soient connectés, sachant qu'ils ont le même sentiment (Figure 2.3).

Le modèle utilisé pour déterminer l'étiquette d'un utilisateur intègre les informations de réseau social (les relations entre les utilisateurs) et les tweets des utilisateurs. Les auteurs ont considéré que les tweets, d'un utilisateur avec une étiquette positive sont positifs.

Les auteurs dans ce travail ont traité le problème de la classification de sentiment pour les textes courts, tels que les tweets, avec une nouvelle manière qui intègre les

Figure 2.3 – Probabilité que deux utilisateurs soient connectés sachant qu'ils ont le même sentiment (Figure2 Tan et al. [2011]). L'axe des abscisses : les sujets traités ; l'axe des ordonnées : les probabilités.



techniques d'analyse des réseaux sociaux et qui ne donne pas une grande importance au contenu de tweets.

2.4 Tweets vs Événements

Les récents qui s'intéressent à l'analyse des textes courts ne visent pas seulement à déterminer la polarité des messages, mais à utiliser les messages pour détecter des événements ou de prédire des résultats.

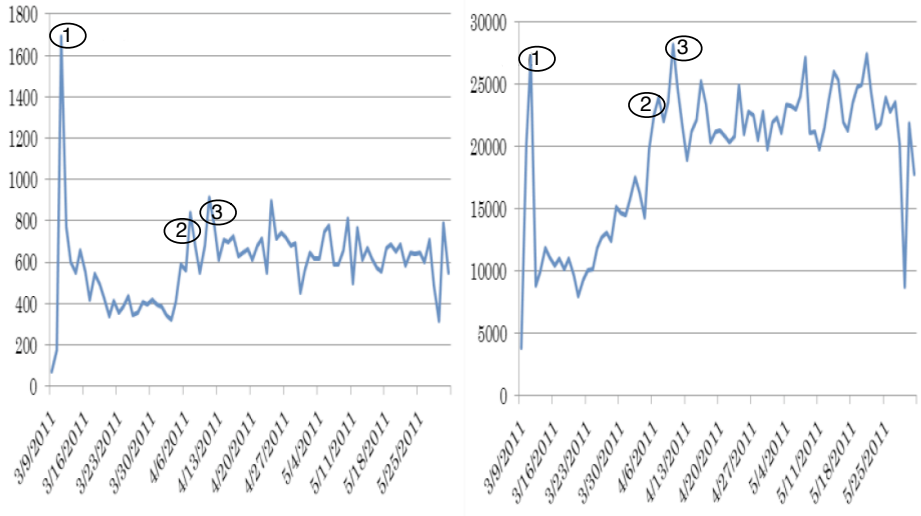
2.4.1 Doan et al. [2011]

Doan et al. [2011] ont analysé les niveaux de sensibilisation et d'anxiété des habitants de Tokyo pour les événements de tremblements de terre, tsunami et les états d'urgences nucléaires au Japon en 2011.

Trois grands tremblements de terre ont eu lieu : le premier a frappé le 11 mars, le second était le 7 avril et le dernier était le 11 avril.

Les auteurs ont utilisé plus que de 1,5 million tweets (48 870 en anglais et 1 611 753 en japonais) envoyés du Japon, pour la période de 9 mars 2011 au 31 mai 2011. La Figure 2.4 présente la distribution des tweets par dates.

Figure 2.4 – Nombre de tweets par date en anglais (à gauche) et en japonais (à droite) (Figure1 Doan et al. [2011]). 1- 31/03/2011 (premier tremblement de terre) ; 07/04/2011 (second tremblement de terre) ; 11/04/2011 (troisième tremblement de terre)



Trois ensembles de mots clés ont été utilisés pour détecter parmi les tweets collectés précédemment ceux qui portent sur la sensibilisation des gens aux tremblements de terre et tsunami, la radiation nucléaire et/ou l'anxiété des habitants (voir Figure 2.5).

Figure 2.5 – Liste des mot clés pour le tremblement de terre et tsunami, radiation et l'anxiété des habitants (tableau1 Doan et al. [2011])

English terms	Japanese terms
Earthquake and Tsunami event	
earthquake, quake, quaking, post-quake,shake, shaking,shock, aftershock,temblor, tremor, movement, sway, landslide seismic, seismography, seismometer, seismology, tsunami, wave	大地震，大震災 震災，地震，余震，揺れ 震度，震源，マグニチュード 津波
Radiation event	
radiation, nuclear, reactor, radioactivity, radioactive, iodine, TEPCO, meltdown, explosion, power plant, micro sievert	放射，放射線，放射能，放射性物質，原発，マイクロシーベルト，ヨウ素，イソジン，ヨウ化カリウム，炉心溶融，メルトダウン
Anxiety event	
die, scary, scared, incredible, worrying, worried, anxious, annoying	死亡，死ぬ，やられてる，やばい，やばかった，ヤバい，やばっ，やべ，怖い，怖かった，怖っ，すごい，すげえ，すげー，すっげー，びびる，びびった，混乱，微妙，避難，助けて，わかり辛い，連絡とれない，大変，心配，恐れ，船酔いしそう

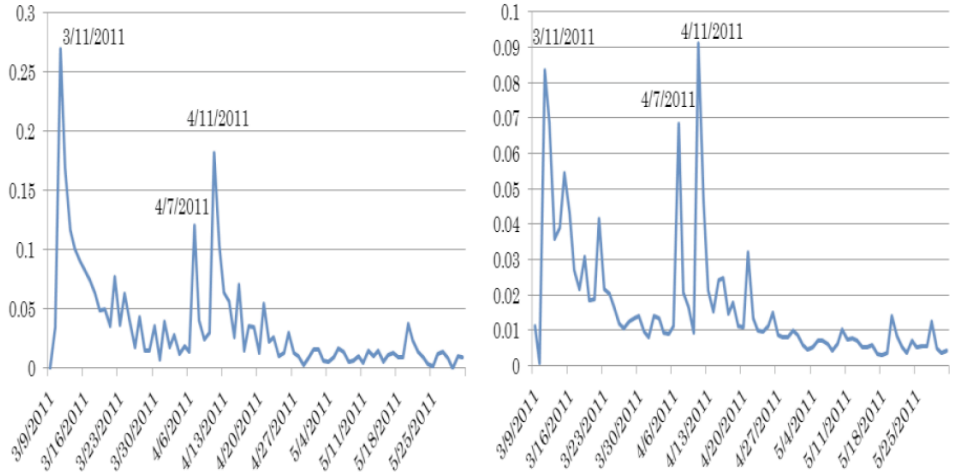
La présence d'un événement *E* dans les tweets est calculée à l'aide de la fréquence relative :

$$F(E) = \frac{nombre\ de\ tweets\ qui\ contiennent\ un\ mot\ clé\ de\ l'événement\ par\ jour}{nombre\ total\ de\ tweets\ par\ jour}$$

La fréquence des mots clés liés aux tremblements de terre est présentée dans la figure 2.6. D'après cette figure, on peut observer une forte corrélation entre le contenu des tweets et la réalité. Les auteurs ont constaté également que le premier tweet, qui porte sur le premier tremblement de terre, était envoyé seulement après une minute et 25 secondes de l'événement.

Ce travail montre que les internautes tiennent à diffuser leurs expériences en temps réel. Il a prouvé qu'à partir des données trouvées dans les *microblogs*, nous pouvons dégager les préoccupations et les intérêts des utilisateurs.

Figure 2.6 – Fréquence des mots clés liés aux tremblements de terre (Figure2 Doan et al. [2011]).



2.4.2 Lampos et Cristianini [2010]

Lampos et Cristianini [2010] ont utilisé Twitter pour mesurer la prévalence de la maladie H1N1 pour la population de la Grande-Bretagne. Les auteurs ont collecté des tweets chaque jour pendant 24 semaines : de 22/06/2009 jusqu'à 06/12/2009 qui proviennent de 5 régions différentes dans la Grande-Bretagne. Par la suite, ils ont supprimé les mots vides (stop words) et ont appliqué un algorithme de racinisation.

La méthode proposée dans ce travail consiste à rechercher dans les tweets les symptômes liés à la grippe H1N1 puis à retourner un score intitulé *Flu_score*. Les auteurs ont utilisé un ensemble $\mathscr{M}$ qui contient 41 indices (temperature, headache, sore throat...) pouvant exprimer la grippe. Le *Flu_score* d'un tweet t est calculé comme suit :

$$s(t) = \frac{\sum_i m_i(t)}{k}$$

D'où m_i est $i^{ème}$ indice et k est le nombre des indices, m_i(t) est égale à 1 si m_i apparaît dans le tweet et 0 sinon. Le *Flu_score* des tweets d'un jour j est calculé comme suit :

$$f_j = \frac{\sum_q s(t_q)}{n}$$

D'où t_q est le $q^{ème}$ tweets au jour j et n le nombre de tweets dans j.
Afin d'évaluer les résultats obtenus, les auteurs ont décidé de les comparer avec les données de l'agence de protection de la santé (APS). Le tableau 2.IV montre les coefficients de corrélation entre les deux résultats dans les 5 régions.

Tableau 2.IV – Coefficients de corrélation obtenus entre les *Flu_score* obtenus et les données de l'APS

Région	**coefficient de corrélation**
A	0.8471
B	0.8293
C	0.8438
D	0.8556
E	0.8178

Les auteurs ont essayé d'améliorer les coefficients de corrélation en attribuant un poids à chaque indice. le calcul de *Flu_score* est devenu comme suit :

$$s_w(t) = \frac{\sum_i w_i \times m_i(t)}{k}$$

$$f_{w,j} = \frac{\sum_q s_w(t_q)}{n}$$

D'où w_i est le poids de l'indice m_i.

Dans le but d'apprendre les poids de chaque indice, ils ont appliqué la méthode des moindres carrés entre les *Flu_scores* (non pondérés) obtenus et les données de l'APS. Ils ont utilisé comme ensemble d'apprentissage les données qui correspondent à une région, puis ils ont évalué les poids inférés sur les données des autres régions. Le tableau 2.V montre les résultats obtenus. La valeur 0.9487 (en gras) est le coefficient de corrélation entre les *Flu_scores* obtenus et les données de l'APS pour la région D en employant les poids appris dans la région B.

Tableau 2.V – Coefficients de corrélation obtenus (par région) avec l'utilisation des poids (tableau2 Lampos et Cristianini [2010]. L'élément (i, j) désigne le coefficient de corrélation entre le *Flu_score* pondéré et les résultats de APS sur la région j, après avoir entrainé les poids sur la région i

Train/Test	**A**	**B**	**C**	**D**	**E**	**Moyenne**
A	-	0.8389	0.9605	0.9539	0.9723	0.9314
B	0.7669	-	0.8913	**0.9487**	0.8896	0.8741
C	0.8532	0.702	-	0.8887	0.9445	0.8471
D	0.8929	0.9183	0.9388	-	0.9749	0.9312
E	0.9274	0.8307	0.9204	0.9749	-	0.9134

Ils ont également essayé de prédire le coefficient de corrélation en utilisant les données qui correspondent à toutes les régions. D'où les données entre les semaines 28 et 41 sont les données de test et le reste est utilisé pour apprendre les poids.

Les auteurs ont tenté d'extraire automatiquement les indices. Ils ont collecté un ensemble de candidats obtenu à partir des articles dans le web liés à la grippe.

L'ensemble contient 1560 candidats. Par la suite, ils ont appliqué le même principe utilisé précédemment (apprendre les poids). Le nombre de candidats retenus (leurs poids >0) est 73. Une corrélation qui dépasse 95% a été obtenue.

Ce travail a confirmé les résultats obtenus dans Doan et al. [2011] (section 2.4.1). Il a prouvé que les utilisateurs tiennent à partager leurs expériences via des plates-formes tel que Twitter. L'analyse de ces expériences peut produire des résultats qui reflètent des données réelles.

Dans le cadre de l'extraction des événement à partir des tweets, Metzler et al. [2012]

ont développer un framework qui un mot clé lié à un événement particulier est de retourner un résumé qui répond à cette requête. Le résumé contient l'heure de début qui indique quand l'événement a commencé à être discuté, une durée qui spécifie combien de temps l'événement a été discuté, et un petit nombre de messages postés pendant cet intervalle de temps. Dans le même contexte, Chakrabarti et Punera [2011] ont proposé une méthode qui permet de produire des résumés à partir des tweets (en temps réel) qui portent sur un événement e.

OConnor et al. [2010] ont montré qu'il y a une forte corrélation entre des données réelles provenant des sondages et l'opinion publique mesurée à partir des tweets. Les élections américaines en 2008 étaient parmi les sujets analysés.

2.5 Conclusion

Dans ce chapitre, nous avons présenté les caractéristiques et les particularités des microblogs, notamment Twitter.

Nous avons décrit des travaux qui s'intéressent à l'analyse des tweets. Ceux-ci nous ont apporté des idées pour le traitement des textes avec une taille réduite. La plupart de ces travaux traitent les tweets écrits en anglais. Néanmoins dans ce travail nous intéressons aux textes en français et en arabe. Dans le chapitre suivant, nous montrerons le corpus que nous avons collecté et les statistiques que nous avons calculées.

CHAPITRE 3

EXPÉRIMENTATIONS

Dans le chapitre précédent, nous avons présenté les différences entre les microblogs et les blogs traditionnels. Nous avons également présenté quelques travaux qui analysent les données provenant des microblogs, particulièrement Twitter. La majorité de ces travaux traitent les tweets écrits en anglais.

Pour collecter des données, nous avons implémenté un programme Java qui a utilisé la bibliothèque Twitter4j. Cette bibliothèque permet d'accéder aux données (tweets, les informations des utilisateurs...) Twitter via son interface de programmation, Twitter API. Nous présentons, dans ce chapitre, les corpus que nous avons collectés et les expérimentations que nous avons faites. Nous avons étudié principalement le contenu des tweets (leurs tailles, les mots les plus fréquents, les mots connus par un lexique français), les préoccupations des utilisateurs selon les hashtags utilisés, le comportement des utilisateurs...

Dans nos expérimentations, nous avons traité les tweets qui portent sur la Tunisie et qui sont écrits avec des langues autres que l'anglais soit le français et l'arabe (voir tableau 2.II).

Nous avons été amenés vers ce type de textes à cause de nos compétences qui nous permettent de comprendre le français et l'arabe, particulièrement la façon d'écrire des Tunisiens qui comporte des abréviations, des fautes de grammaire et d'orthographe, des mots arabes écrits avec des alphabets français et chiffres et différentes langues dans le même tweet.

3.1 Extraction des données

Notre première étape dans ce travail consistait à nous familiariser avec `Twitter API`[1]. Au début, nous avons utilisé `search API` qui permet de retourner des tweets

[1] `https://dev.twitter.com/docs`

qui répondent à une requête q. Si q = 'apple store', search API[2] retourne les tweets qui contiennent les deux termes apple et store. On peut également filtrer les résultats selon plusieurs critères tels que :

Langue des tweets spécifier la langue avec laquelle le tweet est écrit.

La période trouver les tweets écrits entre une date since et une date until.

Type de résultats spécifier le type de tweets retournés les plus populaires (les plus retweetés), les plus récents ou mixtes (mélange entre les plus populaires et les plus récents).

La search API a des limites :

- Le nombre de tweets retournés par requête ne peut pas dépasser 1500.
- Il ne peut pas trouver les tweets qui étaient envoyés il y a plus qu'une semaine.

La figure 3.1 montre un exemple de tweet retourné par search API.

Figure 3.1 – Tweet retourné par search API (format *Json*). Ligne 1 : texte du tweet ; Ligne 4 : langue du text (identifier par Twitter).

```
01 text='Ok, jeudi il fera 31 degré. #Tunisie #auMax', id=193873044287143936
02 toUserId=-1, toUser='null',
03 fromUser='AnisKhez', fromUserId=121504252,
04 isoLanguageCode='it',
05 source='<a href="http://twitter.com/#!/download/iphone"rel="nofollow">Twitter
   for iPhone</a>',
06 profileImageUrl='http://a0.twimg.com/profile_images
   2008493088/229604_10150324090608888_573338887_7473932_392125_n_normal.jpg',
07 createdAt=Sat Apr 21 21:25:10 EDT 2012,
08 location='null',
09 place=null, geoLocation=null, annotations=null,
10 userMentionEntities=[],
11 urlEntities=[],
12 hashtagEntities=[HashtagEntityJSONImpl{start=28, end=36, text='Tunisie'},
   HashtagEntityJSONImpl{start=37, end=43, text='auMax'}],
13 mediaEntities=null
```

Par la suite, nous avons utilisé la streaming API[3] qui permet d'obtenir les tweets en temps réel. On peut filtrer les tweets avec plusieurs (jusqu'à 400) mots-clés. Exemple : 'élections Tunisiennes', 'Tunisie', 'Tunisia', 'Ennahdha', 'CPR', 'PCOT',

[2] https://dev.twitter.com/docs/using-search
[3] https://dev.twitter.com/docs/streaming-api

‘PDP’, ‘PDM’…. Où Ennahdha, CPR, PCOT, PDP et PDM sont des partis politiques tunisiens.

On peut également filtrer les tweets selon leur positionnement géographique. Par exemple, pour récupérer les tweets qui portent sur le mouvement *Occupy Wall Street* et qui proviennent de Montréal, on doit utiliser des mots-clés qui peuvent décrire ce mouvement et les latitudes/longitudes qui correspondent à Montréal. Seuls les tweets créés à l’aide de l’option de géolocalisation sont sélectionnés.

Cet API nous permet d’avoir un nombre plus important de tweets distincts, mais on ne peut pas spécifier la langue de tweets. La figure 3.2 montre un extrait de tweet retourné par streaming API. Le résultat retourné par la streaming API contient plus d’informations que celui retourné par search API telles que : Les informations de l’utilisateur, nombre de fois que le tweet a été retweeté…

Figure 3.2 – Extrait d'un tweet retourné par `streaming API` (format *Json*). Ligne 1 : informations sur le tweet (date, identifiant) ; Ligne 2 : le texte du tweet ; Ligne 21 jusqu'à ligne 32 : informations de l'utilisateur.

```
01 createdAt=Sun Apr 22 00:38:15 EDT 2012, id=193921636305604609,
02 text='RT @tunistribune: Nouvelle publication L'ex-dictateur Ben Ali négocie son retour
   en Tunisie http://t.co/vWM3t4aM',
03 source='web'
...
05 inReplyToScreenName='null', geoLocation=null,
06 place=null, retweetCount=0, wasRetweetedByMe=false,
...
08 retweetedStatus=StatusJSONImpl{createdAt=Sat Apr 21 21:22:36 EDT 2012, id=193872400662802432,
09 text='Nouvelle publication L'ex-dictateur Ben Ali négocie son retour en Tunisie http://t.co/vWM3t4aM',
10 source='<a href="http://www.hootsuite.com" rel="nofollow">HootSuite</a>',
11 geoLocation=null, place=null, retweetCount=0,
...
13 user=UserJSONImpl{id=259789350, name='Tunis Tribune', screenName='tunistribune',
14 location='', description='',
...}}
16 userMentionEntities=[UserMentionEntityJSONImpl{start=3, end=16, name='Tunis Tribune',
17 screenName='tunistribune', id=259789350}],
18 urlEntities=[URLEntityJSONImpl{start=92, end=112, url=http://t.co/vWM3t4aM,
19 expandedURL=http://ow.ly/1iYwmB, displayURL=ow.ly/1iYwmB}],
20 hashtagEntities=[],
21 user=UserJSONImpl{id=226180205,name='mouldi  amamou', screenName='josezit', location='TUNIS TUNISIE',
22 description='Prof Hospitalouniversitaire Medecine# Unv Tn El Manar# Fac Med Tunis# Interniste
23 Réanimateur Médical# Centre Assistance Médicale Urgente# Ministère Santé#',
...
25 profileImageUrl='http://a0.twimg.com/profile_images/1739244207/07012012249_normal.jpg',
...
27 url='http://www.mouldiamamou.com', isProtected=false, followersCount=438,
28 profileBackgroundColor='C0DEED',
...
29 friendsCount=291, createdAt=Mon Dec 13 09:39:59 EST 2010, timeZone='Paris'
30 ,profileBackgroundImageUrl='http://a0.twimg.com/images/themes/theme1/bg.png',
...
32 lang='fr', statusesCount=2196,
...}
```

3.2 Statistiques et interprétations

3.2.1 Expérience 1

Comme mentionné dans la section précédente, nous avons commencé par `search API` pour extraire des tweets. Nous avons sélectionné des tweets écrits en anglais qui répondent à différentes requêtes telles que : `Apple`, `Microsoft`, `Obama`...

Nous avons essayé de déterminer la polarité des tweets en nous basant sur les termes

positifs et négatifs trouvés dans le tweet. La polarité du tweet est :

positive	Si le nombre(termes positifs) − nombre(termes négatifs) > 0
négative	Si le nombre(termes positifs) − nombre(termes négatifs) < 0
neutre	Sinon

L'orientation sémantique (positive, négative) des termes est déterminée à l'aide de lexique utilisé dans *twitrratr*[4]. Nous avons constaté que la plupart des tweets sont considérés neutres. Généralement ces tweets ne contiennent pas des termes présentés dans le lexique utilisé.

3.2.2 Expérience 2

Nous avons encore collecté des tweets qui portent sur les élections tunisiennes d'octobre 2011. Ceci est établi à l'aide de `Search API`. Le mot-clé que nous avons utilisé est `tnelec`, `tn` indique Tunisie et `elec` indique élections. Nous avons exploité ce terme parce qu'il est très employé par les utilisateurs qui parlent de cet événement. Nous avons collecté des tweets en différentes langues : arabe, français et anglais. Le tableau 3.I montre des statistiques sur le corpus collecté. Nous avons remarqué que la plupart des tweets contiennent au moins un hashtag. Ceci peut être expliqué par le mot clé utilisé dans la requête : `tnelec`. Ce mot est souvent utilisé comme un hashtag

Même si nous avions spécifié la langue des tweets, nous avons trouvé que les tweets contiennent des termes qui ne sont pas écrits avec la langue spécifiée. Le tweet suivant est sélectionné avec les tweets écrits en français :

تصويرة بن علي رجعت في حلق الوادي *http://t.co/2w4Ishy0 excellent !!! #tnelec*

Le texte arabe signifie que la photo de `Ben Ali` (président tunisien déchu) a été republiée dans `La Goulette` (ville tunisienne).

Nous avons également effectué des statistiques sur le nombre de mots trouvés dans les tweets (tableau 3.II).

[4] `http://twitrratr.com/`

Tableau 3.I – Statistiques sur les tweets qui portent sur les élections tunisiennes publiés entre le 18 octobre 2011 et 28 octobre 2011. 11 108 tweets en français, 5 552 en arabe et 4 377 en anglais

	Fr	**Ar**	**En**
Nombre de tweets	11 108	5 552	4 377
Nombre de retweets	4 555	2 759	2 141
Nombre d'utilisateurs distincts	2 066	1 649	1 707
Nombre de tweets qui contiennent au moins un hashtag	11 079	5 532	4 353
Nombre de tweets qui contiennent au moins un utilisateur mentionné	2 525	756	1 112
Nombre de tweets qui contiennent au moins un hyperlien	3 490	1 599	1 729
Nombre de hashtags distincts	1 059	373	527

Tableau 3.II – Statistiques sur le nombre des mots trouvés dans les tweets qui portent sur les élections tunisiennes envoyés entre le 18 octobre 2011 et 28 octobre 2011. Le tweet nettoyé ne contient pas les *hashtags*, les *utilisateurs mentionnés* et les *hyperliens*

	Original			**Nettoyé**		
	Fr	***Ar***	***En***	***Fr***	***Ar***	***En***
Nombre minimal de mots	2	2	1	0	1	0
Nombre maximal de mots	33	34	30	30	30	28
Nombre moyen de mots	17	17	17	14	13	12

Les lexiques *Morphalou*[5] et *BDE* ont été utilisés pour vérifier l'existence d'un mot respectivement dans le vocabulaire français et anglais (voir tableau 3.III) . Nous avons considéré comme mot toute suite de caractères non blancs de longueur non nulle.

Tableau 3.III – Statistiques sur le nombre de mots trouvés dans les vocabulaires français et anglais pour les tweets (11 108 en français et 4 377 en anglais) qui portent sur les élections tunisiennes publiés entre le 18 octobre 2011 et 28 octobre 2011)

	Fr	**En**
Nombre minimal des mots	0	0
Nombre maximal de mots	27	25
Nombre moyen de mots	10	9
Nombre de mots distincts	12 638	5 695
Nombre de mots distincts trouvés dans le vocabulaire	6 325	2 838

[5] `http://www.cnrtl.fr/lexiques/morphalou/`

3.2.3 Expérience 3

Cette fois-ci, nous nous sommes intéressé aux tweets qui portent sur la Tunisie, et non pas sur un événement ou une personne en particulier. Pour recueillir des tweets, nous avons utilisé le `streaming API`.

Nous avons utilisé un ensemble de mots-clés, illustré dans le tableau 3.IV, fortement liés à la Tunisie.

Tableau 3.IV – Ensemble de mots-clés utilisés pour extraire des tweets (qui portent sur la Tunisie entre le 07 février 2012 et le 10 mars 2012) à l'aide du `streaming API`

Mots-clés	**Définition**
`marzouki`	Président actuel de la Tunisie
`hammadi jebali`	Premier ministre actuel
`Tunisie`, `tounes`	`tounes` est la prononciation arabe de Tunisie
`tnelec`	Les élections tunisiennes, nous avons utilisé ce terme parce que nous avons trouvé que beaucoup d'utilisateurs en parlent jusqu'à présent
`sebsi`	Ex-premier ministre (après la révolution tunisienne)
`nahdha`, `ennahdha`	Le parti politique qui a gagné dans la dernière élection.
`ghannouchi`	Chef d'`ennahda`
`sidi bouzid`	La région où la révolution tunisienne a commencé
`14jan`	14 janvier est la date de fuite de `Ben ali` (président déchu)

Nous avons réussi à extraire 126 991 tweets entre le 08 février 2012 et le 09 mars 2012. La figure 3.3 montre la distribution des tweets selon les dates. Le 20 février 2012, nous avons eu une panne technique, il y a eu une interruption entre 7 h 35 et 15 h 38. Cela explique la diminution de nombre de tweets à cette date.

Les tableaux 3.VI et 3.V illustrent des statistiques effectuées sur le corpus obtenu.

Figure 3.3 – Nombre de tweets par jour sur la Tunisie (126 991 tweets entre le 08 février 2012 et le 09 mars 2012)

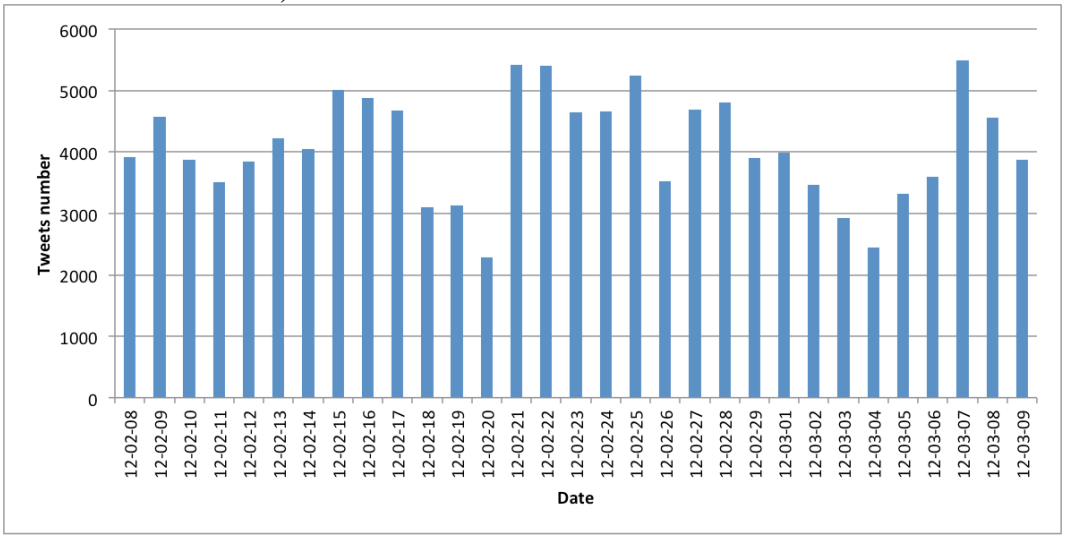


Tableau 3.V – Statistiques sur le nombre des mots trouvés dans les tweets qui portent sur la Tunisie (126 991 tweets publiés entre le 08 février 2012 et le 09 mars 2012). Le tweet nettoyé ne contient pas les *hashtags*, les *utilisateurs mentionnés* et les *hyperliens*

	Original	**Nettoyé**
Nombre minimal de mots	1	0
Nombre maximal de mots	38	37
Nombre moyen de mots	14,9	12,3

Tableau 3.VI – Statistiques sur les tweets qui portent sur la Tunisie publiés entre le 08 février 2012 et le 09 mars 2012).

Numbre de tweets	126 991
Nombre des retweets	23 283
Nombre d'utilisateurs distincts	16 071
Nombre de tweets qui contiennent au moins un hashtag	72 564
Nombre de tweets qui contiennent au moins un utilisateur mentionné	48 750
Nombre de tweets qui contiennent au moins un hyperlien	79 250
Nombre de hashtags distincts	7 879

Étant donné la difficulté de comprendre ce type de texte (courts, fautes d'écriture, convention d'écriture employée par les utilisateurs…), nous avons essayé de nous ba-

ser sur d'autres éléments tels que : les hashtags, le comportement des utilisateurs, les relations entre les utilisateurs...

Nous avons constaté que les utilisateurs utilisent souvent de hashtags, cet élément joue un rôle important pour avoir une idée sur les préoccupations des utilisateurs.

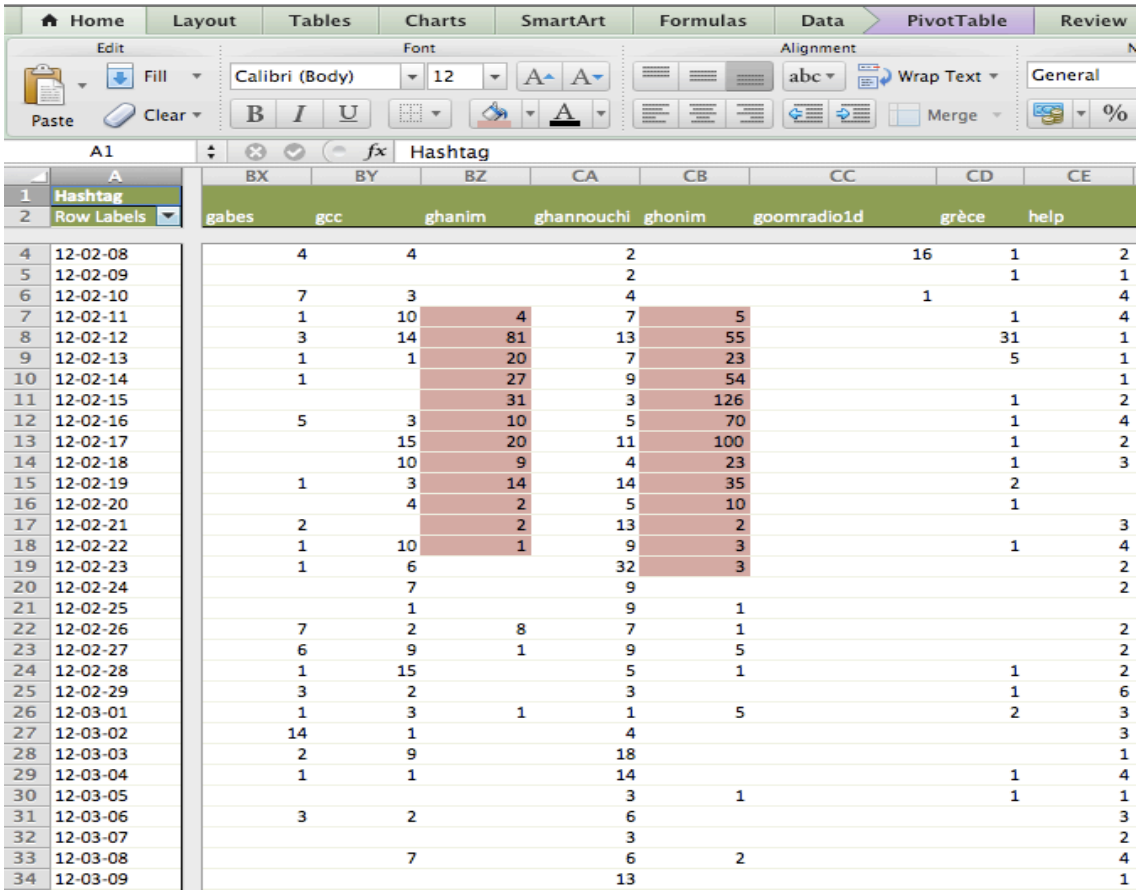
Le tableau 3.VII montre les hashtags, qui ne sont pas les mots-clés utilisés dans la requête, les plus présents dans le corpus. Ces résultats reflètent très bien les préoccupations des Tunisiens.

Tableau 3.VII – Les hashtags (qui ne sont pas des mots-clés) les plus fréquents dans le corpus de tweets qui portent sur la Tunisie (126 991 tweets entre le 08 février 2012 et le 09 mars 2012)

Hashtags	**nombre d'apparition**	**Définition**
`Tngov`	4 224	Le gouvernement tunisien
`Tnac`	3 578	`tn` réfère à Tunisie, `ac` réfère à l'assemblée constituante qui est chargée par la rédaction de la nouvelle constitution
`Ugtt`	1 847	Union générale tunisienne du travail, qui a organisé des grèves et des manifestations contre le gouvernement
`Syria`	1 754	La Tunisie renvoie l'ambassadeur syrien, organisation de la conférence *amis de la Syrie* en Tunisie...
`Mbj`	1 665	Mustapha Ben Jaafar, le président de l'assemblée constituante
`Qatar`	1 555	le Qatar qui va faire des investissements en Tunisie, refuse d'extrader Sakhr Materi (gendre de l'ancien président), soutient (selon plusieurs personnes) le parti au pouvoir (`ennahdha`)...
`Emploi`	1 472	La majorité des tweets qui contiennent des hashtags liés à l'emploi sont des offres d'emploi.
`Recrutement`	1 368	
`jobs`	1 359	

Nous avons visualisé les hashtags présents dans le corpus dans d'un *tableau croisé dynamique* de *Excel* (voir figure 3.4). Nous avons remarqué que certains sujets stimulent l'intérêt des utilisateurs dans une période déterminée.

Figure 3.4 – Capture d'écran du *tableau croisé dynamique* pour les hashtags présents dans le corpus. Les colonnes sont les hashtags et les lignes sont les dates. L'élément *(i,j)* est le nombre de hashtags *j* à la date *i*.



Hashtag	BX	BY	BZ	CA	CB	CC	CD	CE
Row Labels	gabes	gcc	ghanim	ghannouchi	ghonim	goomradio1d	grèce	help
12-02-08	4	4		2		16	1	2
12-02-09				2			1	1
12-02-10	7	3		4		1		4
12-02-11	1	10	4	7	5		1	4
12-02-12	3	14	81	13	55		31	1
12-02-13	1	1	20	7	23		5	1
12-02-14	1		27	9	54			1
12-02-15			31	3	126		1	2
12-02-16	5	3	10	5	70		1	4
12-02-17		15	20	11	100		1	2
12-02-18		10	9	4	23		1	3
12-02-19	1	3	14	14	35		2	
12-02-20		4	2	5	10		1	
12-02-21	2		2	13	2			3
12-02-22	1	10	1	9	3		1	4
12-02-23	1	6		32	3			2
12-02-24		7		9				2
12-02-25		1		9	1			
12-02-26	7	2	8	7	1			2
12-02-27	6	9	1	9	5			2
12-02-28	1	15		5	1		1	2
12-02-29	3	2		3			1	6
12-03-01	1	3	1	1	5		2	3
12-03-02	14	1		4				3
12-03-03	2	9		18				1
12-03-04	1	1		14			1	4
12-03-05				3	1		1	1
12-03-06	3	2		6				3
12-03-07				3				2
12-03-08		7		6	2			4
12-03-09				13				1

Les figures 3.5, 3.6 et 3.7 illustrent la fréquence des hashtags liés à trois événements qui se sont déroulés dans cette période :

- L'arrivée du prédicateur égyptien `Wajdi Ghonim` en Tunisie le 11 février 2012. Ce prédicateur est considéré par certains comme étant radical. Cela est dû à ses prises de position polémiques concernant des sujets à controverse. Une de ses fameuses prises de position se rapporte au sujet de l'excision des fillettes. Certains accusent le prédicateur d'avoir pris une position favorable envers l'excision. 15 février 2012 : deux plaintes ont été déposées contre `Wajdi Ghonim`. 17 février 2012 : des manifestations contre l'arrivée de `Wajdi Ghonim`.

- Les employés municipaux ont effectué une grève qui est organisée par l'UGTT. Cette grève a débuté le 20 février 2012. Suite à cette grève, plusieurs locaux de l'UGTT ont été attaqués par des manifestants. L'UGTT a accusé le parti ennahdha d'en être l'instigateur. Le 25 février, des grandes manifestations (organisées par UGTT) ont réclamé le départ de gouvernement.

- La mise en berne du drapeau tunisien par un étudiant "salafiste" dans le bâtiment de la faculté des lettres, des arts et des humanités de Manouba. Une étudiante tunisienne a essayé de l'empêcher d'enlever le drapeau.
  Les salafistes sont des personnes qui sont jugées par la plupart des Tunisiens comme des musulmans extrémistes et radicaux. Cet acte, qui a eu lieu le 07 mars 2012, a entraîné une vague de colère chez plusieurs Tunisiens qui ont considéré qu'un tel acte est un outrage portant atteinte à leur dignité et la souveraineté du pays.

Les pics observés dans les figure 3.5 illustrent la corrélation entre l'occurrence des événements réels et les hashtags trouvés dans notre corpus. Nous avons constaté que pour la date du 26 février, de nouveaux hashtags relatifs à Wajdi Ghonim sont apparus après une période d'abscence. Cela peut être dû à la visite de l'islamologue Tariq Ramadan le 25 février 2012 où les gens font référence à la visite de Wajdi Ghonim.

Figure 3.5 – Distribution par jour de 36 hashtags (`wajdi_ghanim`, `ghanim`, `ghonim sheikghanim`...) liés à l'événement de `Wajdi Ghonim`. 11 février 2012 : arrivée de Wajdi Ghonim. 15 février 2012 : deux plaintes ont été déposées contre `Wajdi Ghonim`. 17 février 2012 : des manifestations contre l'arrivée de `Wajdi Ghonim`.

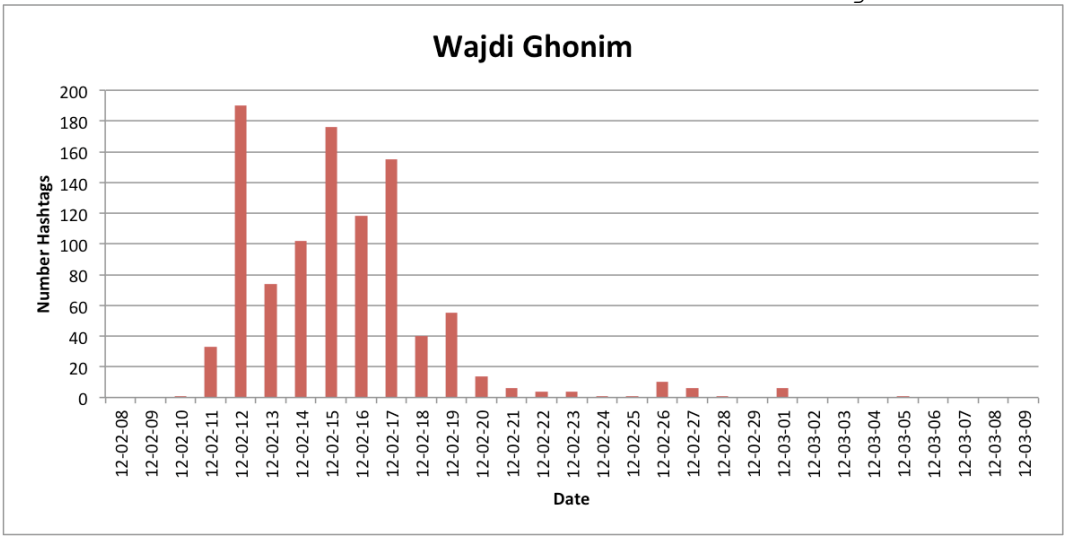


Figure 3.6 – Distribution par jour de 2 hashtags (`manouba`, `mannouba`) liés à l'événement de `Manouba`. 07 mars 2012 : La mise en berne du drapeau tunisien par un étudiant "salafiste".

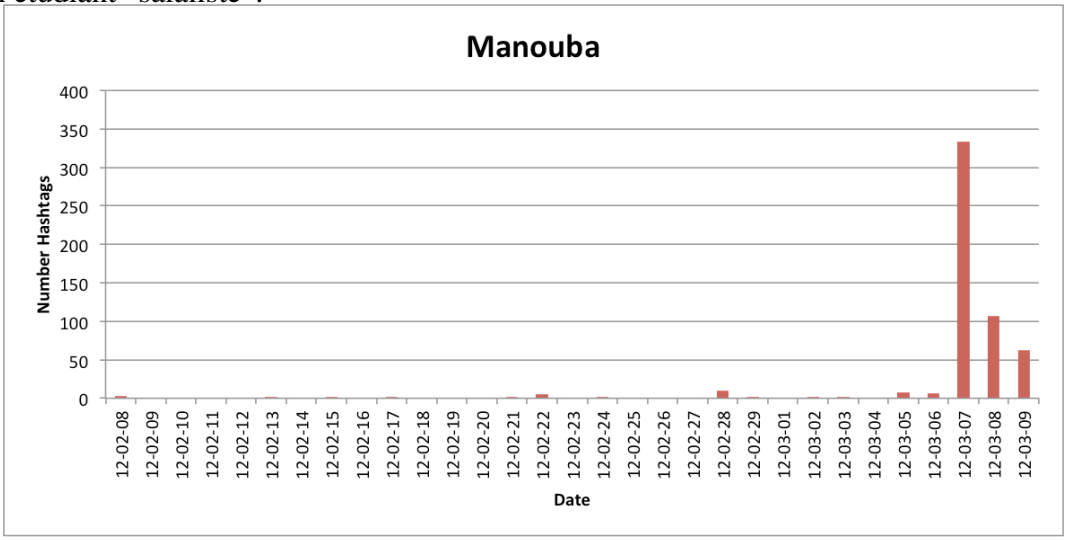
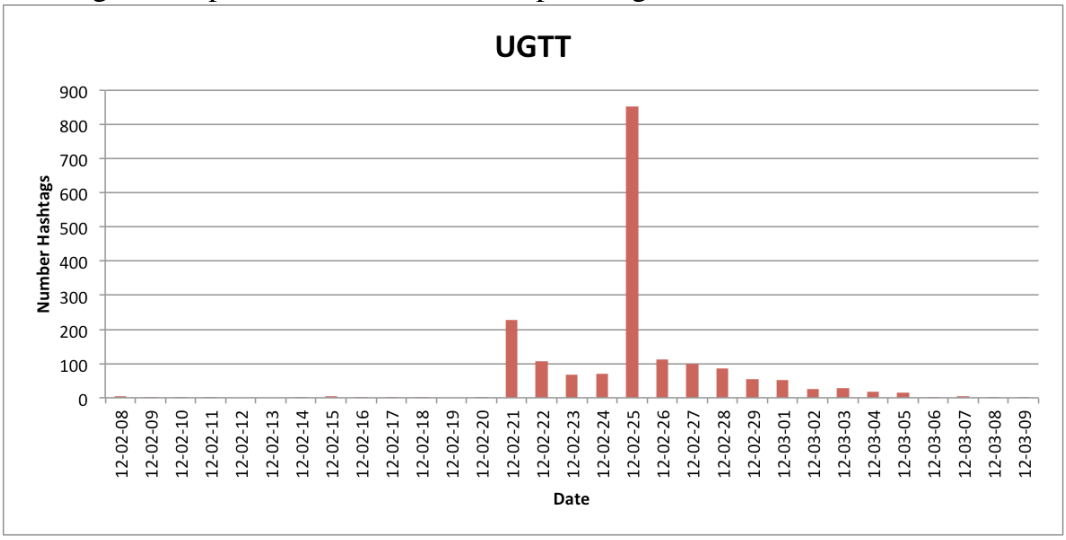


Figure 3.7 – Distribution par jour d'un hashtag (UGTT) lié au sujet de l'Union générale tunisienne du travail. 20 mars 2012 : Les employés municipaux ont effectué une grève qui est organisée par l'UGTT ; Le 25 février, des grandes manifestations organisées par UGTT réclament le départ de gouvernement.



Nous avons constaté qu'il existe également d'autres hashtags qui cooccurrent avec les hashtags principaux liés à ces événements. Ces hashtags (qui cooccurrent avec les hashtags principaux) reflètent la nature des événements et permettent d'éclaircir le contexte dans lequel ils ont eu lieu. De tels hashtags sont présentés dans les tableaux 3.VIII et 3.IX

.

Tableau 3.VIII – Fréquence des hashtags liés au sujet Wajdi Ghonim. Certains appellent le gouvernement, ennahdha et l'assemblée constituante à ne pas inviter les prédicateurs étrangers en Tunisie, d'autres les accusent d'avoir soutenu l'arrivée des prédicateurs étrangers ; le président Moncef Marzouki avait qualifié le prédicateur Wagdi Ghonim de "*microbe*" (cette déclaration avait suscité la polémique) ; Égypte est le pays d'origine de prédicateur ; Mourou est politicien qui a réagi au discours prêché de Wajdi Ghonim ; Une de ses fameuses prises de position se rapporte au sujet de l'excision des fillettes.

	Hashtags cooccurrent avec le sujet		**nombre de fois**
	Sujet	***Hashtags***	
Wajdi Ghonim	Tunisie	Tunisie	723
		Tunisia	
		tun	
		tuni	
		tunisi	
	Ennahdha	ennahdha	104
		nahdha	
		ennahda	
		ennhdha	
	Le gouvernement Tunisien	tngov	85
	L'assemblée constituante	tnac	82
	Moncef Marzouki	marzouki	56
		moncef_marzouki	
	Égypte	egypt	37
		egypte	
	Mourou	mourou	23
	Excision	excision	22
		exision	

Nous avons aussi effectué des statistiques sur les utilisateurs trouvés dans le corpus. Le tableau 3.XI présente les informations qui concernent les 5 utilisateurs qui ont le plus grands nombre de tweets dans le corpus. D'après leurs pseudonymes, nous avons remarqué que ces utilisateurs travaillent préalablement pour des journaux ou des magazines. Nous avons également remarqué que leur nombre d'abonnés est souvent beaucoup plus grand que leur nombre d'abonnements.

D'après le tableau 3.VII, les hashtags liés à l'emploi sont parmi les plus présents

Tableau 3.IX – Fréquence des hashtags liés au sujet Manouba. Certaines personnes demandent au gouvernement, ennahdha et l'assemble constituante d'intervenir et empêcher ce genre de comportement ; les salafistes ont organisé un sit-in à l'université pour revendiquer le droit des étudiantes de porter le niqab ; la mise en berne du drapeau tunisien a été faite par un étudiant salafiste.

	Hashtags cooccurrent avec le sujet		**nombre de fois**
	Sujet	***Hashtags***	
Manouba	Tunisie	Tunisie	702
		Tunisia	
		tunisi	
		tn	
	Le gouvernement Tunisien	tngov	126
	L'assemblée constituante	tnac	108
	Salafistes	salafistes	106
		salafiste	
		salafisme	
		salafis	
		salafist	
	Ennahdha	ennahdha	88
		nahdha	
		ennahda	
	Drapeau Tunisien	drapeau	11
		touchepasamondrapeau	
	Niqab	niqab	6

dans le corpus. Nous avons constaté que 96 % de tweets contenant un hashtag lié à l'emploi (emploi, job, recrutement, candidature) sont envoyés par l'utilisateur tunisieup. Tous les tweets sont des offres d'emploi.

D'après ces statistiques, nous avons déduit que les tweets, provenant de ces utilisateurs, sont généralement objectifs (contiennent des nouvelles). Cette constatation est confirmée par les descriptions, présentées dans le tableau 3.XII, relatives à ces utilisateurs. Même si ces utilisateurs écrivent souvent, nous avons remarqué qu'ils ne sont pas les plus retweetés. Cela peut être expliqué par le caractère objectif de leurs tweets. D'habitude, les utilisateurs retweetent des messages qui représentent leurs avis.

Le tableau 3.XIII illustre les 5 utilisateurs les plus retweetés.

Tableau 3.X – Fréquence des hashtags liés au sujet UGTT. Dans cette période il y avait des tiraillements entre UGTT et (ennahdha, tngov, tnac) ; Le sécrétaire général de l'UGTT (Jrad) est accusé de faire des obstacles au tngov par l'organisation des grèves et des manifestations ; Certains utilisateurs se souviennent que Jrad a soutenu le président déchu zaba

	Hashtags cooccurrent avec le sujet		**nombre de fois**
	Sujet	***Hashtags***	
UGTT	Tunisie	Tunisie	1571
		Tunisia	
		tuni	
		tun	
	Ennahdha	ennahdha	444
		nahdha	
		ennahda	
		enahdha	
		nahda	
	Le gouvernement Tunisien	tngov	196
	L'assemblée constituante	tnac	142
	Zine Abidine Ben Ali	zaba	48
	Manifestations	manif	33
		manifestation	
		maniftunis	
	Abdessalem Jrad	jrad	28

Tableau 3.XIII – Informations sur les tops 5 utilisateurs retweetés : NFER est le nombre de fois que l'utilisateur était retweeté ; NFAR est le nombre de fois que l'utilisateur a retweeté

Pseudonyme	**Nombre de tweets**	**Nombre d'abonnés**	**Nombre d'abonnement**	**NFER**	**NFAR**
ooouups	803	4 666	631	545	71
toomaa_6	22	2 630	1 063	489	1
nawaat	202	40 434	399	481	8
tn_revo	535	199	286	387	59
arabeman2012	1 081	308	219	324	362

Nous avons trouvé que les informations relatives à l'utilisateur toomaa_6 sont un

Tableau 3.XI – Informations sur les tops 5 utilisateurs : NFER est le nombre de fois que l'utilisateur était retweeté ; NFAR est le nombre de fois que l'utilisateur a retweeté

Pseudonyme	**Nombre de tweets**	**Nombre d'abonnés**	**Nombre d'abonnement**	**NFER**	**NFAR**
tunisinews	1 911	557	1	96	0
tunisieup	1 831	1266	631	16	0
tunisienouvelle	1 520	161	268	21	0
actutunisie	1 407	3818	18	76	0
journaltunisie	1 403	1497	255	88	1

Tableau 3.XII – Descriptions des 5 tops utilisateurs

Pseudonyme	**Description**
tunisinews	Latest News About #Tunisia & Arab World. Live 24/24 7/7
tunisieup	Ma vision sur Internet, tout semble à portée de clic. L'information n'a jamais été aussi abondante.
tunisienouvelle	Toutes les informations dans une nouvelle #tunisie
actutunisie	Actualité-Tunisie : Les dernières informations de l'actualité nationale et internationale ...
journaltunisie	Offre une compilation d'articles publiés par un grand nombre de sources d'actualités tunisiennes. Spécial Eléctions de la constituante 2011.

peu bizarre. Il a seulement 22 tweets, mais il est retweeté 489 fois. En fouillant dans les données, il s'est avéré que c'est un utilisateur bahreïnien. Il a utilisé souvent les hashtags `Tunisie`, `tngov`, `Tunisia`, `Arabspring` même si le contenu des tweets portent sur le Bahreïn. Un tel comportement peut être justifié par le déroulement de la révolution bahreïnienne dans cette période et il a utilisé ces hashtags pour faire référence à la révolution tunisienne.

À partir de ces analyses, nous avons déduit que les utilisateurs qui publient souvent, ne retweetent pas beaucoup et ne sont pas très retweetés sont généralement des utilisateurs "*objectifs*" (diffusent des informations). Cependant, les utilisateurs qui retweetent

souvent et qui sont très retweetés sont des utilisateurs "*subjectifs*" (publient des opinions plus que des informations)

Nous avons étudié le type de relation (symétrique, asymétrique ou il n'y a pas de relation) entre deux utilisateurs dans le cas d'un retweet. Nous avons utilisé un échantillon de 166 retweets. Le tableau 3.XIV montre les résultats obtenus. Nous avons constaté que la plupart des retweets sont effectués entre des utilisateurs qui ne sont pas des amis (relation symétrique).

Tableau 3.XIV – Type de relation entre les utilisateurs dans le cas d'un retweet

Type	**Pourcentage**
Asymétrique	36,14%
Symétrique	34,94%
Aucune relation	28,92%

3.3 Construction de corpus d'apprentissage

Jusqu'à présent, il n'y a aucune ressource linguistique annotée pour des messages écrits par des arabophones (tunisiens, marocains, égyptiens, syriens...). Rappelons que pays arabe a son propre dialecte. Dans ce travail, nous nous sommes intéressés actuellement aux tweets écrits par des Tunisiens.

Étant donnée l'absence de ressources, il est difficile d'appliquer des techniques de traitement de la langue naturelle afin de déterminer la langue et la polarité d'un tweet.

Pour cette raison, nous avons décidé de construire notre propre corpus d'apprentissage. Dans cette tâche, nous avons préparé un corpus qui contient plus que 80 000 tweets publiés entre le 08 février et le 09 mars 2012. Le corpus sera annoté par des experts initiés par le dialecte tunisien. Ces experts sont mes copains et mes collègues (des étudiants dans notre laboratoire et dans le département d'informatique de recherche opérationnelle). Leur tâche consiste à déterminer la polarité et la langue d'un échantillon de tweets tiré aléatoirement du corpus non annoté.

L'annotation sera effectuée à travers un site web [6] que nous avons créé pour réaliser cette tâche. La figure 3.8 montre une capture d'écran de la page qui permet aux experts

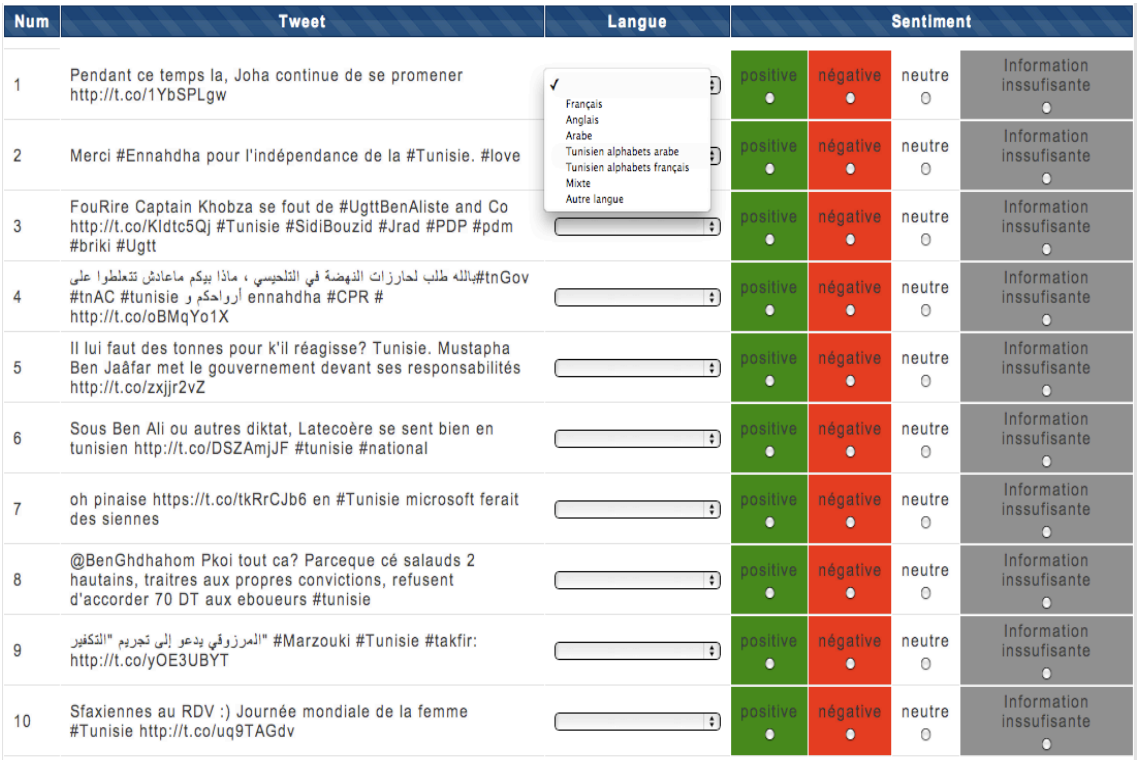
[6] `rali.iro.umontreal.ca:8080/dridihou`

d'annoter les tweets. Ce site contient également une présentation de notre projet et les résultats de nos expérimentations. Néanmoins, pour la crédibilité de notre corpus nous avons décidé que l'accès à la page dédiée à l'annotation soit par authentification (login et mot de passe). Les gens qui ne sont pas inscrits dans la liste des experts n'ont pas le droit d'annoter des tweets. Cependant, les autres personnes peuvent accéder à cette page en utilisant comme login = "*rali*" et mot de passe = "*diro*". Les annotations effectuées par ces personnes ne seront pas considérées.

Une fois notre corpus obtenu, nous prévoyons appliquer des méthodes classiques d'apprentissage machine (*réseaux bayésiens naïfs*, *machines à vecteurs de support*...) (voir 1.2) pour déterminer la langue et la polarité des nouveaux tweets.

Nous comptons rendre cette ressource disponible. Cette initiative pourra encourager pour la création d'autres ressources annotées pour différents dialectes arabes. Le printemps arabe, appelé aussi les révolutions des réseaux sociaux, et les événements après ces révolutions ont montré qu'il est indispensable d'analyser les données des réseaux sociaux publiés par les Arabes afin de déterminer l'opinion publique, détecter leur intérêt...

Figure 3.8 – Capture d'écran de la page dédiée pour l'annotation des tweets



3.4 Conclusion

Dans ce chapitre, nous avons présenté les méthodes que nous avons utilisées pour extraire les données à partir de Twitter et les expériences effectuées sur les données collectées. Nous avons constaté que la taille d'un tweet est généralement trop courte pour y appliquer des méthodes classiques de traitement de la langue. Les hashtags semblent être de bons points de départ pour détecter les événements et comprendre le contexte…

Jusqu'à ici, les analyses ont été faites manuellement. Dans les prochaines étapes, nous avons l'intention de développer des outils d'analyse automatique afin de comprendre le type de données. Nous présenterons dans le chapitre suivant les principales tâches que nous prévoyons effectuer.

CHAPITRE 4

CONTRIBUTION ET CONCLUSION

4.1 Contribution

Dans ce travail, nous nous intéressons plutôt aux événements sociaux et politiques. Notre objectif est de développer un système qui permet de détecter les préoccupations des utilisateurs (p. ex. `élections`, `l'événement de drapeau`, `l'arrivée de Wajdi Ghonim`...) et de prédire l'opinion publique (p. ex. `les partis préférés`, `l'attitude des utilisateurs de l'événement de drapeau`...).

Ce système devrait supporter les tweets écrits en arabe et en français. Nous utilisons le *microblog* Twitter comme source de données.

Étant donné les difficultés (taille réduite, ambiguïté, style d'écriture) d'analyser le contenu de tweet, nous prévoyons créer un modèle qui se base sur le contenu de tweets et d'autres éléments tels que les hashtags et les utilisateurs. Parmi les tâches que nous avons l'intention de faire :

- Éliminer les données qui peuvent fausser notre analyse. Prenons l'exemple de la section précédente, où les tweets envoyés par l'utilisateur `toomaa_6` sont retournés par la requête mais ils ne contiennent pas des informations sur la Tunisie.

- Analyser le comportement des utilisateurs : distinguer entre les utilisateurs objectifs et subjectifs, détecter les utilisateurs qui jouissent de la confiance des autres (dans cette tâche on peut prendre en considération le nombre de fois où un utilisateur est retweeté, le nombre d'abonnés...) etc.

- Normaliser les hashtags qui réfèrent au même sujet. Généralement, les utilisateurs n'utilisent pas le même hashtag pour un sujet particulier. Pour le sujet de `Wajdi ghonim` nous avons découvert plusieurs hashtags qui réfèrent à ce sujet : `ghanim`, `gonhim`, `ghoneim`, `wajdi_ghanim`, `wajdighenim`, `sheikghanim`...

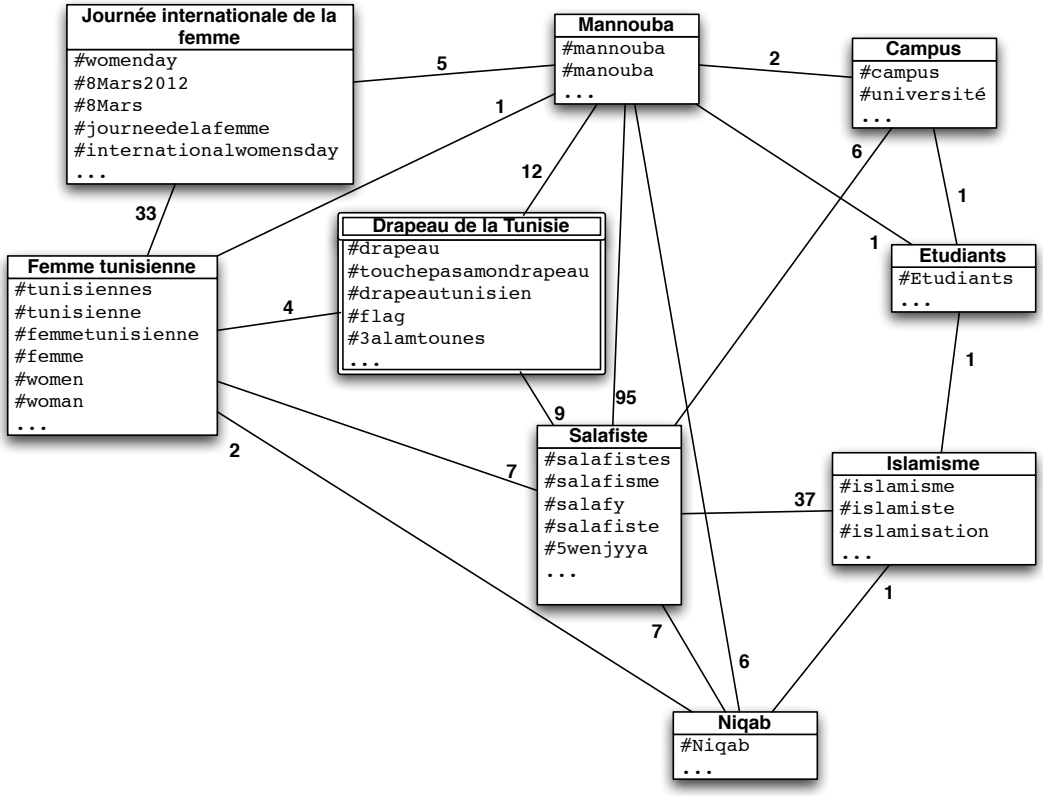
- Représenter les hashtags reliés par un graphe. La relation entre les hashtags permet de mieux comprendre le contexte d'un sujet. Reprenons l'événement de la mise en berne de `drapeau tunisien` dans le bâtiment de la *faculté des lettres, des arts et des humanités de Manouba* (section 3.2.3). Nous avons représenté les principaux hashtags liés à cet événement dans un graphe (figure 4.1). Les noeuds sont les *sujets* où chaque *sujet* contient les hashtags qui le représentent. Les arcs relient les *sujets* qui cooccurrent. Nous avons choisi le `drapeau` comme le sujet principal de ce graphe. La relation entre le sujet *Femme tunisienne* et l'événement de la mise en berne du drapeau peut être expliquée par l'étudiante qui a empêché l'étudiant d'enlever le drapeau. Le lendemain de l'événement est la journée internationale de la femme, plusieurs femmes ont fait un sit-in devant le local de l'assemblée constituante et elles ont dénoncé l'incident de la mise en berne du drapeau.

- Utiliser des techniques de traitement automatique de la langue naturelle (TALN) pour traiter le contenu textuel des tweets. Le traitement de ce type des tweets est plus compliqué que le traitement des tweets écrits avec une seule langue (anglais, français). Si un tweet est écrit seulement en anglais, nous pouvons utiliser des lexiques qui aident le traitement. Même si le tweet contient des fautes et/ou des abréviations, nous pouvons utiliser des normaliseurs permettant de rendre le texte dans une forme standard.

Même si nos données sont différentes de celles traitées par les travaux mentionnés dans le chapitre 2, nous avons l'intention de nous inspirer de ces travaux afin de traiter le contenu des tweets. La meilleure façon de déterminer l'opinion publique est d'utiliser une technique de classification. Cette tâche nécessite parfois un ensemble d'apprentissage (Go et al. [2009], Jiang et al. [2011], Barbosa et Feng [2010]). Afin de construire cet ensemble nous avons créé un site web qui permet aux initiés à la façon d'écriture des tunisiens d'annoter des tweets (voir section 3.3). Dans le même contexte, il est utile de tenir compte de différentes relations entre les données (tweets, utilisateurs…) pour déterminer les polarités des tweets (Jiang et al. [2011], Tan et al. [2011]).

Contrairement aux travaux de Doan et al. [2011] et Lampos et Cristianini [2010] qui utilisent des ensembles de mots-clés pour détecter des événements connus préalablement, dans cette thèse nous avons l'intention de détecter des événements (non connus préalablement) qui stimulent les utilisateurs dans une période.

Figure 4.1 – Graphe représentant des hashtags liés à l'évenement de la mise en berne de drapeau Tunisien à l'université de manouba le 07 mars 2012 par un étudiant "salafiste". Tous le noeuds sont en relation avec les hashtags liés à la `Tunisie`, `Le gouvernement` (`Tngov`), `Ennahdha`, `assemblée constituante` (`Tnac`); `3alamtounes` : `3alam` = `drapeau` et `tounes` = `Tunisie`. Le nombre n sur un arc indique que les deux sujets reliés par cet arc cooccurrent n fois.



4.2 Conclusion

Dans ce rapport, nous avons présenté : les principaux travaux qui s'intéressent à l'analyse des opinions pour les textes assez longs, les différences entre les blogs et les microblogs, la plateforme Twitter, les principaux travaux qui analysent les données de

Twitter. Dans le troisième chapitre, nous avons présenté les différentes techniques utilisées pour collecter les données, les expérimentations faites sur les données collectées. Les résultats ont montré que les hashtags peuvent jouer un bon rôle pour détecter des événements dans une période déterminée. Nos prochaines étapes consistent à détecter ces événements d'une façon automatique puis à produire des résumés qui décrivent le genre de l'événement et l'opinion publique (contre, pour...) sur les événements détectés. Pour résoudre cette tâche, nous prévoyons utiliser des techniques de TALN utilisées par d'autres travaux (chapitre 2), mais pour des tweets écrits d'une façon particulière.

Échéancier

Tableau 4.I – Étapes de réalisation

Automne 2010	j'ai réussi l'examen pré-doctoral (partie 1)
Hiver 2011	j'ai réussi l'examen pré-doctoral (partie 2) et j'ai validé le cours gestion de documents (IFT 6281)
Été 2011	j'ai commencé par me familiariser avec `Twitter API` et de découvrir l'état de l'art
28-29 septembre 2011	nous avons visité Halifax pour se rencontrer avec les responsables de *MediaBadger* et de l'Université de Dalhousie
Automne 2012	j'ai validé le dernier cours : intelligence artificielle (IFT 6010)
09-12 avril 2012	j'ai visité Halifax pour se rencontrer avec les responsables de *MediaBadger* et de l'*Université de Dalhousie*. Dans cette rencontre j'ai montré notre avancement et nos prochaines étapes
Mai 2012	nous avons présenté un papier (*Graduate student symposium*) dans la conférence *Canadian AI 2012*.

Tableau 4.II – Prochaines étapes

1	Construire des données d'apprentissage et appliquer des techniques d'apprentissage machine qui permettent de déterminer la langue et la polarité d'un tweet.
2	Développer une méthode qui permet de détecter des événements à partir des tweets et générer des résumés qui décrivent les événements.
3	Proposer un algorithme qui permet de regrouper les hashtags qui réfèrent au même sujet.
4	Proposer un algorithme qui permet de distinguer les différents types d'utilisateurs : journaux, neutres, qui essayent d'influencer sur les autres, qui jouissent la confiance des autres …

BIBLIOGRAPHIE

L. Barbosa et J. Feng. Robust sentiment detection on Twitter from biased and noisy data. Dans *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pages 36–44. Association for Computational Linguistics, 2010.

D. Chakrabarti et K. Punera. Event summarization using tweets. Dans *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 66–73, 2011.

K. Dave, S. Lawrence et D.M. Pennock. Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. Dans *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

S. Doan, B.K.H. Vo et N. Collier. An analysis of Twitter messages in the 2011 Tohoku earthquake. *Arxiv preprint arXiv :1109.1618*, 2011.

A. Go, R. Bhayani et L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

M. Hu et B. Liu. Mining and summarizing customer reviews. Dans *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

L. Jiang, M. Yu, M. Zhou, X. Liu et T. Zhao. Target-dependent Twitter sentiment classification. *Proc. 49th ACL : HLT*, 1:151–160, 2011.

H. Kwak, C. Lee, H. Park et S. Moon. What is Twitter, a social network or a news media ? Dans *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

V. Lampos et N. Cristianini. Tracking the flu pandemic by monitoring the social web. Dans *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE, 2010.

B. Liu. *Web data mining : exploring hyperlinks, contents, and usage data*. Springer Verlag, 2007.

B. Liu, M. Hu et J. Cheng. Opinion observer : Analyzing and comparing opinions on the web. Dans *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

D. Metzler, C. Cai et E. Hovy. Structured event retrieval over microblog archives. Dans *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 646–655, Montréal, Canada, June 2012. Association for Computational Linguistics.

B. OConnor, R. Balasubramanyan, B.R. Routledge et N.A. Smith. From tweets to polls : Linking text sentiment to public opinion time series. Dans *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

B. Pang, L. Lee et S. Vaithyanathan. Thumbs up ? : sentiment classification using machine learning techniques. Dans *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

A.M. Popescu et O. Etzioni. Extracting product features and opinions from reviews. Dans *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Association for Computational Linguistics, 2005.

C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou et P. Li. User-level sentiment analysis incorporating social networks. *Arxiv preprint arXiv :1109.6018*, 2011.

P.D. Turney. Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews. Dans *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

J. Wiebe, T. Wilson et M. Bell. Identifying collocations for recognizing opinions. Dans *Proceedings of the ACL-01 Workshop on Collocation : Computational Extraction, Analysis, and Exploitation*, pages 24–31, 2001.

H. Yu et V. Hatzivassiloglou. Towards answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences. Dans *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.