Institut National des Langues et Civilisations Orientales

Collecte et analyse exploratoire de données issues de Twitter sur le thème de la mobilité

Amélie Martin
Master 2 Traitement Automatique des Langues, spécialité ingénierie multilingue

Encadré par Coralie Reutenauer et Mathieu Valette
Stage effectué à la Direction de l'Innovation & de la Recherche de la SNCF

21 novembre 2014

Remerciements

Je remercie toutes les personnes qui ont contribué à la réalisation de ce mémoire, et en particulier mes encadrants Coralie Reutenauer et Mathieu Valette pour leur disponibilité et leurs éclairages.

Je remercie également le pôle Innovation et Partenariats de Transilien, ainsi que Sandrine Ségrétain et Françoise Dubois pour les données qu'ils m'ont fournies.

Je remercie les membres des différents groupes du département Mobilités & Services d'Innovation & Recherche (*Conception et Technologies pour le Voyageur*, *Statistiques, Econométrie et Data-mining*, et en particulier *Modélisation et Optimisation de la Décision*) pour l'aide précieuse qu'ils m'ont apportée tout au long de mon stage.

Enfin, merci à mes camarades de Master pour leurs connaissances, leur aide et leur soutien sans faille.

Table des matières

Introduction

Ce travail s'inscrit dans le cadre du programme « Connaissance de la mobilité » de la Direction de l'Innovation et de la Recherche de la SNCF, qui vise à explorer des types de données hétérogènes encore peu exploitées afin de mieux connaître la mobilité. Ce mémoire porte ainsi sur l'analyse de données textuelles.

L'étude des données textuelles issues du web est déjà mise en œuvre dans différentes branches de la SNCF, principalement sur un plan marketing : il s'agit par exemple d'étudier l'opinion des clients sur un produit, ou de mesurer la présence des différents produits/modes dans les conversations web et d'observer les évolutions (essentiellement quantitatives) dans le temps.

Néanmoins, ces analyses ne nous permettent pas d'observer la mobilité dans sa globalité, sur un volet plus qualitatif. Quelles informations sur les pratiques de mobilité peut-on faire émerger des données textuelles ? Quels types de renseignements sur leurs parcours partagent les usagers ? Comment adaptent-ils leurs habitudes en cas d'incident ? Comment sont exprimées les pratiques de fraude sur les réseaux sociaux numériques ? Si l'étude de la production textuelle des usagers, notamment au travers de leurs réponses à des questions ouvertes dans des enquêtes, est également déjà mise en œuvre à la SNCF, certains types d'informations précises sont encore difficiles à observer : nous cherchons donc à, d'une part, explorer les possibilités d'automatisation des traitements, et d'autre part, à dresser une typologie et montrer l'intérêt de données de nature différente encore peu exploitées dans l'entreprise bien qu'elles suscitent un intérêt croissant. Dans cette optique, nous avons choisi de nous intéresser à Twitter.

Ce projet à caractère exploratoire a donc pour but d'identifier et structurer un large panel d'informations pertinentes et originales pour la SNCF dans la production textuelle des usagers (issue de Twitter), aux moyens des techniques du TAL et de la textométrie, et d'amorcer une chaîne de traitement automatique.

Nous présentons dans ce mémoire le contexte dans lequel ce travail a été réalisé, en présentant la notion de mobilité et les différentes priorités de la

SNCF ; puis nous verrons des exemples de travaux sur l'analyse qualitative de messages issus des réseaux sociaux, ou alliant fouille de texte et mobilité ou territoire. Nous présentons ensuite la méthodologie de collecte de données et la nature du corpus obtenu, puis quelques procédés d'élimination du bruit et d'annotation. Nous rendons ensuite compte des différentes observations menées sur le corpus en rapport avec les pratiques de mobilité, avant d'introduire une typologie des tweets pertinents par rapport à notre problématique et premier système de classification automatique des tweets par rapport à cette typologie.

Chapitre 1

Présentation de la problématique et du contexte

1.1 La mobilité

La mobilité désigne la capacité à se déplacer ; on peut parler de mobilité professionnelle, de mobilité sociale, de mobilité réduite ou encore de mobilité spatiale. Cette dernière est a priori celle qui nous intéresse, bien que son cadre ne soit pas clairement défini : on peut en effet considérer la mobilité sociale comme une mobilité spatiale dans un espace non physique.

L'étude de la mobilité quotidienne a démarré dans la deuxième moitié du XXème siècle avec le développement de la société de consommation ou encore la généralisation de la voiture personnelle, s'est principalement développée à partir des années 80, et a connu un regain d'activités à la fin des années 90 avec l'émergence de théories comme celle de la mobilité généralisée, c'est-à-dire la mobilité comme moteur de mutations sociales (Gallez et Kaufmann, 2009). Cette mobilité est une notion pluridisciplinaire, abordée par des géographes (articulation au territoire), des sociologues et des économistes (domaine de la socio-économie des transports).

Nous la définissons schématiquement et dans notre contexte comme l'étude des déplacements et des modes de déplacement articulée à des critères sociologiques et économiques, dans une conception large qui combine espace, temps, et activité.

1.2 « Connaissance de la mobilité »

Le programme « Connaissance de la mobilité » est un programme Innovation & Recherche pluridisciplinaire. Il a pour but d'explorer sous un angle

nouveau des données sources de valeur, actuellement peu ou pas exploitées, afin de produire de nouveaux indicateurs sur les pratiques de mobilité. Ces indicateurs pourront contribuer à élaborer de nouveaux services clients, à mieux concevoir l'offre de transport, à améliorer l'exploitation et à objectiver les choix des autorités organisatrices de transport (AOT). Les investigations couplent des analyses quantitatives (données de capteurs, bases de données clients, flux vidéos...) à des analyses qualitatives.

Le programme se positionne à deux échelles : la gare et le territoire. A l'échelle du territoire, cinq POC (*Proof Of Concept*) pluridisciplinaires qui valorisent des types de données hétérogènes sont prévus.

Si les quatre premiers POC adoptent une démarche plutôt quantitative, le POC « Pratiques de Mobilité », dans le cadre duquel s'inscrit ce stage, a vocation à jongler entre le qualitatif (analyses linguistiques) et le quantitatif (indicateurs statistiques). Il a pour but l'étude des données textuelles produites par les usagers, qu'elles soient issues du web comme les blogs, forums, réseaux sociaux numériques, etc., ou d'autres canaux tels que les enquêtes, par exemple.

1.3 Problématiques des branches SNCF

Quatre problématiques prioritaires pour les branches SNCF ont été identifiées afin de définir les orientations du POC :

- les **stratégies adaptatives** : comment les usagers réagissent et réorganisent leur parcours en cas de perturbation ;
- la **multimodalité**/l'**intermodalité** : comment les usagers articulent les différents modes qui constituent leurs parcours, quels sont ces modes ;
- les **motifs de déplacement** : pourquoi les usagers se déplacent, où vont-ils et quels sont les motifs sous-jacents à leurs déplacements ;
- la **fraude** : comment sont décrites les pratiques de fraude sur le web, comment s'organise la fraude sur le web.

1.4 Quelles données ?

1.4.1 Contraintes et restriction du périmètre

La collecte de nos propres données était soumise à deux contraintes principales : la facilité de collecte, et la pertinence pour l'entreprise.

La première contrainte concerne le plan technique, mais aussi le plan juridique : la collecte et le traitement de données issues du web social, même

publiques, engendrent automatiquement une déclaration au correspondant informatique et libertés (CIL, intermédiaire de la CNIL au sein de la SNCF) qui doit être validée.

D'autre part, il nous a semblé lors des deux séminaires sus-cités, que les données web suscitaient un intérêt fort de la part des différents acteurs SNCF : nous avons donc mis de côté les enquêtes et les réclamations clients.

1.4.2 Présentation de Twitter

Twitter est un site de microblogging créé en 2006. Il permet de publier de courts messages de 140 caractères maximum, datés, et qui peuvent être géolocalisés si l'utilisateur le souhaite. Les tweets ne peuvent être postés anonymement et sont toujours liés à un compte, dont l'identifiant unique commence par le signe @. L'utilisateur peut paramétrer son compte comme étant public (tous les tweets de sa timeline personnelle seront visibles, y compris pour les personnes ne détenant pas de compte Twitter), ou privé (les tweets de sa timeline ne seront visibles que par ses abonnés ou *followers*). Les informations personnelles divulguées par les utilisateurs de Twitter sont beaucoup moins précises que celles diffusées par les utilisateurs de Facebook : (Cheng *et al.*, 2010) mesurent que seulement 26% des utilisateurs (dans un échantillon de 1 million de tweets) révèlent le nom de leur ville de résidence (cependant, les pratiques évoluent rapidement). Le service lui-même pousse beaucoup moins ses utilisateurs à révéler leur identité (un @ et un email suffisent pour s'inscrire).

En avril 2013, environ 5% de la population française détenaient un compte Twitter actif (2.3 millions de personnes environ) ; parmi eux, 33% tweeteraient au moins une fois tous les deux jours (800 000 personnes environ). 61% de ces utilisateurs ont moins de 35 ans, et 33% habitent en Ile-de-france (67 % dans des milieux urbains de plus de 100 000 habitants) [1]. Il est important de ne pas envisager le Twitter francophone comme un tout homogène mais comme un ensemble de sphères souvent imperméables ou se chevauchant partiellement, reflétant la très grande diversité des utilisateurs de la plateforme : journalistes/communicants/professionnels, adolescents et jeunes adultes, communautés de fans/passionnés (du boysband One Direction, par exemple), utilisateurs qui détournent le medium (rôlistes...), comptes automatiques ou semi-automatiques qui relaient de l'information (@raildar_info, qui donne des infos sur le trafic SNCF en temps réel, @quoimaligne ou @controleurRATP qui retweetent en masse des infos lancés par

1. Source : enquête IPSOS/CGI *Usages et pratiques de Twitter en France*, consultable à l'adresse http://www.ipsos.fr/ipsos-public-affairs/actualites/2013-04-25-usages-et-pratiques-twitter-en-France

des usagers)... Ces communautés partagent pourtant bel et bien le même mode et le même espace d'expression (usage des *hashtags*, des *retweets* et des *mentions* notamment), mais ont des pratiques textuelles différentes.

Un tweet est généralement étudié comme un simple texte, mais il est tributaire de son écosystème : il est accompagné de métadonnées et évolue dans des timelines différentes qui influencent son sens (Paveau, 2013). De plus en plus, les citations de tweets en ligne ne se font plus avec un simple copier/coller du texte, mais avec une capture d'écran de sa matérialisation dans l'interface de Twitter, ou à l'aide d'un cadre intégré qui reprend la date, le nom d'utilisateur etc [2]... Nous choisissons cependant dans ce travail de considérer le tweet uniquement pour son contenu textuel, même si nous aborderons brièvement son *investigabilité* (*searchability* selon boyd cité par (Paveau, 2013)) dans la sous-section 4.4 et le lien du texte à ses métadonnées dans les perspectives (section 6).

Nous avons choisi de nous concentrer sur Twitter, pour plusieurs raisons : contrairement aux blogs, aux réclamations clients, ou aux enquêtes, c'est un canal qui est majoritairement non sollicité par SNCF. Il permet de capter des réactions à chaud, en temps réel ou presque, par rapport à une situation perturbée ou à une planification de parcours, par exemple. D'autre part, les outils existants pour la collecte et le stockage, combinés à la documentation très complète de l'API satisfont les contraintes techniques. Enfin, une déclaration au CIL a été effectuée et validée.

Le problème de la représentativité des données textuelles disponibles sur le web avait été abordé par plusieurs acteurs au sein de l'entreprise : jeunesse des utilisateurs, proéminence de l'effet de buzz, *trolls*, inégalités face au numérique... Ce sont des éléments que nous ne laissons pas de côté. Néanmoins, ces données complètent d'autres indicateurs, comme les enquêtes (échantillon représentatif) : l'analyse de tweets a vocation à être croisée à d'autres données pour obtenir une vision globale de la mobilité.

1.5 Objectifs

L'objectif de ce travail est d'ouvrir des pistes intéressantes quant aux pratiques de mobilité au sein des tweets et en accord avec les orientations données par les différents acteurs de la SNCF, et d'amorcer une chaîne de traitement automatique. Le but est de montrer le potentiel des données web non structurées pour la découverte de pratiques difficiles à détecter dans les enquêtes, parce qu'elles sont minoritaires, communautaires ou simplement

2. Un exemple d'outil d'aide à la citation de tweets : http://tweet2cite.com/

ponctuelles, ou encore pour la validation ou l'invalidation par le qualitatif de données quantitatives plus « traditionnelles ».

Ce travail s'est découpé en cinq grandes étapes :

— Une étape préalable de repérage des sources de données pertinentes sur le web ;
— La mise en place d'une collecte de données pérenne et suffisamment exhaustive pour nous permettre d'explorer différentes problématiques et de nous adapter aux besoins des branches qui pourraient émerger ;
— Une phase de débruitage et de pré-annotation d'éléments intéressants (noms de gare, informations temporelles, etc.) ;
— Une phase d'exploration avec l'outil TXM afin de mettre au jour des informations sur la manière dont les utilisateurs-usagers expriment leur mobilité sur Twitter (hashtags courants, thématiques saillantes, etc.) ;
— Une phase de formalisation : nous avons établi une première typologie de tweets relevant de la mobilité et nous avons tenté de construire un modèle d'apprentissage automatique afin de classifier les tweets collectés selon cette typologie.

Chapitre 2

Contexte théorique

2.1 Etude qualitative outillée des messages issus des réseaux sociaux numériques

Outre son importance dans l'analyse de sentiments, l'étude des messages postés sur les grands réseaux sociaux numériques via des méthodes textométriques se développe notamment dans le domaine des humanités numériques. Elle sert de support à des analyses relevant des sciences humaines (sociologiques, psychologiques, linguistiques, historiques, etc.). Nous nous sommes référés à quatre travaux en particulier.

(Ducos *et al.*, 2014) présentent une analyse qualitative des messages publiés sur la page Facebook de soutien au bijoutier de Nice (évènement survenu en 2013). Afin d'identifier les principales thématiques appelées par les différents groupes d'individus se retrouvant sur cette page, les auteurs opèrent une correction orthographique sur les messages, puis une classification hiérarchique descendante qui permet de diviser le corpus en classes. En observant le lexique de chaque classe, les auteurs repèrent au moins neuf thèmes sur lesquels les groupes d'individus s'affrontent (soutien au bijoutier, critique de son geste, réinsertion des délinquants et politique en matière de justice...).

(Smyrnaios et Ratinaud, 2014) opèrent également une classification hiérarchique descendante sur un échantillon de tweets qui traite du pacte budgétaire européen afin de qualifier des communautés d'utilisateurs identifiées grâce à une analyse de graphes basée sur les retweets. Les auteurs trouvent une cohérence entre les classes lexicales qui ont émergé de l'analyse hiérarchique descendante et les communautés qui les emploient. (Ratinaud, 2014) propose une méthode similaire sur le thème du mariage pour tous. Ces deux travaux choisissent volontairement de collecter des tweets sur la base de mots-clés non ambigus (*hashtags* TSCG pour Traité sur la Stabilité, la Coordina-

tion et la Gouvernance, ou mariagepourtous) afin de minimiser le bruit au sein du corpus

D'autres travaux sur Twitter choisissent de se focaliser sur les *hashtags*. (Cervulle et Pailler, 2014) ont également abordé la question du mariage pour tous et de sa résonance sur Twitter en étudiant la « dimension affective » des hashtags utilisés par les différents groupes polémiques.

2.2 Fouille de texte et territoire

Les domaines de la mobilité, du territoire ou du transport sont caractérisés par des analyses plutôt quantitatives : modélisation de flux à partir de relevés terrain ou de résultats d'enquêtes, prédiction de l'affluence sur certains réseaux (via, par exemple, le challenge Data Science qui appelait à utiliser les données ouvertes Transilien pour établir le meilleur modèle prédictif possible). Néanmoins, comme indiqué précédemment, l'intérêt pour les méthodes basées sur l'analyse de contenus couplée à des données quantitatives grandit.

(Kergosien *et al.*, 2014) présentent la méthode OPILAND de fouille de textes issus du web ou encore de la presse afin de qualifier la perception qu'ont les individus de l'aménagement d'un territoire donné. Les auteurs utilisent des méthodes d'extraction d'informations spatiales couplées à des méthodes d'analyse d'opinions sur la base de lexiques. Les auteurs constatent que les termes issus des lexiques d'opinion donnent de faibles résultats sur les textes du domaine de l'aménagement du territoire, et sont amenés à faire de l'extraction de mots polarisés relatifs à ce domaine.

(Cheng *et al.*, 2010) proposent une méthode basée sur le contenu textuel des tweets pour géolocaliser les utilisateurs de Twitter afin de palier le manque de précision des informations fournies par ces derniers. Leur système, testé sur l'anglais et sur le territoire nord-américain, permet de localiser 51% des utilisateurs à moins de 100 miles de leur véritable lieu de résidence.

Enfin, (Morency *et al.*, 2013) ont réalisé un rapport sur les potentialités du web comme outil de collecte de données sur la mobilité, qui présente les avantages et les inconvénients de la collecte de données d'enquêtes via des interfaces web (plutôt que par téléphone). Ce mode de collecte de données hétérogènes, avec notamment la possibilité d'intégrer de la géolocalisation, pose la question de la valorisation de ces données : comment croiser efficacement les données quantitatives et structurées, avec des données non structurées telles que les réponses à des questions ouvertes, dans le contexte de la mobilité ?

Chapitre 3

Collecte et corpus obtenus

3.1 Méthodologie de collecte de tweets

3.1.1 Choix de l'API

Twitter propose deux types d'API publiques et gratuites pour la collecte : REST et Streaming. Au sein de REST, l'API de collecte est l'API Search, qui a un comportement similaire (mais pas tout à fait identique) à la fonction search de twitter.com [1]. La différence majeure entre Search et Streaming réside dans leurs orientations respectives : Search se focalise sur la pertinence des résultats, et Streaming sur l'exhaustivité (c'est-à-dire, faire remonter le maximum de tweets qui correspondent à la requête). Tout comme le moteur de recherche de twitter.com, les résultats d'une requête récupérés sur l'API Search sont non-exhaustifs, et peuvent être classés du plus récent au plus ancien ou selon la popularité. L'API Streaming a un fonctionnement différent (elle demande une connexion HTTP persistante) qui permet de capter les tweets quasiment au fur et à mesure de leur publication. Nous avons donc choisi Streaming pour sa capacité à capitaliser les tweets en continu.

Il existe aussi des méthodes payantes plus exhaustives et surtout non-soumises à la limite de débit [2] imposée aux deux API gratuites, comme Gnip, qui permet de se connecter au Firehose (le flux complet de tweets en temps réel), et également de remonter dans l'historique des tweets publiés. (Morstatter *et al.*, 2013) ont réalisé une étude sur la couverture de Streaming par rapport au Firehose, qui révèle que celle-ci est très variable en fonction des paramètres passés à l'API et du but de la collecte. Il en ressort qu'il est toujours difficile d'appréhender l'exhaustivité et la représentativité d'un dataset

1. https://twitter.com/search-home
2. Avec Streaming, la collecte ne doit pas dépasser 1% du *firehose* à l'instant T.

issu de l'API Streaming.

Il n'est pas autorisé par Twitter de partager un dataset (sur le net par exemple), sauf si l'intégralité des données est remplacé par des ID (du tweet et de l'utilisateur) [3].

3.1.2 Choix de l'outil

De nombreux modules, bibliothèques ou packages ont été développés pour accéder aux API de Twitter. Nous pouvons citer le package R *twitteR*, les packages Python *twitter*, *twython* et *tweepy*, et la bibliothèque Java *twitter4j*, qui permettent tous de gérer la connexion à l'API dans le langage choisi. Il existe également le framework PHP/Javascript *140dev* qui fournit une interface simplifiée pour l'API, et divers outils en ligne comme *TAGs*, qui fonctionne avec Google Spreadsheets (API Search) et présente l'avantage de ne pas nécessiter de connaissances en programmation pour archiver des tweets. Dans la même optique, le site *twapperkeeper.com* permettait également de collecter et stocker des tweets, mais a été fermé suite au changement de politique de Twitter ; le logiciel open source *yourTwapperKeeper*, utilisé par (Smyrnaios et Ratinaud, 2014), vient palier cette fermeture : c'est une version ouverte de twapperkeeper à héberger soi-même. (Smyrnaios et Ratinaud, 2014) évoquent également l'outil *DMI-TCAT* développé par le groupe de recherche Digital Methods Initiative (Borra et Rieder, 2014). Cet outil est également conçu pour être hébergé sur un serveur maison.

Nous avons choisi cet outil pour sa gestion des différentes requêtes via une interface graphique en PHP, et ses outils d'analyse fournis et accessibles via l'inferface graphique : il permet de visualiser la volumétrie des différents corpus, de réaliser différents exports (hashtags et fréquences, échantillons basés sur la date ou sur la présence d'un mot-clé, échantillon géolocalisé...) ou encore d'obtenir des graphes en .gdf visualisables dans Gephi. Son caractère open source permet également de réutiliser et modifier certains de ses composants.

3.1.3 Définition des mots-clés de collecte

La collecte via l'API se fait uniquement à partir de mots-clés ou d'expressions dont la forme est fixe. Par exemple, lancer une requête avec le mot-clé *métro* ne captera pas les formes *metro* ou *métros*.

Notre choix de mots-clés ou d'expressions de collecte n'a pas porté uniquement sur des noms de réseau, des noms de lignes ou des offres SNCF. Afin

3. Developer policy : https://dev.twitter.com/overview/terms/policy

de capitaliser un historique assez important qui couvre tous les aspects de la mobilité, nous avons capté un grand nombre de modes, lignes, produits, ainsi que des mots caractéristiques du vocabulaire de la mobilité.
Les mots entourés d'apostrophes sont des expressions figées qui nous per-

```
Corail, Covoiturage, IDBus, IDTGV, Intercites, Métro, Ouigo,
RER, TGV, Tram, Tramway, Transilien, Velib, Velo, autobus,
autopartage, autostop, avion, aérogare, aéroport, chemin,
circuler, correspondance, desserte, desservir, ferroviaire,
gare, garer, itineraire, ligneh, lignej, lignek, lignel,
lignen, lignep, ligner, ligneu, marche, moto, navette,
parcours, quai, rail, rame, ratp, rera, rerb, rerc, rerd,
rere, retard, roller, route, scooter, skate, sncf, station,
stationnement, stationner, taxi, trafic, train, trajet,
transports, trottinette, vehicule, voyager, voyageur,
voyageurs, vtc
```

FIGURE 3.1 – Liste de mots-clés thématiques de collecte pour le corpus de tweets Mobilité

```
'en bus', 'en ter', 'le bus', 'le ter', 'mon bus', 'mon
ter', 'un ter', 'à pieds', de bus
```

FIGURE 3.2 – Liste des expressions de collecte pour le corpus de tweets Mobilité (mots trop communs pour être un mot-clé de collecte)

mettent de restreindre la collecte : en effet, collecter *bus* seul (mot similaire dans toutes les langues) ou *TER* seul (verbe *avoir* en portugais) ramenait trop de bruit. *de bus* sans apostrophe nous permet de collecter les tweets qui contiennent les mots *de* et *bus*, pas forcément accolés l'un à l'autre (*de* étant un bon indice sur la langue du tweet).

Cette méthodologie de collecte présente l'inconvénient de ramener une masse conséquente de messages non pertinents, notamment du fait de l'ambiguité de certains mots (*train, marche, station, parcours...*), mais également à cause des paramètres de l'API : en effet, *RER* ramène des tweets contenant les verbes *préférer*, *libérer*, *gérer*... du fait de la gestion des caractères diacrités par l'API de Twitter.

Nous avons également constitué un corpus annexe concernant la fraude.

```
controleur, controleurs, fraude sncf, frauder sncf, frauder
train, frode, froder
```

FIGURE 3.3 – Liste des mots-clés de collecte pour le corpus de tweets Fraude

3.2 Présentation du corpus de tweets « Mobilité »

3.2.1 Volumétrie

DMI-TCAT collecte un très gros volume de tweets dans toutes les langues : environ un tiers des données est en français. Sur le mois de septembre, notre requête a collecté 3 230 697 tweets en français (figure 3.4), et au 14 octobre, la base de données comptaient 8 378 253 tweets en français pour une collecte partiellement lancée le 9 juillet.

Période	Nombre de tweets en français
1er au 30 septembre 2014	3 230 697
9 juillet au 14 octobre 2014	8 378 253

TABLE 3.1 – Volumétrie du corpus de tweets Mobilité

3.2.2 Caractéristiques générales

Une large partie des tweets du *firehose* est consacrée au partage de liens externes (encouragé par l'apparition d'un *Tweet Button* sur un grand nombre de sites). La proportion de liens au sein de notre corpus est variable mais n'excède jamais 25% des tweets.

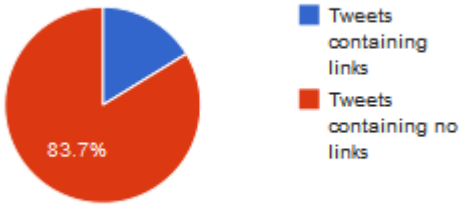

FIGURE 3.5 – Proportion de tweets du corpus Mobilité collectés en septembre contenant un lien (tiré de DMI-TCAT)

D'autre part, du fait de l'ambiguité et du nombre de nos mots-clés de collecte, ce corpus est extrêmement bruité. Les hashtags les plus fréquents dans le corpus au mois de septembre sont un bon indicateur de la proportion de bruit au sein du corpus (tableau 3.2). Par exemple, le hashtag le plus utilisé se révèle être *SS8* (pour l'émission *Secret Story 8*).

Le corpus semble également contenir plus de tweets géotaggés que sur les échantillons cités dans (Cheng *et al.*, 2010) et (Morstatter *et al.*, 2013) (1% et moins pour leurs échantillons contre 3% à 3,5% environ pour le nôtre, voir figure 3.4).

Enfin, en septembre, on relève une surprenante régularité de la proportion de tweets collectés selon les périodes de la journée : on observe un pic systématique à 7h du matin, une légère hausse vers midi et un dernier pic entre 18h et 22h (figure 3.6).

En revanche, les jours de week-end et d'août ne suivent pas le même schéma (figure 3.7 et 3.8), ce qui peut indiquer une certaine dépendance de la courbe aux navettes pendulaires ou au rythme scolaire.

3.2.3 Nettoyages, reformatages

Afin de réduire le bruit au sein du corpus, des pré-traitements ont été effectués : les tweets identiques et les retweets sont supprimés, ainsi qu'un certain nombre d'ambiguités connues ((*en train de*, *ça marche*...) et les tweets contenant les verbes se terminant par `[caractère diacrité+rer]` (erreurs de l'API). Il a également été décidé, après examen des données, de supprimer les tweets contenant des liens : en effet, leur contenu textuel est en majorité en référence à ce lien et ne présente pas beaucoup d'intérêt pour ce travail. DMI-TCAT permet cependant de retrouver le véritable URL des liens raccourcis (Twitter force le raccourcissement des liens via le service t.co), ce qui

permet dans d'autres contextes d'étudier la nature de ce que partagent les utilisateurs.

A l'issue de ces étapes, environ 45% de tweets sont éliminés. Nous choisissons de n'opérer aucune correction orthographique sur le corpus.

3.3 Corpus annexe de tweets « Fraude »

Corpus	Nombre de tweets
Fraude	34 971

TABLE 3.3 – Volumétrie du corpus de tweets Fraude

Nous n'avons pas étudié ce corpus en profondeur : il n'a pas été nettoyé et envoyé dans TXM. Vu son petit volume, une analyse manuelle suffisait : nous l'avons utilisé pour établir une typologie des tweets concernant la fraude (section 5.1).

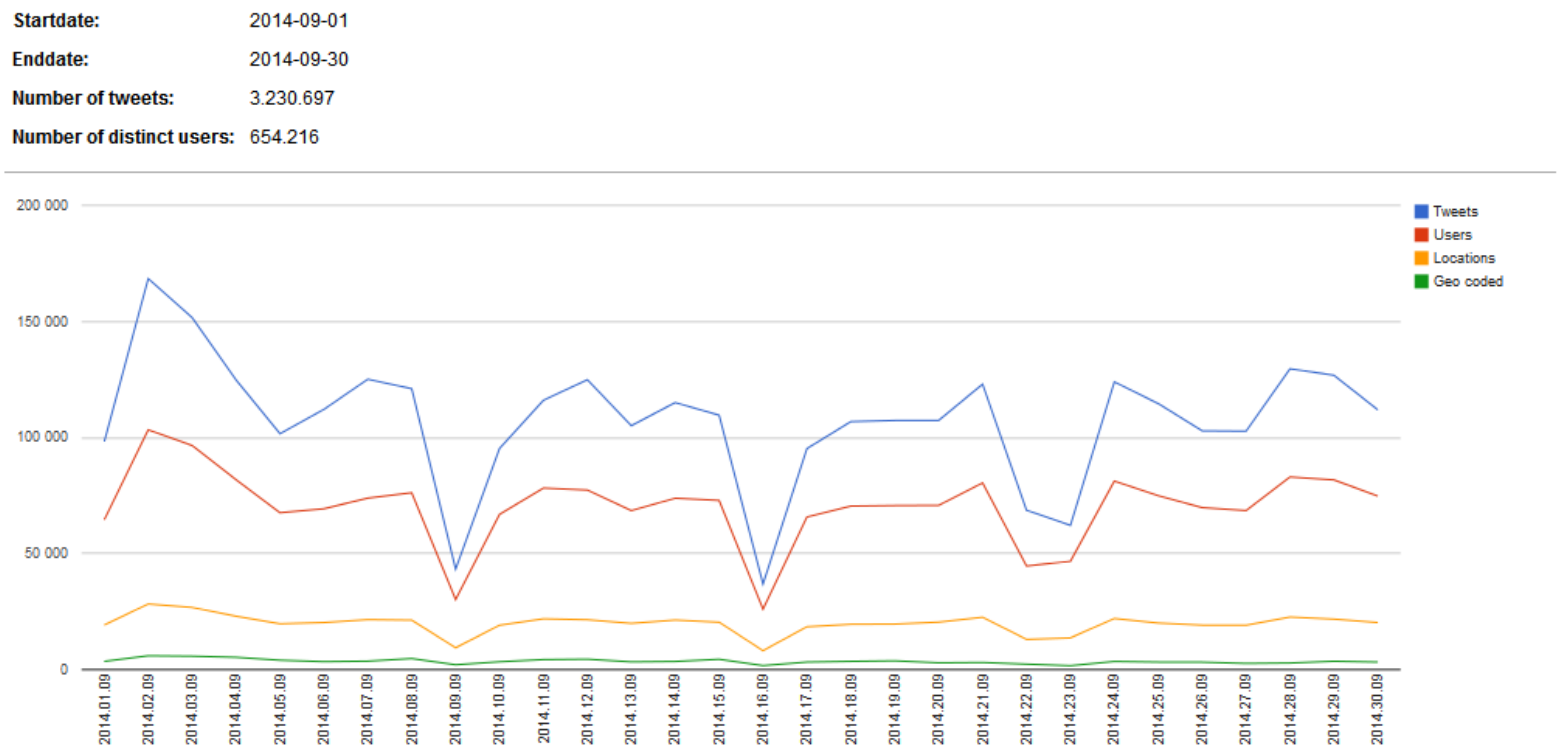

FIGURE 3.4 – Courbe du volume de tweets collectés sur le mois de septembre 2014, tirée de DMI-TCAT (les baisses significatives correspondent aux journées de maintenance)

hashtag	fréquence
SS8	16392
monpremierTweet	15636
SNCF	10984
moto	10786
Moto	8051
RT	6072
velo	4600
Paris	4349
LT	4156
sncf	3917
ratp	3830
RATP	3670
AFP	3514
qml	3487
Auto	3474
gameinsight	3340
Capricorne	2851
5oct	2693
Hollande	2652
lt	2587
skate	2585
ASFC_PBR	2567
LRT	2441
RisingStar	2402
Sarkozy	2278
KohLanta	2240
auto	2235
covoiturage	2192
TheVoiceKids	2170
FinaleSS8	2086
NjoyHit40	2076

TABLE 3.2 – Hashtags les plus fréquents au sein des tweets du corpus Mobilité collectés en septembre

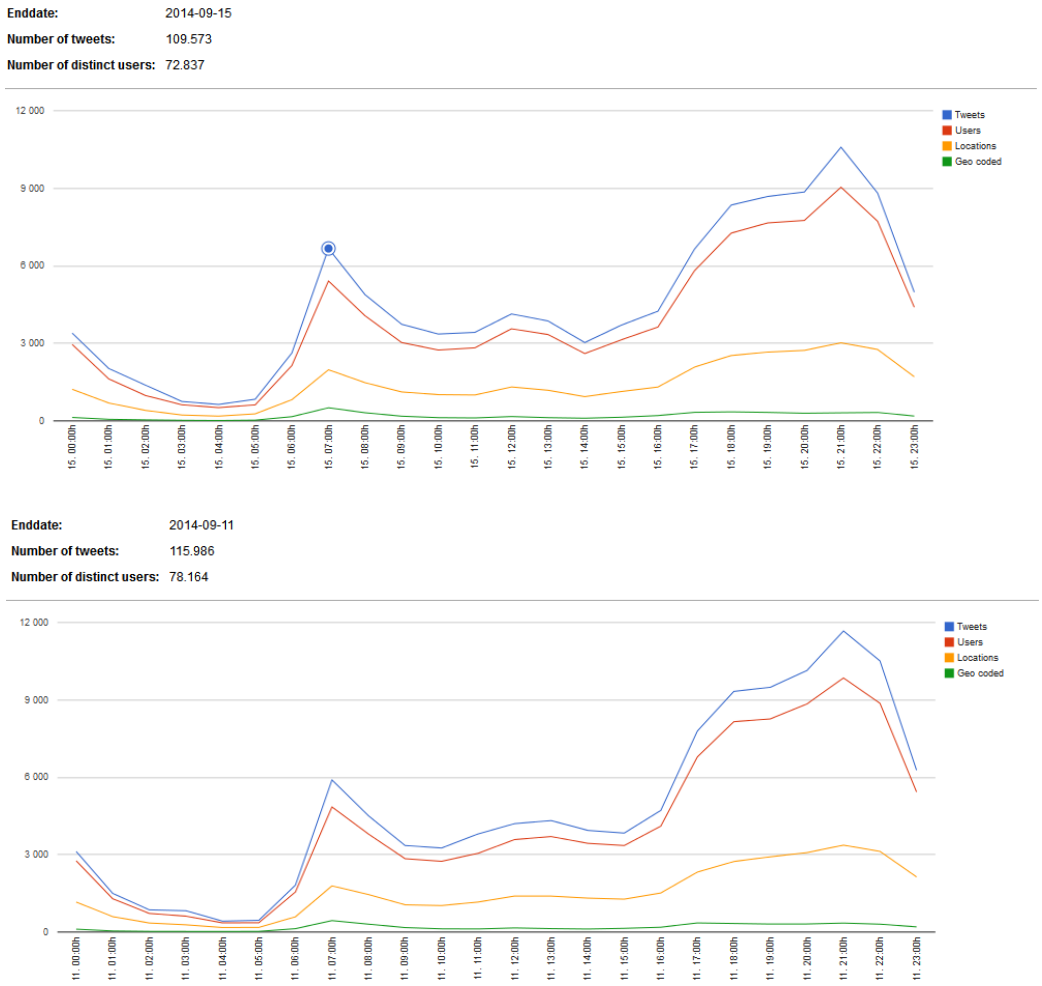

FIGURE 3.6 – Régularité de la courbe de capture de tweets les jours de semaine, heure par heure (lundi 15 septembre, jeudi 11 septembre, tirés de DMI-TCAT)

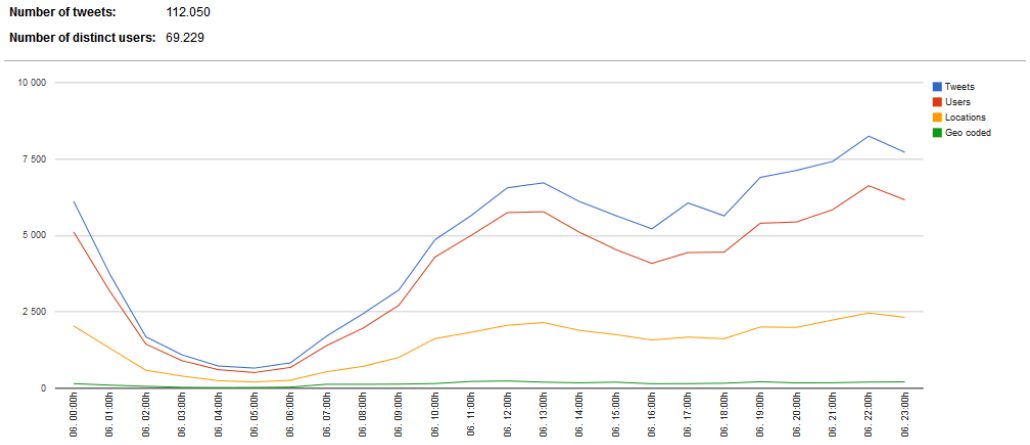

FIGURE 3.7 – Courbe de capture pour le samedi 6 septembre, heure par heure (tiré de DMI-TCAT)

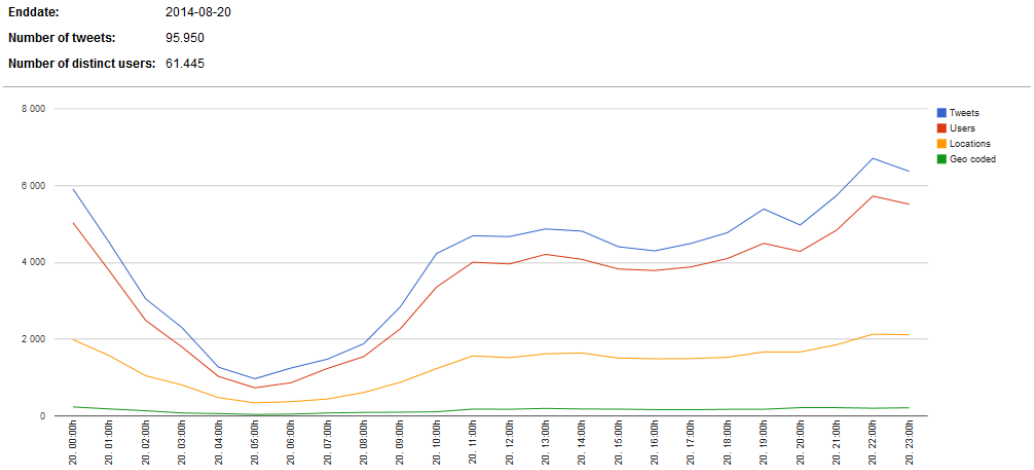

FIGURE 3.8 – Courbe de capture pour le mercredi 20 août, heure par heure (tiré de DMI-TCAT)

Chapitre 4

Annotation et observations

4.1 Classification automatique des tweets pertinents

4.1.1 Annotation manuelle

Malgré les prétraitements, une large part de bruit subsistait au sein du corpus. Nous avons donc entrepris d'annoter manuellement les tweets selon leur pertinence par rapport à notre sujet (1 ou 0). Nous avons ainsi annoté près de 2000 tweets. Selon la période étudiée, le pourcentage de tweets pertinents varie. On parvient, en septembre, à 68.8% de tweets non pertinents, contre presque 80% en juillet (pour les mêmes mots-clés de collecte). Au sein d'une même journée, les variations sont également notables. Nous avons prélevé 300 tweets à 5 périodes différentes de la journée du 2 septembre et les avons classifiés en *pertinents* et *non pertinents*.

Heure	Tweets pertinents	Tweets non-pertinents	Pourcentage de tweets non-pertinents
1h25	66	244	81%
7h25	140	160	53%
12h25	100	200	67%
18h25	92	208	69%
20h25	80	220	73%
Total	468	1032	68.8%

TABLE 4.1 – Nombre de tweets pertinents et non-pertinents au sein de cinq échantillons prélevés le 2 septembre 2014 (le décompte est réalisé sur un corpus déjà dédoublonné et nettoyé).

4.1.2 Classifieur bayésien naïf pour la détection de tweets pertinents

A l'aide du corpus de tweets pertinents/non pertinents, nous avons entraîné un classifieur bayésien naïf basé sur les n-grammes de caractères (séries de n caractères). Afin de mettre en place le meilleur système possible, nous avons réalisé une phase d'évaluation. Nous avons testé quelles combinaisons de n-grammes de caractères donnaient les meilleurs résultats (corpus d'entraînement : 1600 tweets). Après quelques tests, nous avons décidé d'évaluer les quadrigrammes seuls, les trigrammes seuls, et la combinaison bigrammes / trigrammes / quadrigrammes :

Combinaison	4grammes	3grammes	2-3-4grammes
Rappel	0,76	0,65	0,70
Précision	0,72	0,73	0,72
F-mesure	0,74	0,69	0,71

TABLE 4.2 – Evaluation du classifieur pertinent / non pertinent

En augmentant encore un peu plus le corpus d'entraînement (1800 tweets), le classifieur basé sur les quadrigrammes parvient à une f-mesure de 0.75 (contre 0.72 pour les trigrammes). Nous nous sommes donc arrêtés sur cette configuration. Cette méthode un peu « brutale » permet néanmoins d'éliminer de manière automatique entre 50% et 60% de tweets non pertinents (sur un corpus dédoublonné et nettoyé).

4.2 Pré-annotation pour une exploitation dans TXM

4.2.1 Présentation de TXM

Les explorations textométriques menées lors de ce stage ont été réalisées à l'aide du logiciel open source TXM (Heiden *et al.*, 2010), qui réunit plusieurs fonctionnalités de logiciels de textométrie déjà existants et permet de traiter des corpus Unicode au format texte, ou formalisés en XML-TEI, XML-Factiva (du nom de l'agrégateur de contenu Factiva), ou encore dans des formats exportés d'autres logiciels de textométrie français, comme Alceste ou Hyperbase. Il permet de partitionner les corpus sur différents critères, ou encore de créer des sous-corpus qui pourront ensuite être comparés au corpus dans son emsemble.

TXM se présente comme une boîte à outils Java avec un environnement R enrichi d'un package maison (TextometrieR) pour les calculs statistiques. Il utilise l'étiqueteur morphosyntaxique TreeTagger pour lemmatiser et étiqueter les corpus en parties du discours.

TXM permet d'effectuer des calculs sur des « propriétés », c'est-à-dire des formes, des lemmes, des étiquettes morphosyntaxiques ou encore des propriétés personnalisées encodées en XML.

4.2.2 Insertion de balises

Afin de faciliter les explorations dans le logiciel TXM, des échantillons plus réduits ont été prélevé dans la base de données (uniquement basés sur une date, et non sur la présence d'un mot-clé particulier). Ces échantillons ont ensuite été enrichies de balises afin de regrouper des groupes de termes sous la même étiquette.

Les balises utilisées sont :

— lieu : lieux de transit, hubs, comme *aéroport* ou *gare*, *arrêt de bus*, *gare routière* ;

— gare : les noms de gares, avec en attribut la géolocalisation lorsqu'il s'agit d'une gare francilienne [1] (pour une éventuelle visualisation) ; quelques abréviations de gares évoquées dans la sous-section 4.3.2 ont ensuite été ajoutées ;

— mode : le mode de transport évoqué (*train, bus, m.tro, voiture, ferry, rer, tram, tgv, taxi, navette, avion, tramway, ter, à pied, a pied, co-*

1. Ces listes de gares ont été constituées à partir de l'open data Transilien ou de la page http://www.sncf.com/fr/gares.

voiturage, transilien, vtc, v.lo, skate, trottinette) ;
— temporalité : détecter un temps d'attente, une heure, un temps de trajet détectés par l'expression régulière `/\d+\s?(h|min|mn|:|heures?)\s?\d*/`.

D'autres étiquettes ont été utilisées à titre d'essai : *qui*, qui recensait tous les pronoms personnels sujets (en essayant de mettre de côté le *il* impersonnel par exemple), et *agglutination*, qui regroupait les premières personnes tronquées (*jvais*, *jsuis*, *jv...*) afin d'observer leur emploi par rapport aux pronoms personnels standards.

Cette phase de pré-annotation ne se base que sur des méthodes symboliques et est dépourvu de système de désambiguïsation ; néanmoins, sa première version ne comptait que 8 insertions/suppressions/oublis sur 200 tweets. Une des améliorations importantes serait d'inclure une mesure de similarité entre deux chaînes de caractères (type Distance de Levenshtein) pour capter les petites erreurs graphiques.

Avant de nous arrêter sur un simple balisage, nous avons envisagé divers formats de sortie, notamment en étudiant la proposition de schéma TEI pour l'encodage des communications médiées par ordinateur développé dans le cadre du projet DeRiK (Beißwenger *et al.*, 2012), ainsi que les propositions du projet CoMeRe (Chanier *et al.*, 2014). Un corpus de tweets est par ailleurs en cours de constitution au sein de ce projet (bien qu'il ne concerne que des comptes politiques influents), mais en l'absence d'exemples de mise en œuvre et sans priorité quant à la formalisation du corpus, nous avons abandonné l'idée de produire un corpus XML-TEI ou XML formel et nous sommes concentrés sur un format de sortie optimal pour TXM (XML simple), en laissant de côté les métadonnées.

```
1 <s><name type="qui">J'</name>ai un <name type="mode">bus</
    name> pour la <name type="lieu">gare</name> d'<name type="
    sncf" geoloc="48.94700319,2.25717871305">argenteuil</name>
     à <name type="temporalite">59</name>.</s>
```

FIGURE 4.1 – Extrait d'une sortie du pré-annotateur

Cette structure permet d'effectuer des requêtes de la forme `[_.name_type="temporalite"]` dans le moteur de recherche de TXM.

4.3 Observations

Nous avons réalisé une série d'observations générales puis thématiques sur un corpus balisé de 358 168 mots.

4.3.1 Généralités

Noms

En observant les fréquences, on remarque la hiérarchie bus > train > tram > métro > pieds > voiture, et ce malgré les filtres de collecte sur *bus* (on ne collecte que *en bus*, *le bus*, *mon bus*). L'échantillon a été pris au mois de septembre : c'est le mois de la rentrée, beaucoup de collégiens, lycéens, étudiants en parlent sur Twitter, et ce sont bien souvent des utilisateurs « forcés » du bus (présence de *permis* plus loin dans les fréquences). Le peu de prédominance du métro et la bonne place du tram nous poussent à penser que le corpus de tweets dans son ensemble n'est pas si francilien qu'on aurait pu le croire. Le RER est compté comme une abréviation et se situe donc entre tram et métro en termes de fréquence. Le vélo apparaît un peu plus loin, suivi de la moto et du taxi (fréq. < 150). Il ne faut néanmoins pas oublier certains paramètres :

- La couverture 3G/4G dans le métro et le RER, qui ne permet souvent pas de tweeter en temps réel à Paris intra-muros (et même extra-muros sur de nombreuses lignes de RER), contrairement au tram ou au bus (ce qui n'empêche pas les utilisateurs de tweeter n heures après) ;
- L'évènement « rentrée » qui se couple à la jeunesse générale des utilisateurs de Twitter : mineurs, donc obligés de se déplacer en transports ou modes doux (vélo, marche) ;
- Les modes qui requièrent l'attention de leur utilisateur sont peu présents puisqu'on ne peut théoriquement pas tweeter en les utilisant : voiture, vélo, moto, scooter...

Forme	Fréquence	Forme	Fréquence	Forme	Fréquence
bus	9065	temps	380	jour	216
train	3897	monde	365	gars	211
@	2298	route	365	1h	210
retard	1674	vie	360	semaine	204
matin	1326	transports	351	cause	195
tram	1027	min	335	pieds	189
#	946	soir	333	voiture	178
gare	839	merde	332	jours	167
RT	748	avion	309	chauffeur	166
heure	641	lycée	308	pied	164
gens	600	journée	281	putain	162
arrêt	550	mec	259	lundi	160
cours	525	envie	248	ptn	160
métro	509	Putain	240	genre	158
trajet	459	place	235	marche	158
minutes	439	coup	232		
fois	422	mère	220		

TABLE 4.3 – Noms communs les plus fréquents dans le corpus de tweets pré-annoté

Verbes

On distingue plusieurs groupes de verbes qui correspondent aux différentes phases du processus de déplacement.

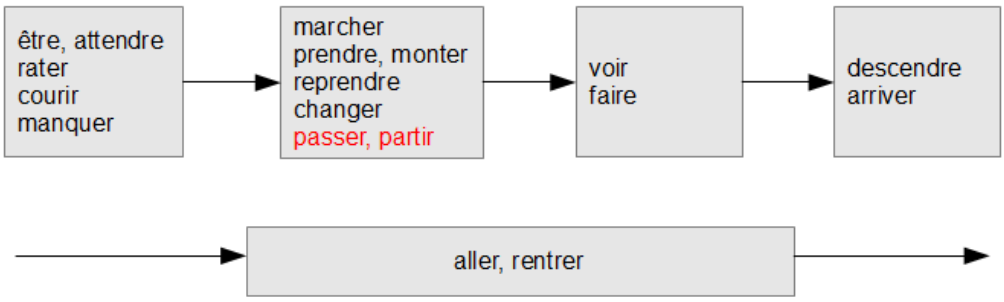

FIGURE 4.2 – Verbes de déplacement

Pronoms personnels

La première personne du singulier est largement majoritaire (*je, j', Je, me*) suivie de *il*, qui désigne principalement après examen des concordances un mode de transport (*il est passé avec 10 min de retard*).

4.3.2 Observations thématiques

Le fil QML

QML signifie *Quoi ma ligne* et est à la fois un compte @quoimaligne et un hashtag #qml. Le compte Twitter @quoimaligne qui retweete les usagers qui relatent leurs mésaventures dans les transports (majoritairement en Ile-de-France à quelques exceptions près). Les utilisateurs du hashtag relaient des infos qui émanent directement de la SNCF, ou racontent une situation qu'ils sont en train de vivre. Les spécificités de ce sous-corpus montrent ce qu'on pourrait appeler un « champ lexical de la situation perturbée SNCF » avec du vocabulaire d'annonces sonores (*retard, voie, supprimer, dir, vers, incident, trafic, omnibus, départ*, codes mission des trains...) alors que les tweets sont rédigés par des usagers. Les spécificités les plus fortes sont les hashtags d'indexation de lignes (#rera, #rerb...) : cela permet de rendre le tweet *investigable* (voir section 1.4.1) ou de contextualiser le message.

Ce fil permet à lui seul d'observer plusieurs manières de diffuser ou demander l'information (comment les usagers perçoivent et se réapproprient le vocabulaire SNCF ?), mais aussi diverses stratégies adaptatives en situation perturbée ainsi que des récits d'incivilités.

La localisation verbale

Si la géolocalisation n'est activée que dans 3% de notre échantillon de tweets, les utilisateurs de Twitter se situent énormément de manière verbale et contextuelle, soit en utilisant le hashtag #tweetloca (ou #tweetlocalisation), soit en précisant simplement où ils sont. On observe un grand nombre de tweets ayant la simple structure *Dans le [mode]* accompagné ou non d'une information topographique qui peut être très contextuelle, et qui nécessite d'avoir une connaissance des plans de transport. Par exemple :

> *J'voudrais remercier le rer d au terminus vlb qui me sauve de 25 min de conversation genante avec ma prof de lycee*

Il est possible de comprendre que *vlb* correspond à *Villiers-le-Bel* grâce à la mention du RER D (nous réabordons ce sujet au point 4.3.2).

Destinations et motifs de déplacement

Afin d'étudier la boucle de déplacement dans son ensemble, nous nous sommes concentrés sur les verbes *aller* et *rentrer*.

On dégage trois grandes classes de cooccurrents (contexte droit) du verbe *aller* :

- Déplacements contraints génériques : *cours*, *lycée*, *taff*, *cour*, *boulot*, *taf*, *fac*, *travail*, *bahut*, *bureau*, *stage*.
- Lieux : *Paris*, *Roissy*, *Champs*, *Châtelet*, *paris*, *Val* [d'Europe, de Fontenay], *panam*, *CDG*, *Versailles*, *Rosny*, *parc* [de Sceaux, Astérix, des Princes], *Montparnasse*, *Juvisy*, *77*.
- Loisirs : *parc* [de Sceaux, Astérix, des Princes], *Disney*, *stade*, *ciné*, *concert*, *soirée*.

Rentrer ayant deux usages, ses cooccurrents (contexte droit) se divisent en deux groupes :

- Rentrer de : *foot*, *vacances*, *soirée*, *taff*, *boulot*.
- Rentrer à/chez : *moi*, *soi*, *maison*, *toi*, *casa*, *oim*, *baraque*.

Dans le cas de *aller*, les scores indiquent que les déplacements contraints sont encore majoritaires ; viennent ensuite les lieux, puis les loisirs.

Micro-descriptions de scènes

Nous nous sommes ici concentrés sur les formes *ya*, *y'a* et *y a*. On observe dans ce sous-corpus différents types de tweets similaires à ceux du fil QML mais dont le lexique est moins marqué « SNCF » : description de la situation du trafic (surreprésentation de *y'a des problèmes sur [mode]*, *y'a des retards sur*, *y'a la grève*, *y'a des travaux*...), ou description de ce qui se trouve sous les yeux de l'usager. Cette dernière catégorie permet de nous éclairer sur les valeurs de l'usager : ce qu'il trouve anormal, comment il perçoit les autres usagers, comment il perçoit une ligne... Le lexique révèle qu'une des plus hautes fréquences de ce sous-corpus est *gens* (freq. 787). En examinant les concordances de ce terme, on tombe sur des descriptions de scènes en temps réel (*je suis dans le train, y'a...*) ou en temps décalé (*ce matin dans le train y'a...* avec conservation du présent) :

> *Dans le rer, j'aime bien observer les gens dans le reflet des vitres. D'ailleurs ce matin y'a un gars qui avait l'air triste...*

> *En mode cuisson a 300 degrés dans le RER ... Et y'a des gens avec des manches longues :0 ALLO ?!?*

mec (fréq. 721) est également dans les plus hautes fréquences :

> *Pourquoi y a un mec qui se balade dans le RER B en caleçon avec les motifs de cartes, un sweet adidas gris et vert et en claquettes ?*

meuf (fréq. 443) :

> *Il y a une meuf dans mon rer qui n'arrête pas d'éternuer. Je sens ses microbes essayer de m'atteindre.*

En revanche, *quelqu'un* n'apparaît que 70 fois :

> *Sinon dans le RER, y'a quelqu'un qui as senti mes cheveux !*

On y détecte diverses incivilités (vomi, pieds sur les sièges, cigarette...), des signes d'intolérance à des degrés différents (odeurs corporelles, microbe, volume des conversations...), mais aussi du racisme, de l'homophobie ou de la transphobie, ainsi que des témoignages de délits (exhibitionnisme, agressions) :

> *bordel dans notre rame et genre en sortant du RER y a un mec qui en a profité pr péta l'iPhone 5 d'une chinoise ...*

> *Jcrois y'a un gay en face de moi dans le RER rien qu'il lâche des regards bizarres j'ai peur*

> *Dans le rer, je suis entouré d'indiens, je n'ai rien contre eux, mais normal qu'il y a quelques moustiques qui se promenent ? Je crois pas*

Multimodalité, intermodalité, stratégies adaptatives

Pour cette thématique, il s'agissait de trouver des motifs récurrents qui permettent d'exprimer les enchaînements plus ou moins complexes de modes. par exemple, l'expression régulière `/pren.+le(.+?)pui(s)?le(.+?)/` permet de capter des chaînes de déplacement complexes :

> *Jprends le train directe de 13h35 Pour aller a la Gare d'austerlitz puis le RER C Pour Francois Mitterrand*

> *Je me prépare vite vite pcq faut que jprenne le 19 après le RerA puis le 312 ! Et tout ça pour aller à champ voir ma meilleure amie*

Cette structure permet également de capter des stratégies adaptatives :

> *Ce soir hors de question que je prenne le RER A. Je prend la ligne L, puis le RER E. J'en ai juste ras le bol.*

> *Problème avec le métro 4, devoir prendre le bus puis le métro puis le rer et à nouveau le bus et au final avoir son arrêt non desservi !*

L'expression régulière `/(pas|plus)\sd.\s?[mode]/` permet également de relever des stratégies adaptatives :

> *Y'a pas d'rer C, j'suis obliger d'aller jusqu'à st-lazare puis Argenteuil pour aller à Epinay. Saha l'détour !*
>
> *Demain je dois aller à Opéra ya pas de train jusqu'à Paris Est, je dois aller à Chessy pour taper le rer A putain !*

Les tweets mentionnant deux modes de manière suffisamment rapprochée (fenêtre de 0 à 5 mots) se révèlent peu nombreux : le passage au quantitatif (déterminer sur l'enchaînement bus-RER est particulièrement fréquent par exemple) paraît pour le moment difficile.

Forme	Fréquence	Forme	Fréquence
bus / tram	4	train ou bus	3
train et le bus	4	tram bus	3
train métro	4	Bus train	3
bus / métro	3	métro et le RER	3
bus et le train	3	métro RER	3
bus et le tram	3	bus + rer	2
bus et tram	3	bus + train	2
bus train	3	Bus + tram	2
train bus	3	Bus + Tram	2
train et le métro	3	bus et le métro	2
train metro	3	bus / train	2
		tram - bus	2

TABLE 4.4 – Expressions d'enchaînement de deux modes

Expressions argotiques ou familières, abréviations

Nous avons relevé des **usages argotiques**, comme *leurs*, *leurleurs*, ou *charognards* pour désigner les contrôleurs :

> *En se moment quand jprend la ligne H que ya les leur mes jme fai jamais controler !*

On trouve également des mentions de gares raccourcies, issues de l'oral (*saint-laz*) ou plus ou moins encouragées par la contrainte des 140 caractères (*psl*, *mlv*). Ces sigles ne sont cependant pas générés par la contrainte des 140 caractères : *MLV* ou *GDN* préexistaient à Twitter. Nous avons notamment extrait les abréviations de 3 caractères les plus courantes :

abréviation	nom complet
gdn	Gare du Nord
cdg	Charles de Gaulle
bfm	Bibliothèque François Mitterrand
blr	Bourg la Reine
mlv	Marne la Vallée
gdl	Gare de Lyon
psl	Paris Saint Lazare
vsg	Villeneuve - Saint-Georges
vlb	Villiers-le-Bel [ou Vélib]
psy	Poissy [ou psy]
bnf	Bibliothèque François Mitterrand
sqy	Saint-Quentin en Yvelines
mlj	Mantes-la-Jolie
clh	Châtelet les Halles
pds	Parc de Sceaux
nlg	Noisy le Grand
vdf	Val de Fontenay

TABLE 4.5 – Exemples de formes courtes pour désigner des gares (trois lettres)

Temporalités

L'expression d'informations temporelles est un enjeu important pour la mobilité : les utilisateurs de Twitter donnent-ils des informations temporelles précises dans le contenu même de leurs tweets (outre l'horaire d'émission présente dans les métadonnées) ?

Forme	Fréquence	Forme	Fréquence
min	330	20min	69
1h	253	1h30	68
:	225	18h	66
8h	170	15min	65
9h	122	30min	64
2h	115	17h	62
7h	105	6h	62
30	95	15	59
10min	88	16h	57
10	81	5min	55
20	74	4h	51
10h	71	5h	49

TABLE 4.6 – Expressions temporelles numériques et leurs fréquences au sein du corpus de tweets annotés

Les deux points témoignent de la présence assez forte de la forme 00:00, indiquant un horaire précis, contrairement aux formes 00h, 00min ou 00h00 qui expriment plus souvent des durées.

Forme	Fréquence
7 : 30	6
1 : 30	4
18 : 10	4
12 : 00	3
12 : 20	3
17 : 10	3
17 : 30	3
18 : 30	3
5 : 30	3

TABLE 4.7 – Formes de type 00:00 les plus fréquentes

Les fréquences sont très faibles, on observe en fait beaucoup d'hapax (des horaires précis qui n'apparaissent qu'une fois).

Forme	Fréquence
5min	55
10 min	54
30 min	48
20 min	46
15 min	37
5 min	31
40 min	20
45 min	20
2 min	15
2min	13
25 min	12

TABLE 4.8 – Formes de type 00min les plus fréquentes

Forme	Fréquence
1h	253
8h	170
9h	122
2h	115
7h	105
1h30	68
6h	62
4h	51
5h	49
8h30	46
3h	40

TABLE 4.9 – Formes de type 00h(00) les plus fréquentes

4.4 Corpus « Fraude »

S'il existe des applications mobiles de signalement des contrôleurs, cette pratique est également très répandue sur Twitter. Elle repose principalement sur le caractère *investigable* des tweets, via le moteur de recherche mais également via les hashtags (cliquables, qui permettent d'afficher le fil construit autour de l'utilisation de ce hashtag). Les manières d'indexer les signalements de contrôleurs sont néanmoins pléthoriques et moins organisées que

le fil QML par exemple. Sur 30 tweets d'avertissement, on relève ces *technomots*[2] (Paveau, 2013) d'indexation :

Type	Technomot
Compte Twitter fédérateur	@ControleurRATP
Lignes	#rerb
	#t4
	#LigneH
	#LigneJ
	#rera
	#RERD
Lieux	#Paris
	#chatelet
	#Argenteuil
	#ErmontEaubonne
	#95
Généralistes	#sncf
	#ratp
	#Controleurs
	#InfosControleurs
	#controleur
Créations	#PointFraude
	#infosduhood
	#Pointfraude
	#teamfrodeurs
	#vigilenceensemble
	#planqué

TABLE 4.10 – Liste de technomots utilisés dans les tweets de signalement des contrôleurs

@ControleurRATP est un compte très utilisé par ce réseau : il retweete les usagers qui lui adressent ces avertissements, mais il est également possible d'accéder à un fil complet de signalement des contrôleurs en recherchant tous les tweets mentionnant @ControleurRATP (et, contrairement aux apparences, ces tweets concernent aussi bien la SNCF que la RATP). Les hashtags

2. *mots clicables qui prennent des dimensions non existantes hors ligne, ajoutées aux fonctions langagières préalables*, selon Paveau M.-A., 26 août 2012, Linguistique et numérique 4. *Les écritures de Protée : identités pseudonymes, La pensée du discours [carnet de recherche]*, http://penseedudiscours.hypotheses.org/?p=10057, consulté le 16 octobre 2014

dits « créatifs » n'ont pas forcément vocation à indexer le tweet à proprement parler (pas autant que #controleur par exemple, qui renvoit à un canal complet) et ont un côté plus ludique que pragmatique.

On trouve également dans ce corpus des récits de fraude :

> *PAS DE CONTROLEUR OUUUUHYEAH 22E ECONOMISÉS*
>
> *Jetais en train de frauder oklm jvois les leurs ils sont devant la porte. Jai couru dans lautre sens crary*

Mais aussi des reproches de la part des payeurs :

> *jamais ils sont là les controleurs quand j'paye mon billet de train fdp*
>
> *Putain a chaque fois que jprends le train ya jamais les controleurs jcommence a avoir le seum de payer le ticket*

Les concordances du motif *contr.leur* révèlent différents types de tweets concernant la fraude, autres que la délation de contrôleurs : les reproches (pas assez ou trop de contrôles), les astuces...

Contexte gauche	Pivot	Contexte droit
la best des baratineuses pour échapper à l'amende des	contrôleurs	de bus. On attend le bus.....
de tension a la gare Lille Flandres. Échapper aux	contrôleurs	de tram : fait. Ça contrôle sec
bus. Crary 150bal de train et même pas un	contrôleur	# SncfDeMerde. Le bus est trop rapide j'vai
un train la ca me stress. RER C les	contrôleurs	sont de sortie. On prend le train de 15h21
# tarn. Putain mais dans mon bus ya un	controleur	en civil genre heureusement que javais validé. putin je
du mois de Septembre de la RATP : Planquer 10	contrôleurs	pour piller les étudiants sortant de la fac ....
Ma vie ca sresume aux horaire de bus. #	contrôleurs	a magenta juste avant d'accéder au RER E.
trajet reste a déterminé selon notre nombre ?. Les	contrôleurs	SNCF.... les pires FDP discriminants. @ B_ang_bang
plus qu'à prier qu'il y est pas de	contrôleurs	... Ma cousine 31 / 31 pour le permis Taxi
aller au casino ce soir ; j'ai parie no	controleur	dans le train et j'ai eu raison. @
à la défense c'est une galère y'a des	contrôleurs	partout. Strasbourg tram. Bon truc trop chiant quand
. J'espère que je vais pas tomber sur les	contrôleurs	putain. Ils augmentent tellement tous que c'est devenu
ce fdp. J'me suis fais arreter par les	controleurs	car j'avais pas de carte de bus donc jdois
retard. Merci les boloss qui s'embrouillent avec les	contrôleurs	. MERCI. Des le matin jai couru jetais en
RDV ! Elle est toujours en retard. @ ControleurRATP	controleur	à reuilly diderot un peu partout. 15h00. J'
: A la défense sorti F y'a plein de	contrôleurs	attention # rera # RATP. Et en faire la
. Être debout dans ce train de merde. Attention	contrôleurs	dans les RER D entre Gare du Nord et Villiers-Le-Bel
Putain je prie pour que demain y'a pas de	controleur	dans mon bus. J'ai mon permis moto.

TABLE 4.11 – Concordances (partielles) de *contr.leur.?*

Chapitre 5

Typologie de tweets et classification automatique

5.1 Typologie des tweets pertinents « Expérience Voyageur »

5.1.1 Typologie

Une fois les données collectées, nettoyées, observées et analysées, il s'agit d'identifier formellement quels types de tweets concernent notre problématique, afin de définitivement mettre de côté les tweets non-pertinents et réduire au maximum la masse de données à traiter.

Nous avons établi une typologie des types d'informations détectables dans les tweets sur la base de nos observations articulées aux besoins identifiés dans les branches SNCF. Nous recherchons les tweets « Expérience Voyageur », majoritairement à la première personne. Elle n'est pas exhaustive et répond à deux besoins : détecter des informations assez fines, et la nécessité d'établir des contrastes entre les types de tweets.

TABLE 5.1: Typologie de tweets « Expérience voyageur »

Classe	Sous-classe	Description
Moi/ma subjectivité (perception, goûts, préférence ; système de référence personnel, représentation structurelle)	Mes habitudes/mon quotidien	Description d'une habitude, d'actions répétées

	Ce que je trouve anormal, ce qui m'est désagréable	Attitude négative vis-à-vis de quelque chose, difficultés à accepter
	Préférences de mode	Opinion sur un mode (par rapport à un autre, ou pas)
	J'apprécie	Commentaire positif
Savoir et faire savoir : partage de l'information	Demande d'information	Questions directes, questions lancées à l'assemblée, interrogations
	Diffusion d'information	Bons plans, incidents, réponse à une question
	Exhortation	Encouragement subjectif à utiliser tel mode ou emprunter tel itinéraire
Mobilité en temps réel : Nature du parcours / description et justification du parcours	Localisation (où suis-je à l'instant T)	Tweet-localisation
	Où je vais/qu'est-ce que je vais faire	Motifs de déplacement et destinations
	J'explique mon parcours ou un segment du parcours	Enchaînements de modes, de gares, d'horaires...
Mobilité en temps réel : occupations / activités pendant le parcours	Ce que j'ai vu	Description d'une situation où l'utilisateur est resté passif
	Ce que j'ai fait	Description d'une situation où l'utilisateur a été actif
Perturbation du parcours	Je suis responsable	L'utilisateur perturbe son parcours (réveil, retard, flemme...)
	Je ne suis pas responsable : Incident	Le parcours de l'utilisateur est perturbé par un incident imprévu

	Je ne suis pas responsable : Administratif	Le parcours de l'utilisateur est perturbé par des problèmes « administratifs » (travaux, changements d'horaires, suppression d'arrêts, gestion de la carte de transports...)
	Je ne suis pas responsable : Information	Le parcours de l'utilisateur est perturbé à cause d'une mauvaise diffusion de l'information
Fraude	Avertissement/délation	Signalement de contrôleurs
	Reproches	Reproches adressés à la SNCF (manque de contrôle, trop de contrôles...)
	Astuces	Truc, manière de frauder, failles dans le système
	J'ai fraudé/je fraude	
	Je projette de frauder/j'aimerais frauder	
	J'ai fraudé sans savoir/J'ai été obligé de frauder	L'usager a fraudé par obligation ou par accident (perte du pass navigo, mauvaise information...)
	Comportement des contrôleurs	Comportement inhabituel des contrôleurs
Autres	SNCF	Tweets infos sur le statut de la SNCF, tweets adressés à la SNCF directement ou indirectement
	Expression de la durée	Tweet avec une expression de durée de parcours, durée d'attente

	Tierce personne	L'utilisateur raconte / commente le parcours d'une tierce personne
Inclassables		

5.1.2 Phase d'annotation manuelle

Nous avons ensuite annoté les tweets pertinents selon les sous-classes de la typologie. Nous avons classifié 591 tweets, au format CSV : pour chaque tweet, chaque sous-classe est activée (1) ou inhibée (0). Un tweet peut être affecté à plusieurs sous-classes.

Après annotation, on note une prépondérance de la sous-classe "J'explique mon parcours", suivie de "Activités pendant le parcours : ce que j'ai vu" et "Ma subjectivité : ce qui m'est désagréable" (figure 5.1).

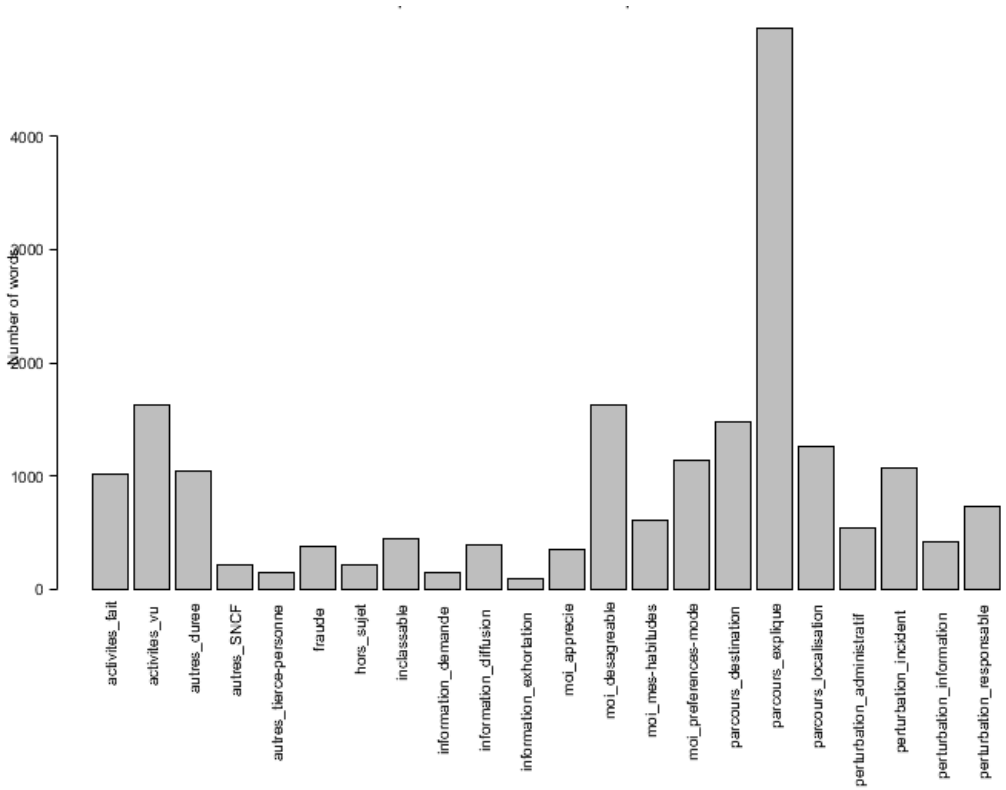

FIGURE 5.1 – Proportion de chaque sous-classe au sein du corpus annoté

5.2 Classification automatique à partir des observations

5.2.1 Choix des classes

La typologie ci-dessus n'a pas été conçue pour la classification automatique, mais simplement dans un but descriptif. Néanmoins, nous avons tenté de classifier les tweets selon les différentes sous-classes de la typologie, pour deux raisons : amorcer l'automatisation du repérage d'informations fines au sein des tweets, et obtenir un aperçu du taux d'homogénéité des différentes sous-classes.

Nous avons tenté deux approches : l'approche classique en classification multinomiale qui consiste à entraîner un classifieur binaire distinct pour chaque classe (méthode One-vs-All, OvA) [**méthode 1**], et une approche plus brutale qui consiste à considérer chaque groupe de classes possible comme étant une seule classe distincte. Par exemple, si le tweet *Wallah sa pue le rayon poisson dans le tram* est classifié à la fois dans la classe "Ma subjectivité : ce que je trouve désagréable" (moi_desagreable) et dans la classe "Nature du parcours : localisation" (parcours_localisation), on considérera que "moi_desagreable/parcours_localisation" comme étant une seule classe. Les multiples combinaisons donnent au final près de 170 classes différentes [**méthode 2**].

Nous avons réutilisé les éléments du classifieur pertinent/non pertinent et nous sommes donc concentrés sur le bayésien naïf.

5.2.2 Tests

Baseline

Nous choisissons comme paramètres de base les quadrigrammes de caractères et leur fréquence, et non les occurrences des formes, du fait du caractère extrêmement court d'un tweet (13 mots en moyenne par tweet sur un échantillon de 100 tweets de notre corpus). Le corpus d'apprentissage est composé de 691 tweets annotés, et le corpus de test comporte 60 tweets.

Pour la méthode 1, les seuls quadrigrammes donnent des résultats plutôt encourageants malgré le nombre de classes ramené à la taille du corpus. Les classes les plus ouvertement homogènes, c'est-à-dire qui comportent un vocabulaire assez restreint et des structures récurrentes (comme "parcours_localisation"), obtiennent naturellement les f-mesures les plus élevées (tableau 5.2).

Classe	f-mesure	Nombre de tweets réellement dans la classe
moi_mes-habitudes	0	1
moi_desagreable	0.67	6
moi_preferences-mode	0.67	1
moi_apprecie	x	0
information_demande	1	1
information_diffusion	0.8	3
information_exhortation	x	0
parcours_localisation	0.71	10
parcours_destination	0.86	4
parcours_explique	0.84	24
activites_vu	1	9
activites_fait	0.67	5
perturbation_responsable	0.5	3
perturbation_incident	0.57	5
perturbation_administratif	0.5	3
perturbation_information	x	0
fraude	0	1
autres_SNCF	x	0
autres_duree	1	1
autres_tierce-personne	1	1
inclassable	0.36	9
hors_sujet	0	12

TABLE 5.2 – Evaluation de la classification automatique basée sur des quadrigrammes de caractères (*baseline*)

Nous pensions à première vue que la méthode 2 donnerait de très mauvais résultats. Au final, la première configuration à base de quadrigrammes permet d'annoter correctement 29 tweets sur 60 (avec toutes leurs classes), ainsi que 15 tweets comportant au moins une classe correcte. Après examen des résultats de classification, on remarque que le classifieur parvient même à classifier certains tweets dans des classes complexes, comme "information_diffusion / parcours_destination / parcours_explique / perturbation_incident".

Ajout de critères textométriques

En reprenant des éléments de la méthode présentée dans (Eensoo et Valette, 2012), nous avons répertorié des descripteurs de classe afin de tenter d'améliorer notre classifieur.

Ce qui m'est désagréable Interjections familières et leurs variantes raccourcies (*putain*, *ptn*) ; formes à tonalité négative (*galère, horrible*) ; expressions de la répétition (*tous, chaque*) ;

Mes habitudes Expressions de la répétition, temporalités à l'échelle de la journée (qui ont un lien avec le pendularité : *matin, 7h*) ;

Préférences de mode *aimer/détester*, *jamais/toujours* ;

J'apprécie *emoji*, formes à tonalité positive ;

Demande d'information Le point d'interrogation ; formes diverses (*horaires, où, appli*)

Diffusion d'information Vocabulaire caractéristique d'annonces SNCF (*interrompu, supprimé, modifiée, circulation*) ; hashtags liés aux lignes SNCF ;

Exhortation L'arobase ;

Localisation *je suis dans* sous toutes ses formes (agglutinations) ; #tweetloca ;

Où je vais, qu'est-ce que je vais faire (destination) *(pour) aller, (pour) rentrer* ; expressions caractéristiques sur Twitter (*Go [lieu]*)

J'explique mon parcours Informations temporelles ; verbes cooccurrents des modes (*prendre, monter*) ;

Pendant le parcours : ce que j'ai fait Pronoms personnels réfléchis ;

Pendant le parcours : ce que j'ai vu *y'* ; différentes désignations familières d'autrui (*mec, meuf, gars*) ;

Perturbation du parcours : responsable Adjectifs possessifs ; *louper*, *rater*, *courir* ;

Perturbation du parcours : pas responsable Univers *travaux*, *horaires* pour [administratif] ; Univers *retard*, *bondé* pour [incident] ; Univers *quel*, *savoir*, *horaires* pour [information] ;

Autres : SNCF Vocabulaire caractéristiques d'annonces SNCF ; mention de comptes Twitter officiels SNCF ;

Autres : Expression de la durée Informations temporelles ;

Autres : Tierce personne Pronoms personnels sauf première personne.

Nous n'avons pas traité les classes "Hors sujet" et "Inclassable" qui sont, par nature, difficiles à décrire.

Nous avons appliqué ces critères à la méthode 1 et avons constaté une légère amélioration de la f-mesure sur certaines classes, comme "Ce que je trouve désagréable" (0,8) et "Destination" (1). On ne constate aucune baisse de la f-mesure sur l'ensemble des classes.

En revanche, l'application de ces critères « en bloc » (puisque nous avons un seul classifieur pour l'ensemble des classes) pour la méthode 2 fait légèrement baisser les performances : on passe de 29 à 27 tweets classifiés correctement.

5.2.3 Discussion

Dans un premier temps, nous nous sommes demandés si les critères textométriques allaient être pertinents pour des messages de 13 mots en moyenne. Néanmoins, même peu significatifs, les résultats de la méthode 1 restent encourageants pour des corpus d'apprentissage et de test assez restreints. Nous abordons le travail en cours pour améliorer le classifieur dans les perspectives (section 6).

Le fait que des classes complexes récurrentes émergent avec la méthode 2 pourrait témoigner d'un besoin de refondre la typologie. Après annotation, nous constatons qu'il y manque certaines informations : par exemple, une sous-classe qui concernerait le confort physique, le ressenti (odeurs, chaleur...), qui est également une préoccupation pour les opérateurs de transport. Dans cette optique de refonte, il serait intéressant de tenter une opération additionnelle de clustering sur le corpus de tweets.

Chapitre 6

Perspectives

6.1 Amélioration de la classification automatique

La classification automatique des tweets représente un travail en soi. Pour l'améliorer, il serait nécessaire d'utiliser des méthodes itératives (étiquetage, validation, injection dans le corpus d'apprentissage). Une interface PHP simple de validation a été réalisée pour alimenter une base de données de tweets étiquetés. A termes, cette base de données pourrait servir à la recherche précise d'informations en rapport avec un thème particulier de la typologie.

6.2 Multiplier les types de données : étendre la collecte

Pour contrebalancer les biais de Twitter déjà évoqués, il serait intéressant d'effectuer les mêmes analyses sur d'autres données textuelles issues du web, dont les pratiques textuelles diffèrent : blogs, commentaires de blogs, forums, autres RSN (réseaux sociaux numériques), etc. Nous avons envisagé deux autres sources principales.

6.2.1 Blogs de ligne Transilien/TER

Les blogs de ligne sont une initiative lancée par Transilien et TER afin d'améliorer les échanges entre les clients et le responsable de ligne. Ce dernier y publie des billets concernant les travaux, la ponctualité moyenne, les animations dans les gares de la ligne, les perturbations de la journée avec

explications détaillées... Les utilisateurs du blog et généralement clients de la ligne sont libres de commenter ces billets. Ce sont ces commentaires, sous forme de questions, de coups de gueule ou de coups de cœur qui nous intéressent.

Actuellement, les blogs Transilien sont au nombre de neuf (RER A, C, D, E ; lignes H, J, L, N et U, P) [1], et on compte un peu moins d'une quarantaine de blogs TER. Ils sont fréquentés par beaucoup d'habitués, mais aussi des gens de passage. Les commentaires font majoritairement mention de trajets domicile-travail quotidiens, avec quelques incursions de voyageurs ponctuels en quête d'informations. Les blogs TER regorgent de cas isolés qu'on peine à observer ailleurs (hormis dans les réclamations clients), et surtout pas sur Twitter (proéminence des utilisateurs urbains). Lorsque les commentaires sont des réclamations, ils peuvent être assez longs et détaillés avec de nombreuses mentions de gares et de numéros de trains, des marqueurs de modalité classiques, etc... L'orthographe et la syntaxe sont assez standards.

6.2.2 Forums

Les forums sont également une source de données importante lorsqu'il s'agit de capter des microchronologies, des récits de parcours, des avis... Les formats sont en général très similaires d'un forum à l'autre, mais la syntaxe et l'orthographe dépendent énormément de la politique des administrateurs et modérateurs.

Nous avons envisagé trois types de forums : de grands forums généralistes [2], des forums de consommateurs et des forums spécialisés pour le voyage. Le moteur de recherche spécialisé dans les forums Boardreader [3] permet de se faire une idée des sources de données possibles.

1. Il existe également un blog pour le RER B mais qui ne fait pas partie des blogs gérés par Transilien.

2. Si jeuxvideo.com ou hardware.fr n'apparaissent pas comme des forums généralistes de prime abord, les discussions qu'on peut y lire abordent tous les sujets du quotidien.

3. http://boardreader.com

Type	Forum
Forums généralistes	forum.aufeminin.com
	www.jeuxvideo.com/forums
	forum.doctissimo.fr
	forum.hardware.fr
	www.commentcamarche.net/forum/
	linternaute.com/forum/
Forums de consommateurs	forum.quechoisir.org
	www.60millions-mag.com/forum/
Forums sur le voyage	voyageforum.com
	routard.com/comm_forum_de_voyage
	tripadvisor.fr

TABLE 6.1 – Liste embryonnaire de forums envisagés pour une collecte

6.3 Croiser les données

Le croisement de données est l'enjeu majeur de ce travail : il s'agit dans un premier temps de croiser les données qui ont émergé de la fouille de textes avec les métadonnées des textes eux-mêmes (date d'émission du tweet, plateforme d'émission du tweet, fuseau horaire...). Dans un second temps, ces informations sont à croiser avec d'autres informations quantitatives issues d'autres sources (statistiques, enquêtes...).

Dans ce contexte, un sujet de thèse est en cours de lancement sur la qualification des déplacements multimodaux à l'aide de données textuelles, basé sur les pistes ouvertes dans ce mémoire. Il s'agirait de mieux connaître ces chaînages complexes encore difficilement appréhendés par les AOT et les acteurs du transport, en analysant diverses sources de données textuelles (tweets, blogs, etc.) et en croisant ces analyses à des données structurées tels que les plans de transport, des bases de données d'horaires, ou encore des tendances qui auraient émergé d'enquêtes d'opinion.

Conclusion

Nous avons pu mettre en place une collecte pérenne qui nous a permis d'effectuer des explorations textométriques, à travers lesquelles nous avons pu identifier quelles informations potentiellement intéressantes étaient présentes dans ces données. Nous avons pu formaliser ces informations en établissant une typologie et en amorçant un travail de classification automatique qui permettra, à termes, de produire des indicateurs thématiques sur la mobilité.

Nous avons montré que les données textuelles, même dans leurs manifestations les plus laconiques, constituent une source importante d'informations sur les pratiques de mobilité. Elles permettent de découvrir des habitudes, des opinions, des préférences concernant les modes ou même les pratiques de fraude, qu'il serait plus difficile, plus lent, ou plus coûteux de déceler dans d'autres types de données.

A termes, le croisement de ces données avec d'autres initiatives permettrait de produire des indicateurs précis quant aux évolutions à venir en matière de mobilité.

Table des figures

Liste des tableaux

Bibliographie

Beisswenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. et Storrer, A. (2012). A TEIschema for the representation of computer-mediated communication. *jtei*, 3.

Borra, E. et Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 66(3):262–278.

Cervulle, M. et Pailler, F. (2014). #mariagepourtous : Twitter et la politique affective des hashtags. *Revue française des sciences de l'information et de la communication*, (4).

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., R. Wigham, C., Hriba, L., Longhi, J. et Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. Final version to Special Issue of JLCL (Journal of Language Technology and Computational Linguistics (JLCL, http://jlcl.org/): BUILDING AND ANNOTATING CORPORA OF COMPUTER-MEDIATED DISCOURSE: Issues and Challenges at the Interface of Corpus and Computational Linguistics (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel).

Cheng, Z., Caverlee, J. et Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA. ACM.

Ducos, A., Bonnet, V., Marchand, P. et Ratinaud, P. (2014). Classification d'un corpus hétérogène : la page facebook de soutien au « bijoutier de nice » (septembre 2013). *In Actes des 12èmes Journées internationales d'Analyse statistique des Données Textuelles*, Paris.

Eensoo, E. et Valette, M. (2012). Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des

sentiments. *In* Antoniadis, Georges, Blanchon, Hervé et Sérasset, Gilles, éditeur : *TALN 2012*, volume 2, pages 367–374, Grenoble, France. GETALP-LIG.

Gallez, C. et Kaufmann, V. (2009). Aux racines de la mobilité en sciences sociales. *In De l'histoire des transports à l'histoire de la mobilité ?*, Histoire, pages 41–55. Presses Universitaires de Rennes.

Heiden, S., Magué, J.-P. et Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *In* Sergio Bolasco, Isabella Chiari, L. G., éditeur : *Statistical Analysis of Textual Data - Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, volume 2, pages 1021–1032, Rome, Italie. Edizioni Universitarie di Lettere Economia Diritto.

Kergosien, E., Laval, B., Roche, M. et Teisseire, M. (2014). Opiland : identification de la perception des territoires par la fouille de texte. *MASHS*, RNTI-SHS-2:185–212.

Morency, C., Verreault, H. et Bourbonnais, P.-L. (2013). Évaluation des potentialités du web comme outil de collecte de données sur la mobilité. Rapport final. Étude réalisée pour le compte du ministère des Transports du Québec.

Morstatter, F., Pfeffer, J., Liu, H. et Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. *CoRR*, abs/1306.5204.

Paveau, M.-A. (2013). Technodiscursivités natives sur Twitter. Une écologie du discours numérique. *Épistémé (Revue internationale de sciences humaines et sociales appliquées, Séoul)*, 9:139–176.

Ratinaud, P. (2014). Visualisation chronologique des analyses alceste : application à twitter avec l'exemple du hashtag #mariagepourtous. *In Actes des 12èmes Journées internationales d'Analyse statistique des Données Textuelles*, Paris.

Smyrnaios, N. et Ratinaud, P. (2014). Comment articuler analyse des réseaux et des discours sur twitter. *tic&société*, 7(2).

Annexe A

Exemples de tweets pour chaque sous-classe de la typologie « Expérience Voyageur »

Moi, ma subjectivité

Mes habitudes, mon quotidien

vous plaignez pas je suis dans le bus à 7h tout les jours moi

La chose qui me soule le plus c'est de devoir reprendre ce RER B tous les jours

Je suis toujours la première a arrivée a l'école . Si je prends le métro d'après je dois courir et j ai pas envie

Ce que je trouve anormal, désagréable

Bref chaque année transpole ils font de pire en pire avc nos horaires de bus

Le rer c'est vraiment la merde ca me manque le bus

Libourne (33). TER pour Périgueux bondé. La galère des trains du quotidien, ce n'est pas seulement en Île-de-France.

Préférences de mode

Putain 1h de bus tous les matins et soir alors que en voiture il y en a pour 20minutes ..

Le rer c'est vraiment la merde ca me manque le bus

je crois que cette aprèm je vais marcher au lieu de prendre le train

J'apprécie

Par contre fini les 15min de marche le matin et le soir et ça c'est cool.

Et askip bis maintenant que c'est la rentrée et que y'a les nvx horaires de RER c'est le bonheur

Au moins à cette heure là les bus ne sont pas bondés faut voir le bon côté

des choses

Savoir et faire savoir : partage de l'information

Demande d'information

Les horraires de bus c'est ceux d'hiver maintenant?

Ça coûte combien un aller/retour Amiens en TER ?

j'sais pas si c'est mieux de faire Cergy -> st lazare puis prendre le M12 pour porte de versaille ou d'aller à la défense et prendre le T2

Diffusion d'information

Génial ! Le #SNCF_P reprend du service ! Trafic normal axe Paris Est/-Château Thierry #OK

Je fais fougères rennes dans l'aprem si ça intéresse quelqu'un contactez moi par pigeon voyageur

D'après le site TER, des trains directs sont rétablis entre Lyon et Saint-Etienne demain @MuseomixRA #onycroit

Exhortation

@Capetlevrai pour aller au salon ? si tu cherches un moyen pour arriver près du salon vasy en métro

@Dreyzuu J'pense que tu devrais prendre le T2 mec

Mobilité en temps réel : Nature du parcours / description et justification du parcours

Localisation

#tweetloca - Mairie des Lilas

Dans le train pour Bourg...

Gare de Nancy

Où je vais, qu'est ce que je vais faire

Dans le bus je rentre chez moi!

J'me tape 2h30 de route alle-retour juste pour une conférence inaugurale

J'explique mon parcours ou un segment du parcours

Je dois prendre le bus a 7h07 je vais décédé

Demain je me tape tous le RER A

C'est pas "métro-boulot-dodo" moi, c'est "bus-train-tram-bus-boulot-bus-tram-train-bus-dodo"

Occupations / activités pendant le parcours

Ce que j'ai vu

dans le bus un malade gueulait il insultait tout le monde parce qu'il payait 105 euros sa navigo et que les bus étaient en retard

Ya une dame dans le train qui vient de se mettre de l'huile de ricin dans les cheveux ca sent trop fort c'est horrible!!

Y'a très peut de monde ce matin dans les bus ;0 !

Ce que j'ai fait

Je vais essayer de bosser dans le bus/tram pour mon dst en italien.

1% en écoutant la musique dans le bus aller retour et utiliser la 3g 3-4 fois dans la journée mais il a résisté mon petit Sony

Ce qui est bien c'est que pendant les 45min de bus je peux dormir

Perturbation du parcours

Je suis responsable

Premier jour je me trompe déjà de bus moi

D'accord ça commence déjà je suis en retard j'vais louper mon train

Je ne suis pas responsable : incident

Voilà j'attend mon train 10 min de retard donc je vais être en retard mais c'est pas grave ils ont l'habitudes. Ça commence bien !

Le tram bloqué 20min à faubourg de saverne le matin c'est la première cause de mes retards

Je ne suis pas responsable : administratif

Jeudi je commence à 10h mais mon bus est à 8h09 ! Parce qu'à 9h45 il y en a un mais c'est trop tard ..

Horaires de rentrée. Le bus 1 ne change pas. La correspondance aka le bus 2 passe 5mn plus tôt. Je le rate. Je perds donc 25mn. Glorieux.

Encore des travaux a la #SNCF Qui dit travaux dit retard de train

Je ne suis pas responsable : information

Je sais même pas quand est mon bus –'

Je descend le contrôleur du quai me dit qu'il part je remonte. Le train est plein et on n'est pas partis... 2/2 #ter stE Lyon

Fraude

Avertissement / délation

Controleurs en gare d'Aulnay Sous Bois #SNCF #RERB #T4

Masse de controleur a bras de fer les gens! Attention les cowboy son de sortie

@ControleurRATP contrôleurs a gare de Nogent sur marne et dans le RER a direction Boissy-Saint-Léger

Reproches

Putain a chaque fois que jprends le train ya jamais les controleurs jcommence a avoir le seum de payer le ticket

jamais ils sont là les controleurs quand j'paye mon billet de train fdp

@SNCF_QR constats du jour : contrôle billets dans TER vide en milieu de journée et tjrs fraude massive sur @RERD_SNCF nord bondé le soir.

#RERD: tu quittes le quai pour voir plus d'info en gare tu ne peux plus reprendre ton train... Les fraudeurs n'ont pas le pb #Navigo #QML

Astuces

Avec la carte #voyageur #sncf : voyagez gratos lol. Le contrôleur ne verifie même pas le billet car pas le matos pour...

Sabi la derniere fois pour froder le metro jme souvien elle a essayer de passer a 4pattes dans le truk elle est rester bloqué

Hella le train est blinde l'avantage pour les fraudeurs comme les leurs ne nous contrôlera pas mais je vais rester debout tout le trajet

J'ai fraudé/je fraude

Merci au controlleur de mon train d'avoir dormi tout le voyage me permettant de frauder sans problème #sncf

Controleur en civile on s'est bien fait niquer

Jetais en train de frauder oklm jvois les leurs ils sont devant la porte. Jai couru dans lautre sens crary

Faut qui me prête son passe navigo j'en ai bsoin yen a marre de froder

Je projette de frauder/j'aimerais frauder

Froder le tgv c'est mon reve mais jsais pas ou aller apres

Jpense que je vais prendre le train en fraude sinon j'aurais plus de thune arriver là bas mdr

Faut qui me prête son passe navigo j'en ai bsoin yen a marre de froder

J'ai fraudé sans savoir/j'ai été obligé de frauder

Comme par hasard yavait les controleurs et javais pas le bon ticket 33euros azii seumer

Le controleur n'est pas venue ducoup je n'ai pas payer mon billet Temp pis là je suis chez moi !

Comportement des contrôleurs

Merci au controlleur de mon train d'avoir dormi tout le voyage me permettant de frauder sans problème #sncf

L'agent de la SNCF qui m'encourage à frauder le tgv c'est pas beau ça ?

Les controleurs a Colombes snt chaud lapin ! Le mec me fais " Jvs controle pas votre poitrine généreuse ma eblouis"

Autres

SNCF

Encore erreur d'affichage on retient pas les leçons chez vous @lignej a pontoise .

@RERC_SNCF la personne chargée d'ouvrir les accès a la gare d'Issy ne s'est pas réveillée ce matin. Ajoutons de la galère aux regards....

Expression de la durée

Vaguere le train il vien dans 15 min

Ca fait 15min jsuis ds le RER j'ai déjà 2controles

Tierce personne

Le bus de mon frère il est pas passer

@moha_ddict Quand je bossais sur Monac une collegue sans permis se prenais pas la tête. Taxi pour venir taffer.40euros par jours

Inclassables

Je suis arrivé a temps pour le train ouf

Elle s'arrêtait plus la prof un truc de dingue mais AMEN j'ai eu mon bus heureusement qu'il était en retard