Titre: Title:	Valorisation de l'information sur les marchés financiers par l'utilisation des mégadonnées
Auteur: Author:	William Sanger
Date:	2014
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Sanger, W. (2014). Valorisation de l'information sur les marchés financiers par l'utilisation des mégadonnées [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/1519/

UNIVERSITÉ DE MONTRÉAL

VALORISATION DE L'INFORMATION SUR LES MARCHÉS FINANCIERS PAR L'UTILISATION DES MÉGADONNÉES

WILLIAM SANGER

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION DU DIPLÔME

DE MAÎTRISE ÈS SCIENCES APPLIQUÉES

(GÉNIE INDUSTRIEL)

AOÛT 2014

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

VALORISATION DE L'INFORMATION SUR LES MARCHÉS FINANCIERS PAR L'UTILISATION DES MÉGADONNÉES

présenté par : SANGER William

en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées

a été dûment accepté par le jury d'examen constitué de :

Mme BERNARD Sophie, Ph.D., présidente

Mme DE MARCELLIS-WARIN Nathalie, Doct., membre et directeur de recherche

M. WARIN Thierry, Ph.D, membre et codirecteur de recherche

M. JOANIS Marcelin, Ph.D, membre

DÉDICACE

À ma grand-mère,
pour qui les voyages forment la jeunesse
et déforment les valises…
j'ai les pupilles remplies de ton amour,
ce qui me permettra d'entreprendre
les plus belles aventures.

REMERCIEMENTS

La recherche est une aventure intellectuelle passionnante. C'est un voyage initiatique sans commune mesure. Le cerveau explose et scintille. Ce mémoire couronne deux années de pur plaisir où les remises en question furent perpétuelles et formatrices. Il est le résultat de l'apport d'une multitude de personnes que je tiens maintenant à remercier. Pour qu'un enfant grandisse, il faut tout un village.

Nathalie de Marcellis-Warin et Thierry Warin, mes directeurs de recherche. Ils ont su me prendre sous leurs ailes pour m'épanouir à leur côté. J'ai découvert que les défis les plus fous pouvaient être relevés, et que les limites n'étaient faites que pour être repoussées. Je les remercie chaleureusement de tout cœur.

Le 414 Wiseman est devenu la maison où je me ressource de l'amour qui m'y ait prodigué. Sans cette énergie constamment renouvelée, aucune ligne de ce texte n'aurait pu voir le jour. Pour cela, Hélène Boisvert, Raphaelle Occhietti et Serge Occhietti ne sauront être assez remerciés. Un merci particulier à Raphaelle pour avoir relu l'ensemble de mon mémoire.

Gabrielle Crétot-Richert, pour partager mes joies et mes peines, puis pour réussir à me ramener sur terre par sa tendresse quand je divague trop.

Silvia Sugihara, qui m'appuie après chaque hésitation et me pousse à me surpasser.

Antoine Troadec et Bertrand Nembot pour leur accompagnement à travers cette aventure, puis dans celle de Fintech 2013 et pour les suivantes. Marine Hadengue, qui m'enseigne jour après jour comment concilier doctorat et développement personnel avec harmonie.

Et puis les amis, vous qui êtes là quand je m'enfonce trop dans les livres, mais qui répondez toujours présents lors d'un voyage ou pour un nouveau projet, et qui d'une tape sur le dos ou avec une tisane remontent le moral… Andrea Sandrine Matthieu Bénédicte Lad'a Ivka Mehdi Gaëtan Christophe Guillaume et Fred.

Finalement, Fintech Montréal, pour croire en Montréal et à l'esprit entrepreneurial de ses jeunes ; CIRANO, pour m'avoir fourni un cadre exceptionnel d'expression, pour croire en la relève scientifique du Québec et y former les ressources clefs de demain ; et Polytechnique Montréal, pour façonner avec créativité et audace les *geeks* décideurs qui sauront répondre aux prochains défis.

RÉSUMÉ

La finance 3.0 est encore dans ses balbutiements. Pourtant les mégadonnées représentent une opportunité sans précédent pour l'industrie. À travers cette recherche, nous caractérisons l'impact financier de deux types de messages publiés sur Twitter, notamment sur quatre types de rendements de compagnies inscrites au S&P500. Dans le cas des rendements journaliers, l'influence des messages financiers s'avère significative et peut être prise en compte dans les modèles prédictifs. En ce qui concerne les rendements nocturnes, les deux types de messages peuvent être utilisés (messages financiers et messages mentionnant les noms des compagnies), même si l'impact de ces messages n'est pas de même ampleur. Nous étudions aussi les rendements anormaux de ces compagnies, les volumes d'actions échangées et l'impact de la publication des rapports financiers. Au niveau méthodologique, l'utilisation de modèles économétriques probit a permis de mettre au point des tableaux de contrôle optimisant les opportunités de gain en fonction des jours de la semaine et des types d'industries visés. La deuxième partie de cette étude se penche sur l'analyse du contenu des messages financiers, et l'identification des utilisateurs du réseau. Les mesures de réputation traditionnelles s'avèrent inefficaces pour obtenir un signal nettoyé de tout bruit. À l'inverse, une approche par cartographie permet de mettre en lumière les nœuds sensibles du maillage des utilisateurs en identifiant les éléments clefs à suivre. Finalement, nous proposons une stratégie d'investissement ayant présenté des rendements supérieurs à l'indice de référence, puis nous concluons par des recommandations quant à l'utilisation des médias sociaux pour les investisseurs, les firmes inscrites en bourse, les organismes régulateurs et l'industrie financière en général.

ABSTRACT

Finance 3.0 is still in its early stages. However, using Big Data represents huge opportunities for the financial industry. In this study, we compare the influence of two kinds of message sent on Twitter (a microblogging social network) over different types of return concerning firms listed on the S&P500. For daily returns, one should consider financial tweet in particular, whereas for overnight returns, both financial texts and messages naming a company could be considered. We investigate the impact of these messages on abnormal returns, on exchange volumes and during the release of quarterly reports. Investment dashboards have been implemented following these findings, allowing one to optimize its gain opportunities depending on the investment day or the industry targetted. The second part of this study explores the content of financial tweets and the description of their senders. Typical reputation measurements could not provide enough insights due to the high level of noise in the data. However, by sketching the network of users, identifying key components was possible. Finally, we propose an efficient trading strategy outperforming the S&P500 index, and we conclude the study by recommandations concerning the use of social media for investors, firms, regulators and the financial industry in general.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ABNORMALRETURN	Rendement d'une compagnie par rapport à un indice boursier (S&P500)
CAC40	Indice boursier regroupant 40 capitalisations françaises à Paris
CAPM	Capital Asset Pricing Model
DAX	Indice boursier regroupant 30 capitalisations allemandes à Francfort
DJIA	Dow Jones Industrial Average
INDICENAME	Indice de 0 à 100 concernant les tweets mentionnant les noms des compagnies (Name)
INDICETICKER	Indice de 0 à 100 concernant les tweets financiers (Ticker)
INTRADAYRETURN	Rendement entre la valeur de fermeture d'une compagnie par rapport à la valeur d'ouverture du même jour
FTSE	Indice boursier regroupant les 100 capitalisations britanniques à Londres
FOLLOWER	Un utilisateur suivant les messages envoyé par un autre utilisateur
GNH	Facebook's Gross National Happiness
HASHTAG	Mot de référence sur Twitter caractérisé par l'utilisation du symbole « # »
NAME	Tweet mentionnant le nom d'une compagnie inscrite en bourse
MCO	Moindres carrés ordinaires
MEDAF	Modèle d'évaluation des actifs financiers
OVERNIGHTRETURN	Rendement d'une compagnie entre la valeur d'ouverture d'une journée et la valeur de fermeture de la veille
POMS	Profile of Mood States
R&D	Recherche et développement
RETWEET	Message retransmis sur Twitter

TICKER	Tweet financier (contenant l'identifiant d'une compagnie boursière précédé de *$*)
TWEET	Message de 140 caractères publié sur Twitter
TWEETER	Verbe (usage courant), le fait d'envoyer un message sur Twitter
R	Logiciel libre de traitement de données et d'analyses statistiques
S&P500	Indice boursier regroupant 500 capitalisations boursières à New York
SELFIE	Autoportrait réalisé à partir d'un téléphone cellulaire
STREAMR & TWITTER	Packages liés à R permettant d'accéder à Twitter
SVM	Support Vector Machines
VIX	Indice de volatilité du S&P500
VOLUMERETURN	Rapport entre le volume d'actions échangées d'une journée par rapport au volume d'actions échangées la veille
WEKA	Logiciel libre d'apprentissage automatique développé par l'université de Waikato (Nouvelle-Zélande)

LISTE DES ANNEXES

INTRODUCTION

Un long chemin a été parcouru en 16 ans. Alors qu'hier seules 24 millions de pages étaient cataloguées sur le site *google.stanford.edu*, aujourd'hui ce sont plus d'un milliard de requêtes quotidiennes qui sont effectuées sur ces mêmes serveurs (Brin & Page, 1998 ; Varian, 2014a). Un long chemin illustrant l'évolution exponentielle des technologies d'Internet dans nos sociétés. La quantité de données générée à l'heure actuelle n'a jamais été aussi importante : l'ensemble des connaissances humaines accumulées jusqu'en 2003 est dépassé par la production de 48 petites heures de données en 2014. À l'ère des mégadonnées (*Big Data*), tout est question de rapidité : le temps devient la richesse première.

Les mégadonnées sont caractérisées comme étant un ensemble de données structurées (texte, indicateurs financiers) ou non structurées (tweet, géolocalisation, photo, rythme cardiaque) produites en grande quantité. Elles sont générées en temps réel et sont la plupart du temps longitudinales (Warin et al. 2014). Alors que la capacité technique de les interpréter se démocratise, elles apparaissent comme les prémisses d'une nouvelle révolution radicale de processus, modifiant structurellement les approches industrielles. La finance est un domaine d'application particulièrement intéressant.

En remontant aux années 1950, on remarque que la finance moderne repose sur des hypothèses fortes desquelles découlent les modèles financiers actuels. Parmi celles-ci, soulignons-en deux. Tout d'abord, un investisseur est un être rationnel maximisant ses rendements, tout en minimisant son exposition aux risques (Markowitz, 1952). Vingt ans plus tard, Eugène Fama définit les niveaux d'efficience du marché quant à l'information disponible et son reflet dans les prix des actions (Fama, 1970). Il caractérise les marchés financiers selon trois niveaux d'efficience dans lesquels les opportunités de gain ne peuvent être obtenues car (1) l'information est disponible à tout individu et (2) tout nouvelle information est reflétée dans les prix des actions, et donc est absorbée par le marché.

Ces deux hypothèses sous-tendent que les marchés financiers ne sont pas prédictibles, notamment en décrivant les fluctuations comme associées à un mouvement aléatoire. Cette vision est néanmoins remise en cause, comme expliquée en détail par Malkiel (2003). L'auteur présente les travaux effectués soulevant les faiblesses du postulat de Fama. Entre autres, il souligne la présence de momentum dans les prix des actions et le fait que l'amplitude d'une réaction face à certaines

nouvelles n'est pas optimale (provoqué par un optimisme ou un pessimisme accru de la part des investisseurs). L'explication de tels phénomènes peut provenir de la finance comportementale. En effet, celle-ci exploite les émotions comme vecteur de changement décisionnel de la part des investisseurs, et présente donc les marchés financiers comme proxy de l'humeur générale (Nofsinger, 2005). Les bulles financières (octobre 1997, bulle Internet ou la crise financière américaine de 2007), autrefois considérées comme données aberrantes dans une série temporelle, sont incorporées dans un cadre et expliquées par ce qu'Akerlof et Shiller décrivent en 2009 sous le terme d' « esprits animaux », c'est-à-dire comment les émotions sous-tendent les mécanismes économiques et financiers.

Et les mégadonnées dans tout cela ? Avec 500 millions de gazouillis de 140 caractères publiés quotidiennement, le pouls de la population n'a jamais été aussi accessible et présent. Et Twitter n'est qu'une infime partie des données disponibles sur Internet. Le *Big Data* représente une des voies d'avenir pour la finance, tant pour la gestion des risques, qu'au niveau légal concernant les questions de vie privée, en passant par la possibilité de révéler de nouvelles relations entre les évènements (Tapiero, 2013). De plus, Tetlock et al. (2008) soulignent le fait que des données qualitatives (textuelles) concernant des firmes inscrites en bourse permettent de quantifier des caractéristiques fondamentales de ces firmes, et que le contenu informationnel n'est pas incorporé de manière optimale dans leurs prix. Twitter, par le biais des tweets publiés en ligne, semble s'apparenter à ce type d'information à caractère informel. Les interactions complexes entre utilisateurs et le volume de messages publiés contribuent au fait que les marchés financiers ne peuvent incorporer une telle somme d'information de manière efficiente.

L'utilisation des médias sociaux en finance a fait l'objet de nombreuses études au cours des dernières années, études reprises par des fonds d'investissements à saveur technologique. Toutefois, ces annonces médiatiques ne débouchent pas vers les résultats prometteurs des recherches scientifiques. Le problème pourrait être d'ordre méthodologique. En effet, la plupart des travaux reposent sur les techniques d'apprentissage automatique ou sur les techniques d'analyse de sentiment. Les études économétriques poussées sont encore rares dans la littérature.

Ici se situe notre problématique. Avec un tel flot de données, *comment valoriser cette information pour les marchés financiers ?* Bien qu'encore à ses balbutiements, la finance 3.0 offre une

opportunité de recherche sans précédent tant les répercussions sur les pratiques industrielles évolueront dans les prochaines années.

Ce mémoire de maîtrise se divise en quatre parties principales. Dans le premier chapitre, une revue de la littérature scientifique sera effectuée, notamment sur l'utilisation des données produites sur Internet dans le domaine financier, sur la structuration des données et sur les opportunités d'utilisation de ces dernières.

Au deuxième chapitre, nous étudierons en détail l'impact financier de deux types de messages publiés sur Twitter à travers les rendements journaliers et nocturnes des compagnies du S&P500.

Le troisième chapitre servira à caractériser les rendements anormaux des compagnies du S&P500, les volumes d'actions échangés, l'impact de nouvelles officielles, puis nous établirons une stratégie d'investissement basée sur ces résultats.

Finalement, le quatrième chapitre abordera les notions de réputation et d'influence des utilisateurs, puis nous cartographierons le réseau financier de Twitter afin d'en identifier les nœuds sensibles. Le mémoire se terminera sur des recommandations par rapport à l'utilisation des mégadonnées en finance.

CHAPITRE 1 REVUE DE LA LITTÉRATURE

Le premier chapitre de ce mémoire est consacré à établir la revue de littérature de notre sujet. Dans un premier temps, nous aborderons Internet comme ressource première pour la finance. Ensuite, nous étudierons les méthodes de structuration des données puis nous terminerons par présenter les études réalisées sur les opportunités des mégadonnées.

1.1 Internet, une source intarissable de données pour la finance

Cette partie de la revue de littérature concentre son attention sur trois périodes chronologiquement successives de la recherche académique. Ces périodes s'apparentent au rythme d'adoption et à la démocratisation des technologies Internet par les consommateurs. Ainsi, une revue exhaustive des travaux de recherche portant sur l'utilisation de la ressource Internet en finance sera présentée, en mettant d'abord l'accent sur les forums et les blogues Internet, puis sur les moteurs de recherche, et finalement sur les médias sociaux.

1.1.1 Forums et blogues internet

Au début des années 2000, les forums de discussions et blogues spécialisés ont commencé à émerger progressivement sur Internet. Parmi ceux-là figurent les sites de discussions dédiés à la finance, suivant les actualités quotidiennes. Deux aspects stratégiques de l'utilisation de cette nouvelle source d'information ont été entre autres étudiés, résumés par les questions suivantes : Doit-on prendre en compte les commentaires sur Internet ? Qui doit-on suivre en particulier ?

1.1.1.1 Performances boursières

En étudiant plus de 3000 compagnies cotées en bourse, Wysocki (1998) démontre pour la première fois l'importance des commentaires publiés sur *Yahoo! Finance*. Il décrit plusieurs caractéristiques des compagnies générant le volume de commentaires le plus important. Elles présentent une évaluation boursière élevée, de forts volumes d'échange, des rendements exceptionnels, une faible part d'investissement institutionnel et une activité de vente à découvert importante. Plus particulièrement, il trouve que lorsque le nombre de messages publiés double entre la fermeture des marchés et le lendemain, les rendements entre la fermeture de la veille et la fermeture du lendemain sont augmentés par 0,18%.

Grâce à une interface facilitant la prise de données, plusieurs groupes de recherche se sont concentrés sur un des premiers forums financiers, *RagingBull.com*. Tumarkin et Whitelaw (2001) étudient la relation entre des rendements anormaux et le nombre de messages publiés sur le site web. En utilisant des études d'évènements, ils montrent que l'impact de ces messages ne peut être anticipé à plus d'une journée d'avance. Néanmoins, des opinions positives et une activité de publication élevée sont corrélées positivement avec des rendements élevés. La même année, Antweiler et Frank (2004) analysent plus de 1,5 millions de messages publiés sur *Yahoo! Finance* et sur *RagingBull.com*. Les auteurs ont noté une corrélation entre les volumes échangés de 45 titres boursiers cotés au DJIA et le nombre de messages s'y référant, et dans une plus faible mesure entre la volatilité de ces titres et le nombre de messages envoyés. Ces deux relations s'observent pour une fréquence temporelle de l'ordre de la journée et moins. Ils en déduisent que l'ensemble de ces messages ne peut être considéré comme du bruit, et qu'ils véhiculent un contenu pouvant renfermer de l'information utile pour les investisseurs.

Un autre blogue fait l'objet d'études, notamment le blogue à caractère technologique *Engadget.com*. De Choudhury et al. (2008) réussissent à prédire dans 78% des cas l'amplitude des mouvements boursiers de compagnies technologiques et dans 87% des cas la direction de ces mouvements boursiers. Ils posent l'hypothèse que les mouvements boursiers peuvent être corrélés à l'actualité des compagnies sur les blogues, avec une anticipation allant cette fois-ci jusqu'à une semaine. Ils analysent cinq types d'informations publiés sur le site, notamment : le nombre de messages envoyés, le nombre de commentaires, la longueur des commentaires, la durée de réponse aux articles et la force des commentaires. Ils analysent à la fois des données non structurées (messages publiés sur le blogue) et des données structurées (financières). La direction des fluctuations boursières des compagnies est à chaque fois comparée à un indice de référence, dans ce cas-ci les fluctuations du NASDAQ. Ainsi, les rendements boursiers peuvent être analysés par rapport à ceux du marché, tel que :

$$y_t^c = \frac{\varphi_t - \varphi_{t-1}}{\varphi_{t-1}}$$

$$y_t^\eta = \frac{\psi_t - \psi_{t-1}}{\psi_{t-1}}$$

$$y_t = y_t^c - y_t^\eta$$

Avec : φ_t le taux de rendement d'une compagnie au jour t, ψ_t le taux de rendement de l'indice de référence (NASDAQ) au jour t et y_t la direction relative du mouvement boursier d'une compagnie.

Finalement, Gruhl et al. utilisent en 2005 le forum de discussion financière *HotCopper.com* afin de mesurer l'impact de la propagation de rumeurs sur les volumes d'actions échangées. Ils arrivent à la conclusion que les marchés boursiers considèrent les messages postés sur Internet, ce qui provoque une augmentation des volumes d'actions échangées anormalement élevés à la suite de messages publiés sur le site Internet.

L'ensemble de ces travaux de recherche arrive à la conclusion que les messages envoyés sur Internet contiennent une information pouvant être utilisée par et sur les marchés financiers. Ils se révèlent être une source supplémentaire d'informations pouvant adéquatement anticiper les mouvements bousiers. La question suivante complète les précédents travaux : parmi l'ensemble des messages publiés, y en a-t-il à écouter en particulier ?

1.1.1.2 Identification des faiseurs d'opinion

Buechel et al. (2012) s'intéressent à deux aspects de cette problématique. Tout d'abord, à quel point le *leadership* d'opinion repose sur la conformité des avis partagés par l'ensemble des individus et ensuite, comment une fausse représentation interfère avec l'humeur générale (concept aussi appelé sagesse des foules, ou « *wisdom of crowd* »). Ces deux aspects de recherche permettent d'aborder la notion de bruit afin d'éviter l'introduction de distorsions dans l'information fournie. Les auteurs arrivent à la conclusion que les personnes créant trop de bruit (c'est-à-dire ayant une présence trop important par rapport à leur capacité à fournir une information précise) devraient avoir des opinions plus conformes au groupe dans son ensemble ; à l'inverse, les acteurs ayant une information privilégiée devraient avoir tendance à éviter l'effet de groupe, et se démarquent par leurs opinions.

La définition des faiseurs d'opinion déteint ainsi sur la capacité des individus à avoir de l'influence sur le consensus de leur groupe. Un individu fortement conformiste aura tendance à rester fidèle aux opinions émises par le groupe auquel il appartient tandis qu'un faiseur d'opinion sera porté à exagérer sa position afin de contrebalancer le consensus admis par son groupe. Cette définition

rejoint celle de Brink et al. (2011) qui décrivent un faiseur d'opinion comme une personne pouvant influencer le comportement de ses pairs, mais aussi des personnes la suivant.

En étudiant les blogues Internet, Song et al. (2007) élaborent sur la problématique d'identification des influenceurs au sein d'un réseau. Ils proposent une méthode de catégorisation à partir d'une mesure d'influence appelée *InfluenceRank*. Cette mesure repose sur deux aspects : (1) la position de l'individu au sein d'un réseau et (2) le caractère inédit de l'information partagée. Ces deux aspects permettent de déterminer si un individu (ou un blogue) peut être considéré comme un faiseur d'opinion au sein de son réseau.

Cette position privilégiée est abordée dans les travaux de Domingos et de Richardson (2001). Les techniques de fouille de données leur permettent non seulement de déterminer si les individus sont susceptibles d'acheter ou non des produits, mais aussi d'identifier ceux dont le réseau sera le plus à même d'adopter une pratique similaire (et donc d'être influencé). Kempe et al. (2003) complètent cette approche en démontrant l'efficacité d'identifier les personnes les plus influentes afin de maximiser les retombées d'une stratégie marketing.

L'historique de publication des bloggeurs constitue une autre clef d'identification des influenceurs. Nakajima et al. (2005) décrivent deux types de contributeurs des blogues à fort potentiel d'influence : les *agitateurs* (« *agitators* ») comme étant les personnes stimulant les discussions et les *synthétiseurs* (« *summarizers* ») comme les acteurs résumant les échanges de messages. Les caractéristiques suivantes décrivent les deux types d'acteurs :

- Agitateurs : un nombre élevé de personnes les suivent et de messages les citent ; une incidence forte sur le volume de messages publiés à la suite de la publication d'un de leurs messages ; une forte similarité quant au contenu des messages publiés à la suite d'un de leurs messages.
- Synthétiseurs : un nombre élevé de messages cités à travers leur message.

En résumé, les travaux concernant les blogues Internet ont permis de déterminer que : (1) les messages publiés contenaient de l'information essentielle pouvant influencer ou mimer les comportements boursiers et (2), parmi l'ensemble de l'information mise en ligne, l'identification d'utilisateurs particuliers (les influenceurs) permettait d'accéder à un réseau étendu d'utilisateurs afin de véhiculer un message.

1.1.2 Moteurs de recherche

La littérature concernant l'utilisation des données des moteurs de recherche est très récente. Les difficultés techniques permettant l'accessibilité des données expliquent ce manque dans la littérature scientifique, les données étant une propriété des compagnies mettant à disposition leur moteur de recherche (Google, Yahoo!, Bing, Baidu, Yandex…). Même si plusieurs travaux émergent des départements de R&D de ces compagnies, la majorité des études académiques se penchent sur l'utilisation de *Google Trends* pour expliquer les comportements financiers.

Google Trends (*google.com/trends*) est un service offert par Google qui permet de visualiser et d'exporter les données agrégées des requêtes provenant des utilisateurs du moteur de recherche. Le caractère massif des données à disposition est indéniable, avec par exemple plus d'un milliard de requêtes envoyées quotidiennement aux serveurs de Google (Varian, 2014a).

Ces données générées en temps réel permettent de « prédire le présent » selon Choi et Varian (2012). Alors que certaines données macroéconomiques sont publiées à des fréquences allant de la semaine à l'année, l'utilisation par les auteurs des requêtes formulées à travers le moteur de recherche a permis de raffiner l'unité temporelle d'observation de l'ordre de la journée. Choi et Varian réussissent à expliquer qu'une augmentation de 1% du volume de recherche correspondant à une marque de voiture provoque l'augmentation de 0,5% des ventes de cette marque au cours du même mois. Cette méthodologie est déclinée à travers plusieurs autres secteurs industriels : ventes de voitures, ventes de maisons et ventes de billets sont expliquées par le volume de requêtes les concernant, et les statistiques mensuelles rapportant les résultats officiels peuvent ainsi être anticipés.

Deux avenues sont envisagées quant à l'utilisation des moteurs de recherche à des fins financières. La première, la plus fournie en recherche, concerne l'explication des volumes d'actions échangées par le volume de recherches effectuées auprès des moteurs de recherche. Le second pan de la littérature étudie la prédiction de rendements boursiers.

En utilisant les recherches effectuées sur *google.com*, Preis et al. (2010) notent la corrélation entre les volumes d'actions échangées des compagnies du S&P500 et les requêtes les concernant. Cette relation est vérifiée en utilisant des données agrégées à la semaine par *Google Trends*. Une augmentation du volume de requêtes concernant ces compagnies est corrélée avec une augmentation du volume d'échanges.

Dimpfl et Jank (2011) viennent confirmer les études de Preis et al.. Ils montrent que les volumes de requêtes sont corrélés avec les moments de forte volatilité sur les marchés. En analysant la volatilité du Dow Jones (New York), du CAC40 (Paris), du DAX (Francfort) et du FTSE (Londres), ils remarquent que l'ajout des données de Google Trends améliore leur modèles prédictifs : le logarithme des volumes de recherche contribue à hauteur de 9% à 23% quant à la variance du logarithme de la volatilité des indices boursiers. La relation entre volume de requêtes et volatilité n'est cependant pas à sens unique ; les deux variables s'influencent mutuellement, car « l'attention des investisseurs augmente lors de périodes de forts mouvements boursiers et vice-versa, une forte volatilité est provoquée par une augmentation de l'attention des investisseurs ».

Cette méthodologie semble robuste, Bank et al. (2010) proposent une étude axée sur les capitalisations boursières allemandes. Une augmentation du volume de recherche détectée à travers Google Trends est en relation avec une augmentation du volume d'actions échangées, ce qui en augmente la liquidité. Les auteurs expliquent ce phénomène par le fait que les asymétries d'informations sont réduites, les moteurs de recherche modélisant le comportement d'investisseurs non-experts. Leur approche considère les noms des compagnies au lieu des indices boursiers les identifiant (codes mnémoniques).

En utilisant le nombre brut de requêtes acheminées vers les serveurs de Yahoo!, Bordino et al. (2012) se concentrent sur le potentiel d'anticipation des fluctuations boursières des 100 plus importantes compagnies composant le NASDAQ (compagnies financières exclues). Les auteurs observent les volumes d'échanges de ces actions et les mettent en relation avec les volumes de requêtes effectuées quotidiennement.

Les résultats de leurs travaux permettent de mettre en évidence le potentiel prédictif des moteurs de recherche. En effet, les volumes de recherche sont corrélés avec les volumes d'échange des compagnies du NASDAQ, anticipant par une à trois journées d'avance les volumes d'échange. De plus, en observant le profil des utilisateurs, ils remarquent que la plupart ne s'informent que sur une compagnie par année. Cette distribution du nombre de compagnies recherchées permet d'identifier ces utilisateurs comme des utilisateurs naïfs (non-experts), permettant d'obtenir à partir des moteurs de recherche un certain pouls de la société, ou la sagesse des foules.

Par rapport aux rendements, une sélection plus restreinte de recherches est disponible.

Les périodes entourant les annonces officielles des compagnies provoquent d'importants mouvements boursiers. Da et al. (2011) utilisent *Google Trends* afin d'accéder à une information provenant directement des consommateurs dans le but d'anticiper ces fluctuations boursières. En analysant le nombre de requêtes de produits phares de 865 compagnies, ils trouvent que (1) les annonces de résultats financiers inattendus peuvent être anticipées avec les données de *Google Trends*, notamment les chiffres d'affaires et les rendements des actions lors des annonces de résultats trimestriels ; (2) les profits sont difficilement prévisibles ; (3) cette méthode d'anticipation ne marche pas avec tout type d'entreprise. Les entreprises à forte croissance, proposant une faible variété de produits et gérant adéquatement leurs bénéfices sont celles dont l'effet des annonces officielles peut être le mieux anticipé.

Cette approche se retrouve dans les travaux de Ramos et al. (2013) qui mettent en évidence qu'une augmentation des recherches auprès du moteur de recherche Google provoque une augmentation de la volatilité des titres boursiers considérés, une augmentation de leur volume d'échange et une diminution des rendements associés. Ils prennent en compte aussi des biais comportementaux des investisseurs, en considérant les maxima et les minima sur les 52 dernières semaines des titres boursiers : lorsque les maxima sont atteints, le pouvoir prédictif des données issues de *Google Trends* s'en retrouve renforcé ; a contrario, il diminue au moment des minima.

À l'image de Google avec *Google Trends*, le moteur de recherche chinois Baidu propose aux utilisateurs un service d'agrégation de données (*Baidu Index*). Ce service diffère de son équivalent américain, permettant d'obtenir directement les fréquences de recherche (sans les indexer par semaine pour *Google Trends*). Zhang et al. (2013) se sont intéressés au potentiel prédictif des recherches effectuées par les utilisateurs de Baidu en observant la relation entre rendements anormaux et fréquences de recherches concernant les compagnies enregistrées sur les marchés boursiers chinois (Main Board, ChiNext et SME Board). Leur modélisation se présente ainsi :

$$AR_t = \lambda + \beta_{TV} TV_t + \beta_{IA} IA_t + \varepsilon$$

Avec AR_t les rendements anormaux des compagnies au temps *t*, TV_t les volumes échangés au temps *t* et IA_t représentant l'attention des investisseurs au temps *t* (c'est-à-dire la fréquence de recherche fournie par *Baidu Index* sur les compagnies). En ajoutant la variable correspondant à l'attention des investisseurs, les auteurs améliorent l'efficacité de leur modèle prédictif par 26% par rapport à un modèle ne comportant que les volumes d'échange des actions.

Pour conclure, Mao et al. (2011) comparent plusieurs sources d'information provenant d'Internet : sondages, recherches effectuées sur Google, messages envoyés sur Twitter. En comparant l'efficacité des différentes sources d'informations, ils trouvent que (1) Google et Twitter peuvent servir d'outils de prédiction en finance et (2) la granulosité temporelle plus subtile de Twitter permet d'anticiper les fluctuations boursières avec une avance de une à deux journées, ce qui est impossible avec *Google Trends*. Les médias sociaux, de par leur instantanéité, se présentent alors comme une voie de prédilection afin de valoriser l'information pas encore prise en compte par les marchés financiers.

1.1.3 Médias sociaux et sites Internet (Facebook, Twitter, Wikipédia, eToro)

L'Internet 2.0 se caractérise par la production de contenu non plus uniquement de la part d'initiés, mais en provenance du grand public. Des sites tels que Wikipédia, Facebook, Imgur ou Twitter rendent extrêmement simple la mise en ligne d'informations sans connaissance particulière en programmation.

Cette section de revue de littérature se concentre donc sur la recherche effectuée sur ces sites et applications Internet, et leur utilisation dans le domaine financier. Elle sera divisée en quatre parties, la première se concentrera sur Facebook, réseau social global ; la seconde sur l'encyclopédie participative Wikipédia ; la troisième sur eToro, site de microbloggage à caractère financier ; et finalement la dernière partie étudiera en détail les travaux concernant Twitter, objet de ce mémoire.

1.1.3.1 Facebook

Très peu de recherches ont été effectuées sur le réseau social le plus important d'Internet (en 2014). Cette observation peut-être expliquée par le fait que Facebook réussit à protéger les informations de ses utilisateurs du grand public, car les utilisateurs décident du cercle de personnes ayant accès à leurs informations (contrastant avec le côté ouvert de Twitter par exemple). Seuls les travaux de Karabulut et de Kramer et al. ont été trouvés au moment de l'écriture.

Karabulut (2011) propose une mesure de sentiment basée sur la mise à jour des statuts personnels de 160 millions d'utilisateurs de Facebook, nommée *Facebook's Gross National Happiness* (GNH). Le GNH a la capacité de prédire les changements tant dans les taux de rendements que dans les volumes d'actions échangées des compagnies américaines inscrites en bourse.

Plus précisément, une déviation standard supplémentaire du GNH à un jour donné est corrélée avec une augmentation de 11,23 points des rendements boursiers le jour suivant. Ainsi, Karabulut obtient une mesure indirecte du sentiment des investisseurs. Le GNH est mis à jour quotidiennement par Facebook et se base sur la fréquence de mots positifs (ou négatifs) concernant les statuts mis en ligne, tel que :

$$GNH_t = \frac{\mu_t^p - \mu^p}{\sigma^p} - \frac{\mu_t^n - \mu^n}{\sigma^n}$$

où μ_t^p et μ_t^n représentent respectivement la fréquence journalière relative des mots positifs et négatifs dans les statuts des utilisateurs de Facebook ; σ^p (σ^n) et μ^p (μ^n) les déviations standards et les moyennes des fréquences journalières de mots positifs (ou négatifs) sur l'ensemble de la période étudiée.

Cette méthode est appliquée avec les marchés boursiers allemands et britanniques, puis considère deux types de rendements (prix de fermeture par rapport au prix d'ouverture du jour même ; prix de fermeture par rapport au pris de fermeture de la veille). Lorsque le GNH augmente d'une déviation standard, les marchés britanniques gagnent en moyenne 11,85 points de base tandis que les marchés allemands augmentent de 13,96 points de base.

En mars 2014 une étude menée conjointement entre les chercheurs de l'Université Cornell et de Facebook étudie les conditions de propagation des émotions au sein du réseau social. Kramer et al. ont mené une expérience auprès de 689 000 utilisateurs dans laquelle une partie de l'information publiée sur leur fil d'actualité a été omise. Les résultats de cette recherche ont montré qu'en retirant une partie des informations négatives accessibles à un utilisateur, celui-ci aura tendance à produire un contenu jugé positif. Vice-versa, en occultant une partie des informations positives, la production de contenu jugé négatif sera plus élevée. Ces résultats mettent en évidence la propagation des émotions auprès des réseaux d'utilisateurs.

Toutefois, cette recherche a provoqué de nombreuses réactions, tant au niveau académique que dans les sphères médiatiques. En effet, afin d'acquérir les données personnelles des utilisateurs, les protocoles d'éthique en recherche de l'Université Cornell n'ont pas été utilisés. Les chercheurs se sont contentés des décharges que les utilisateurs remplissent lors de leur inscription sur le réseau social. Cette utilisation des mégadonnées, et la modification du contenu proposé aux utilisateurs du réseau social pose plusieurs questions éthiques (Shroeder, 2014). Tout d'abord, cette pratique

pointe le rôle du réseau social dans sa gestion des données personnelles de ses utilisateurs. Ensuite, la recherche met en exergue le besoin de régulation dans la mise en place d'expérimentations liées à l'utilisation des mégadonnées, notamment en sciences sociales. Finalement, comme le souligne le Pr. Schroeder de l'Oxford Internet Institute, les techniques utilisées par les auteurs de cette recherche montrent l'imbrication des média sociaux dans la vie de tous les jours des internautes, mais surtout le fait que l'information proposée aux utilisateurs a été modifiée à grande échelle (plus de 700 000 utilisateurs ont vu leur fil d'actualité modifié au cours de cette expérience).

1.1.3.2 Wikipédia

Rubin et Rubin (2009) se penchent quant à eux sur l'encyclopédie participative en ligne Wikipédia. Chaque utilisateur d'Internet a la possibilité de contribuer aux articles publiés en modifiant les pages accessibles. Les auteurs supposent que la fréquence de mise à jour des pages Wikipédia des firmes inscrites en bourse peut être un proxy afin de mesurer le degré auquel la population est associée au traitement de l'information concernant ces firmes.

Leur supposition initiale est la suivante : plus la page Wikipédia d'une compagnie est éditée, plus le nombre d'individus ayant confiance dans les informations la concernant est élevé.

Ils posent et vérifient trois hypothèses. (1) Plus les informations sont mises à jour, moins les erreurs des analystes financiers sur ces compagnies sont importantes. (2) De plus, les mises à jour fréquentes sont corrélées avec une dispersion de prédiction plus faible. (3) Finalement, ils trouvent une corrélation avec les changements d'écart entre l'offre et la demande des prix des actions des compagnies lors des annonces officielles et la fréquence de mise à jour des pages des compagnies. Cette variable concernant les mises à jour des pages Wikipédia est mesurée à une fréquence mensuelle, entre juillet 2005 et décembre 2006 et concerne les entreprises du DJIA.

1.1.3.3 eToro

eToro est une plateforme d'achats et de ventes d'actions en ligne où les utilisateurs peuvent tisser des liens entre eux, notamment en mimant les échanges effectués. Cette plateforme permet de prendre des positions sur les marchés financiers, et offre la possibilité d'acheter ou de vendre à découvert. Comme décrit par Pan et al. (2012) eToro démocratise l'investissement boursier en le rendant « accessible et fun ». Quelques études sur le service Internet existent, menées par l'équipe du Media Lab du MIT.

Altshuler et al. (2012) ont mis au point un modèle de diffusion d'anomalies au sein des réseaux afin de détecter le seuil où l'information deviendra tendance (« *trending* »). Les auteurs ne se concentrent pas sur l'identification des noeuds les plus influents d'un réseau, mais plutôt sur la capacité de prédiction de viralité du contenu publié. Les échanges de plus d'un million et demi d'utilisateurs ont été analysés afin de déterminer les conditions pour qu'un élément devienne viral après avoir été partagé par au moins 5% des individus du réseau.

La seconde étude publiée porte sur le rôle des liens sociaux dans les mécanismes financiers. Selon Pan et al. (2010), l'influence sociale des individus joue un rôle déterminant quant à la surréaction du marché. La réputation des utilisateurs les plus renommés n'est pas due à la performance de leurs investissements, mais plutôt aux liens tissés entre les individus.

Deux questions d'étude émergent : (1) Est-il possible d'inférer des positions d'investissement à partir de la sagesse de la foule ? (2) De quelle manière l'influence au sein du réseau altère les dynamiques du groupe.

Les auteurs trouvent en premier lieu que l'ensemble des individus performe (retours sur investissements en moyenne positifs) mieux qu'un utilisateur isolé (retours sur investissements en moyenne négatifs). Ils notent aussi la présence d'influence sociale importante en lieu et place de pensée rationnelle, causée notamment par les incitatifs financiers rattachés à la plateforme. Finalement, l'influence des individus apparaît comme catalyseur de spéculations, menant à la provoquant une réaction disproportionnée du marché. Ainsi, les investisseurs sont plus prompts à adopter des comportements risqués lorsqu'ils suivent l'avis de leurs pairs.

1.1.3.4 Twitter

Mis en ligne en 2006, Twitter est devenu au fil des années un médium incontournable d'Internet. Présent lors des campagnes électorales, utilisé à la suite de désastres naturels ou afin de surveiller la propagation de maladies, la versatilité de cette plateforme de microblogage n'est plus à démontrer. En 2014, ce sont plus de 500 millions de messages qui sont envoyés quotidiennement entre 241 millions d'utilisateurs à travers le monde.

L'usage de Twitter est régi par une série de règles qui lui sont propres. Chaque message ne peut dépasser 140 caractères. Afin de référencer les messages envoyés, l'usage de mot-clef précédé du croisillon (« # ») est utilisé, appelé mot-dièse ou *hashtag*. Pour parler directement à certains

utilisateurs, il est nécessaire d'écrire leur nom de compte après le symbole « @ ». La géolocalisation des messages est possible, permettant de suivre avec précision l'origine des messages envoyés. Finalement, une convention concernant les messages à caractère financier existe : pour parler d'un cours boursier, il est nécessaire d'utiliser le symbole « $ » avant le code mnémonique des compagnies (pour Apple : $AAPL ; pour Google : $GOOG). Le caractère ouvert du réseau social a permis l'émergence de nombreuses publications scientifiques au sein de différents domaines, notamment en finance.

Les travaux de Bollen et de Mao apparaissent comme références dans la littérature scientifique. Après avoir montré que les marchés financiers pouvaient être anticipés par les moteurs de recherches puis plus efficacement par les réseaux sociaux comme Twitter (Mao et al., 2011), ils démontrèrent que les émotions liées aux messages publiés constituent un proxy précis pour prédire la direction des marchés. Grâce à leurs algorithmes d'analyse sémantique, ils classent les messages selon six types d'émotion et révèlent que les messages liés aux émotions du contrôle de soi (l'état d'esprit « calme » étant l'émotion associée à ce type de comportement) ont le plus d'incidence sur les résultats boursiers. Leurs prédictions peuvent anticiper les résultats boursiers du DJIA par quatre jours dans 86,7% des cas (Bollen et al., 2011). L'analyse de sentiment sera un thème abordé dans la section 3.1.3 de ce chapitre.

Ces travaux servent de point d'ancrage à plusieurs autres publications. Ainsi, Mittal et Goel (2011) effectuent une quasi-réplique de leurs travaux et obtiennent un pouvoir prédictif des performances boursières de 75,56%. Leurs données ont été collectées pendant les six derniers mois de 2009 et concernent l'indice du DJIA. Ils tentent de mettre en relation le sentiment associé à l'ensemble des messages publiés durant cette période de temps et la valeur de l'indice boursier. Quatre types de sentiment sont obtenus à partir de leur algorithme : joie (« Happy »), calme (« Calm »), alerte (« Alert ») et gentillesse (« Kind »).

De leurs recherches découlent plusieurs résultats. En premier lieu, les auteurs confirment les études de Bollen et de Mao stipulant que Twitter capture le sentiment général de la foule. De plus, deux types d'émotions, calme et joie, permettent d'anticiper par trois à quatre jours les résultats du DJIA. Finalement, ils réussissent à mettre en place un algorithme d'investissement boursier basé sur ces résultats, mais sous-performent les rendements de l'indice boursier par 50%.

En 2012, Brown se penche sur l'étude de la corrélation entre deux métriques reliées à Twitter (sentiment et volume de messages) et les performances boursières de compagnies (volume d'actions échangé et mouvements de prix). L'auteur présente les voies de recherche futures concernant l'utilisation de Twitter à des fins prédictives : réputation des utilisateurs et performances boursières, prise en compte des messages retransmis (retweets) dans la modification de la valeur du sentiment associé à un titre boursier, élargissement du nombre de compagnie considérées à l'ensemble du S&P500...

La même année, une recherche concernant uniquement la compagnie Apple a été menée par Smailovic et al.. Récoltant les messages financiers contenant la mention $AAPL », les auteurs montrent que la corrélation et la causalité entre le sentiment des messages et les performances boursières sont optimales pour une période de deux jours. La méthodologie de leur analyse de sentiment se base sur la classification par séparateurs à vaste marge (« Support Vector Machine », ou SVM) et permet de catégoriser les messages de manière positive ou négative.

Dans une étude détaillée publiée en 2010, Sprenger & Welpe fouillent le contenu de plus de 250 000 tweets sur une base quotidienne pour prédire plusieurs métriques boursières sur les 100 compagnies les plus échangées du S&P500. Ils prouvent que les messages concernant les cotations boursières contiennent de l'information apte à être utilisée mais qui ne se retrouve pas nécessairement dans les indicateurs du marché (même si cette information sera incorporée rapidement).

Ils soulignent la difficulté de suivre des utilisateurs en particulier afin de trouver les messages optimaux sur lesquels baser des décisions d'investissements. Par contre, les utilisateurs effectuant des investissements boursiers fructueux se voient accorder une part d'attention plus importante au sein du réseau social, notamment en gagnant en nombre de followers.

En analysant le sentiment associé aux messages, ils trouvent que l'optimisme est associé aux rendements anormaux, c'est-à-dire les rendements supérieurs à une déviation standard par rapport à la moyenne des rendements. Selon leurs recherches, les volumes d'échange peuvent être anticipés avec une journée d'avance : une augmentation de 1% des messages publiés est associée à une augmentation de 10% du volume échangé.

En se concentrant sur les évènements boursiers, Ruiz et al. (2012) tentent d'extraire deux types d'information à partir des messages : le premier concerne l'activité générale du réseau social, le

nombre de messages publiés et le nombre de messages retransmis ; le second type d'information est une approche graphique de retransmission des messages, mettant en valeur les noeuds centraux au sein du réseau des utilisateurs.

Cette fois-ci, l'échantillon de compagnies étudié est de 150 compagnies issues du S&P500. Les données ont quant à elles été récoltées quotidiennement au cours des six premiers mois de l'année 2010. Une plus forte corrélation (cinq fois plus importante que pour les rendements quotidiens) a été révélée entre le volume quotidien d'actions échangées et les messages publiés. Néanmoins, les résultats obtenus pour les rendements quotidiens ont pu être utilisés afin d'établir des stratégies d'investissements.

Les études suivantes tentent d'exploiter les informations contenues dans les messages publiés sur Twitter et possèdent le même schéma de recherche : collecte de données, structuration de l'information en analysant le sentiment y étant associé, mise en place de stratégies d'investissement. Par exemple, Chen & Lazer (2011) modifient l'approche d'analyse de sentiment de Mao et de Bollen en la simplifiant et réussissent à développer une méthodologie d'investissement battant le marché. Zhang et al. (2011) tirent de l'analyse des messages publiés quotidiennement un indice représentant ce qu'ils nomment « *collective hope and fear* ». Ils trouvent que ces émotions fortes sont positivement corrélées avec l'indice boursier VIX, modélisant la volatilité des marchés, et négativement corrélés au DJIA, au S&P500 et au NASDAQ. Finalement, Porshnev et al. (2013) poussent plus loin la démarche de Bollen et de Mao. En se concentrant sur un panel de messages 76 fois plus important, ils réussissent à prédire la direction du DJIA dans 70% des cas, la direction du NASDAQ dans 58,08% des cas et la direction du S&P500 dans 68,63% des cas. Ces résultats inférieurs à ceux des travaux de Bollen et de Mao peuvent être expliqués par le fait que la composante de prédiction de la recherche de référence ne s'est déroulée que sur une courte période de temps.

L'ensemble de ces travaux académiques permet de mettre en lumière le potentiel de Twitter à renfermer de l'information pertinente pour les marchés financiers. La diversité des méthodologies employées laisse néanmoins supposer un manque de méthode robuste unanimement adoptée par l'ensemble des groupes de recherche. Ainsi, plusieurs limitations sont à soulever.

L'échantillon de compagnies étudiées n'est jamais semblable, allant d'une compagnie à 150, en passant par les plus échangées au cours d'une période de temps. De par la composante

technologique reliée à l'obtention des données, la plupart des travaux de recherche proviennent des départements de sciences de l'informatique ; les techniques d'économétries poussées ne constituent pas une méthodologie adoptée par défaut dans la littérature, sauf pour les études de Sprenger et al. (2011) et de Porshnev et al. (2013). Finalement, une approche par secteur industriel manque à la littérature, ainsi qu'une comparaison entre messages financiers et messages normaux en utilisant un échantillon de données étendu. Ce mémoire de maîtrise tentera de répondre à ces zones d'ombre.

Fouille de données et analyse de sentiment ne sont que deux composantes des techniques qui permettent de valoriser de l'information non structurée. Le chapitre suivant se concentrera donc sur ces différents outils disponibles.

1.2 Techniques de valorisation de l'information

Trois approches seront abordées dans ce chapitre. (1) La première concerne les techniques informatiques, telle que l'acquisition et la fouille de données, puis l'analyse de sentiment. Nous aborderons ainsi les différentes méthodes permettant d'assigner une valeur à un message selon les types d'algorithmes utilisés. (2) La seconde approche met en relief l'emploi de cadres d'analyse issus de la physique : lois de puissance et ruptures structurelles. (3) La troisième et dernière approche se rapporte à l'utilisation de modèles financiers afin de structurer les données massives avec l'adaptation de la théorie moderne du portefeuille et du modèle d'évaluation des actifs financiers.

1.2.1 Techniques informatiques

Acquisition de données

Qualifié de réseau social « ouvert », Twitter rend public les données produites par ses utilisateurs. Plus important encore, la mise en place d'interface de programmation (« *Application Programming Interface* », ou API) rend possible l'acquisition systématique de ces données.

Plusieurs types d'interface sont disponibles aux programmeurs afin d'extraire l'information désirée. Le flot de données principal, *Firehose*, consiste en la totalité de l'information produite par les utilisateurs. Néanmoins, la seule façon d'y accéder est de passer par un service-tiers tel que *gnip.com* ou par des sites d'agrégation de données tel que *topsy.com* ou *peoplebrowsr.com*. Chacun comporte ses avantages et ses coûts dépendamment des services utilisés.

L'autre option offerte aux programmeurs est d'utiliser deux types d'interface de programmation, *REST API* et *Streaming API*. Le volume de messages publié étant considérable (500 millions de messages par jour), il est donc possible de connaître en temps réel ce qui se passe sur Internet, et plus globalement sur tout point du globe à tout moment (Bifet & Frank, 2010). Nous élaborerons plus en détail les différences entre ces deux types d'interface.

Après la création d'un compte de développeur (*dev.twitter.com*) et l'identification à travers le protocole de sécurité OAuth, les utilisateurs ont le choix d'utiliser deux types d'interface pour accéder automatiquement aux messages publiés sur Twitter.

(1) L'interface de programmation **REST API** peut être considérée comme une recherche dans la mémoire vive de Twitter. Seule une faible partie de l'information est disponible et l'utilisation est limitée par des contraintes fortes. Néanmoins, son utilisation reste simple grâce aux commandes de requêtes préétablies[1]. En d'autres termes, c'est une recherche dans l'historique des messages publiés, mais elle ne peut être utilisée pour remonter à des périodes plus lointaines qu'une semaine ou si le nombre de messages téléchargés est supérieur à 7000 (la première limite atteinte est celle qui arrêtera le téléchargement de données). Twitter fournit une description détaillée des limites auxquelles sont soumis les utilisateurs de ce type d'interface[2]. Des algorithmes de programmation sont disponibles pour le logiciel R réunis sous le package *twitteR* mis au point par Gentry (2013).

(2) La seconde interface de programmation permet d'obtenir des informations en temps réel (**Streaming API**). L'utilisateur se connecte au flot de données de Twitter, effectue une requête et le réseau social renvoie tous les messages correspondant à cette requête pendant une période de temps souhaitée[3]. Selon la documentation officielle, entre 1% et 40% des messages publiés sont disponibles. À nouveau, des algorithmes de programmation sont disponibles pour le logiciel R réunis sous le package streamR mis au point par Pablo Barbera (2014)

Cette seconde méthode d'acquisition des données offre une plus grande latitude pour l'analyse en temps réel de données, notamment par la création de boucles d'acquisition dans lesquelles sont

---

[1] https://dev.twitter.com/docs/api/1.1

[2] https://dev.twitter.com/docs/rate-limiting/1.1/limits

[3] https://dev.twitter.com/docs/api/streaming

insérées les lignes de codes nécessaires au traitement de l'information. Un des exemples d'application est TwitInfo, service mis en place au MIT afin de visualiser en temps réel l'information sur Twitter (Marcus et al., 2011). Le CIRANO a notamment suivi en direct les débats électoraux au Québec en avril 2014 afin de refléter la résonance des thèmes de campagne sur les médias sociaux (Warin et al., 2014). C'est cette seconde méthode qui sera utilisée afin d'acquérir les données nécessaires à la réalisation de ce mémoire.

Fouille de données

De par l'ampleur du nombre des messages mis en ligne, leur annotation ne devient possible que par l'assistance de systèmes informatiques. Les algorithmes de fouille de données (*datamining*) permettent de mettre en valeur les relations entre les données.

L'utilisation de techniques de fouilles de données permet de réduire le coût de traitement de l'information, d'augmenter les revenus engrangés et surtout de maintenir un niveau de suivi optimal et en temps réel (D. Zhang & Zhou, 2004). D'après les auteurs, les domaines d'application en finance des techniques de fouilles de données concernent la prédiction de prix des titres boursiers, la gestion de portefeuille, la prédiction de banqueroutes, le marché des changes et la détection de fraudes. Tous ces domaines d'application nécessitent le traitement de données massives, qu'elles soient structurées comme des données financières mises en ligne à chaque milliseconde, ou non structurées, comme des brèves financières, ou dans notre cas des tweets financiers.

Par exemple, Mittermayer (2004) a mis au point un système de classification de nouvelles financières sur lequel se base des algorithmes d'investissements boursiers. Son système, NewsCATS, a classé plus de 150 000 nouvelles financières en trois catégories (positive, négative, neutre) puis recommande par la suite l'achat ou la vente de titres financiers.

En 2002, Kloptchenko utilise des techniques de traitement automatique de l'information afin d'analyser rapidement les rapports financiers de trois compagnies de télécommunications (Nokia, Ericsson et Motorola). Ces techniques leur permettent d'interpréter les données tant qualitatives que quantitatives afin d'anticiper les performances financières futures. Ils remarquent que les données quantitatives (ratios financiers) ne reflètent que les performances passées des compagnies alors que les données qualitatives (tel que le ton sur lequel le texte est écrit) peuvent être révélatrices de l'état d'esprit de la compagnie, et donc de ses performances futures. L'analyse de sentiment que

nous aborderons dans la prochaine partie permettra d'optimiser cette approche de traitement de données qualitatives.

Varian (2014b) résume adéquatement le potentiel d'utilisation des techniques de fouille de données. La puissance de calcul aujourd'hui disponible permet d'extraire l'information encore cachée hier. Les outils technologiques destinés à manipuler les données massives sont multiples, et les bases de données sont coordonnées notamment à travers le format MySQL (ou NoSQL pour des bases de données de l'ordre des téraoctets). Plusieurs outils propriétaires ou libres existent, comme *Hadoop*, utilisé pour la parallélisation des calculs entre ordinateurs. La science des données s'occupe notamment de mettre en place des modèles prédictifs et de chercher les relations entre les données, entre autre en utilisant des arbres de régression.

De plus, la communauté des logiciels libres participe à l'implantation et l'adoption de systèmes d'analyse performants. À travers le site *r-bloggers.com*, ce sont plus de 400 sites qui mettent à jour des articles et des packages pour le logiciel de traitement statistique R. WEKA a été téléchargé plus de 1,4 million de fois (Hall et al., 2009), et consiste en un logiciel de traitement de l'information mettant en œuvre la plupart des techniques de fouille de données et d'apprentissage automatique de manière intuitive.

Cette puissance de calcul, combinée à des outils performants, permet de traiter massivement et rapidement l'information non structurée. Nous approfondirons une des techniques utilisées à travers la littérature scientifique reliée à la finance et à Twitter : l'analyse de sentiment.

Analyse de sentiment

Pouvoir analyser rapidement un texte et en extraire la polarité du sentiment associée a été l'objet de recherches poussées depuis les années 90. De nombreuses techniques ou algorithmes différents existent, tous possédant leurs avantages et leurs limites selon les textes étudiés. Twitter offre un nouvel espace de jeu pour les chercheurs. Des messages en quantité (presque) illimitée, accessibles, et de très courte longueur représentent un réel défi technique. Il a été démontré mathématiquement par Engle et Ng (1991) que les nouvelles positives sont liées à de fortes répercussions sur les prix et provoquent un impact à court terme seulement. À l'opposé, l'effet de nouvelles négatives tend à durer plus longtemps sur les prix et les volumes d'actions échangées (tiré de Devitt & Ahmad, 2007).

Ainsi, le besoin de méthodes efficaces de classement de l'information a été souligné dans la littérature scientifique (Das & Chen, 2001 ; Das & Chen, 2007). Sans être un recensement exhaustif sur le sujet, nous tenterons de décrire les différentes approches adoptées par les groupes de recherche.

Seo et al. (2002) déterminent 5 types de classification de texte alors utilisés à travers la littérature scientifique avant la démocratisation des médias sociaux, soit : (1) la classification naïve bayésienne (« *Naive Bayes* ») ; (2) l'algorithme de Winnow ; (3) l'utilisation de séparateurs à vaste marge (« *Support Vector Machines* », ou SVM) ; (4) l'algorithme du plus proche voisin (« *nearest neighbor classification* ») et (5) les modèles à entropie maximale. Les auteurs proposent par la suite une technique de classification de messages qui consiste à regrouper préalablement des nouvelles similaires pour en extraire des groupes de mots faisant référence à des concepts semblables. Les auteurs obtiennent ainsi des mots de référence relatifs à 5 classes de sentiment. Leur algorithme permet par la suite de classer adéquatement 79% des nouvelles financières émises. Nous verrons que plusieurs autres techniques ont été utilisées par la suite.

La ***classification naïve bayésienne*** (« *Naive Bayes* ») repose sur l'hypothèse (forte) qu'un texte ne peut appartenir à plusieurs catégories en même temps. En théorie, ce cas de figure ne se révèle pas complètement vrai, compte tenu de la complexité du langage et des niveaux de lectures pouvant y être associés. C'est notamment une des limitations de cette approche concernant l'analyse des tweets, par les double-sens pouvant être inscrits en 140 caractères. Cependant, en pratique, cette méthode se révèle efficace.

Go et al. (2009) testent trois types d'algorithmes en 2009 sur des tweets. Ils comparent trois algorithmes d'approche, soit (1) la méthode de classification naïve bayésienne, (2) la méthode par modèle à entropie maximale et (3) la méthode de séparateurs à vaste marge. Ces trois méthodes ont montré des résultats similaires, allant entre 80 et 83% de réussite quant à la prédiction de sentiment (positif ou négatif) par rapport aux messages publiés. Afin de réduire le bruit associé aux courts messages, les auteurs suggèrent de prendre en compte les émoticônes dans les catégories de référence.

Cette technique a été utilisée pour le traitement des tweets avec leur implication en finance, notamment par Antweiler & Frank (2004), Brown (2012), Go et al. (2009), Porshnev et al. (2013) et Sprenger & Welpe (2010).

L'approche par ***séparateurs à vastes marges*** (« *Support Vector Machines* », ou SVM) est une technique largement utilisée dans la littérature, notamment par son côté intuitif (Cortes & Vapnik, 1995; Joachims, 1998; Yang & Liu, 1999)).

Dans un premier temps, des lexiques de références sont bâtis par rapport à des concepts prédéfinis (positif et négatif dans la plupart des cas). Par la suite, on comptabilise le nombre de mots du texte étudié se référant aux concepts de référence pour en tirer une valeur de sentiment. Cette technique a été utilisée pour la classification de critiques cinématographiques (Pang et al., 2002), la classification de nouvelles financières (Mittermayer, 2004 ; Fung et al., 2002) et la classification de tweets (Smailovic et al., 2012 ; Go et al., 2009).

Une autre technique se base sur une méthodologie psychologique et tente d'associer à un texte six émotions distinctes : tension ; dépression ; rage ; vigueur ; fatigue et confusion. Cette technique utilisent ***les profils d'émotions*** (« *Profil of Mood States »*, ou POMS).

C'est l'approche utilisée par les travaux de Bollen et al. (2009), ils mettent en relation le sentiment lié aux messages publiés sur Twitter et l'impact sur les fluctuations boursières et sur les prix du pétrole. Ils arrivent à la conclusion que les évènements sociaux, politiques (élections présidentielles iraniennes de 2009), culturels et économiques ont un impact sur la modélisation des émotions présentes sur Twitter.

Leurs travaux de recherche devinrent la référence des études financières sur Twitter. Leurs expériences furent reprises par plusieurs groupes de recherches qui ne purent toutefois obtenir d'aussi bons résultats (Mittal & Goel, 2011 ; Porshnev et al., 2013).

Le bruit associé aux messages sur Twitter est un concept récurrent dans la littérature. Barbosa et Feng (2010) explorent une méthode utilisant une ***double classification*** prenant en compte le bruit inhérent aux messages téléchargés sur Twitter. Le fait que ces messages ne soient composés que de 140 caractères limite la portée de certaines techniques selon les auteurs (notamment la technique des n-grammes qui vise à inférer la probabilité de présence de lettres après une unité de texte sélectionnée). Afin de répondre à cette problématique, ils proposent de filtrer dans un premier temps les messages pour déterminer s'ils sont de nature objective ou subjective. Dans la seconde option, ils raffinent leur analyse pour détecter la polarité du message, c'est-à-dire le fait qu'il soit positif, négatif ou neutre. Ils basent leurs algorithmes en bâtissant leur échantillon test à partir de

trois sources de données différentes pour plus de robustesse dans leurs mesures (*twendz.com*, *twittersentiment.com*, *tweetfeel.com*).

Cette technique de double classification est utilisée dans les travaux de Pang & Lee (2004) et de Wilson et al. (2005).

1.2.2 Modèles physiques

Une autre méthode pouvant structurer des données est l'application de méthodes issues de la physique. La finance possédant une part d'aléas à travers ses processus de par l'implication d'êtres non rationnels (car humains!) et la physique moderne permet de quantifier ou de modéliser les phénomènes stochastiques. À la frontière entre sciences sociales et sciences pures, cette définition empruntant l'approche développée par Nofsinger est notamment abordée par Alex Pentland du MIT sous la dénomination de *Physique Sociale* (Pentland, 2014).

Des modèles physiques, deux approches seront abordées : les lois de puissance et les ruptures structurelles. L'avantage de ces méthodes est de faire *parler le bruit*, d'extraire le superflu du signal essentiel.

Les lois de puissance permettent d'expliquer la relation entre deux facteurs X et Y selon une relation du type $Y = kX^{\alpha}$, avec α étant l'exposant de puissance. En d'autres mots, quand X est multiplié par 2, alors Y est multiplié par un facteur 2^{α}.

Cette approche a été fortement utilisée en finance afin d'expliquer les bulles sur les marchés boursiers. Traditionnellement admis comme des données aberrantes, les effets des bulles financières peuvent être expliqués par le caractère exponentiel des lois de puissance.

Gabaix et al. (2003) proposent un modèle expliquant les fondements des lois de puissances se retrouvant dans les fluctuations des marchés boursiers. En effet, les fluctuations des actions inscrites en bourse semblent répondre à des comportements décrits par des lois de puissance, et ces lois se retrouvent à plusieurs échelles, autant sur de courtes périodes de temps que sur le long terme, et s'appliquent aussi bien à de petits marchés financiers qu'à des marchés financiers importants (Preis et al., 2011).

Les résultats des travaux empiriques (sans être exhaustifs) sur les lois de puissance au sein des marchés boursiers sont les suivants :

- Loi de puissance des rendements : la probabilité que le rendement d'un actif boursier soit plus élevée qu'une valeur *x* est fonction d'un facteur x^{-3}(Lux, 1996 ; Gopikrishnan et al., 1999)
- Loi de puissance des volumes échangés : la probabilité que la distribution des volumes échangés soit plus élevée qu'une valeur *y* est fonction d'un facteur $y^{-1,5}$ (Gopikrishnan et al., 2000)
- Loi de puissance du nombre de transactions : la probabilité que le nombre de transactions soit plus élevée qu'une valeur *z* est fonction d'un facteur $z^{-3,4}$ (Plerou et al., 2000)

Gabaix et al. (2003) soutiennent que les plus grandes fluctuations vont de paire avec les mouvements d'investissement des acteurs importants des marchés financiers, soit après les décisions des fonds d'investissement.

Une autre relation découlant des lois de puissance existe, notamment en ce qui concerne le prix des actions et leur demande respective. Cette dernière influence les variations de prix pour une période de 15 minutes avant et après la transaction des actions (donc pour une période de 30 minutes au total). Plerou et al. (2001) remarquent que les plus importants changements de prix des actions surviennent lorsque la demande est minimale.

Afin de pouvoir estimer et tester le comportement de données en fonction de lois de puissance, Gabaix (2008) propose une méthodologie utilisant l'estimateur de Hill dans un premier temps et ensuite une régression logarithmique. D'un point de vue pratique, afin de pouvoir quantifier les changements brusques de tendances dans les marchés financiers, Stanley et al. (2010) proposent d'observer les extrema locaux dans les séries temporelles associées aux cours de bourse.

Les auteurs définissent comme extremum local du prix d'une action sur une période Δt s'il n'y a pas de prix de l'action plus élevé à travers l'intervalle $t-\Delta t \leq t \leq t+\Delta t$. Inversement, les mimima locaux sur une période Δt sont considérés lorsqu'il n'y a pas de prix d'action plus bas sur le même intervalle.

Plusieurs lois empiriques ont ainsi été prouvées. L'approche universelle des lois de puissance se retrouve non seulement dans le domaine financier, mais aussi dans celui du commerce international, des régulations ou de la biologie.

Les ruptures structurelles sont des phénomènes qui apparaissent au sein de séries temporelles de données disjointes. Afin de rejoindre toutes les observations entre elles, la méthode la plus efficace n'est plus une droite de régression linéaire mais deux droites aux pentes distinctes. Ce phénomène doit être pris en compte, et fait l'objet de recherche économétriques et financières, notamment par le fait que les modèles prédictifs peuvent s'avérer erronés à la suite de ruptures structurelles (Timmermann, 2001).

Des études menées sur les marchés boursiers mondiaux mettent en lumière la robustesse de cette méthodologie. Moon & Yu (2010) utilisent ex-post les tests de rupture structurelle afin d'identifier la date exacte à laquelle un choc structurel est survenu sur les marchés boursiers chinois entre 1999 et 2007. Bahng (2004) identifie quant à lui trois bris structurels dans les rendements mensuels des marchés suisses entre 1988 et 2000 ; Chancharat et al. (2009) se concentrent sur les marchés thaïlandais. Ces derniers mettent en évidence le fait qu'après une rupture structurelle, les stratégies d'investissement surpassant le marché ne peuvent être mises en œuvre, notamment par l'impossibilité d'utiliser des données historiques pour de futures prédictions.

Quel est le lien entre ces modèles et les données massives non structurées (et Twitter en particulier) ? Il serait intéressant d'appliquer ces méthodologies aux séries de données recensant les nombres de messages publiés sur Twitter. En effet, plusieurs fonds d'investissements technologiques se sont basés sur les résultats prometteurs de Bollen et de Mao pour établir leurs stratégies. Après des résultats encourageants, leurs rendements ne se sont pas avérés aussi élevés que prévu. Une des explications peut être la démocratisation accrue de Twitter auprès de la population, entraînant des ruptures structurelles dans les séries de données, et de ce fait même, rendant les modèles prédictifs obsolètes.

1.2.3 Structuration des mégadonnées par les modèles financiers

Les techniques informatiques et les modèles physiques permettent de traiter d'une part une quantité massive de données de manière efficace, et de l'autre d'appréhender une composante aléatoire au sein des modèles. Un troisième type de méthodologie issu des modèles financiers peut être utilisé.

Ces modèles permettent de traiter de manière robuste une quantité presqu'infinie de données : les cotations boursières sont rendues publiques à chaque milliseconde. De plus, ils permettent de

quantifier et de comparer un niveau de risque associé aux titres boursiers. Ces deux caractéristiques s'avèrent cruciales pour le traitement des données non structurées.

À la lumière de l'impressionnante efficacité de l'industrie financière actuelle, il apparaît que ces modèles financiers sont robustes pour (1) traiter massivement de l'information en temps réel et (2) extraire une valeur de risque associée aux titres financiers. Warin & Sanger (2014) ont tenté d'utiliser ce cadre d'analyse pour structurer les données massives issues des messages de Twitter.

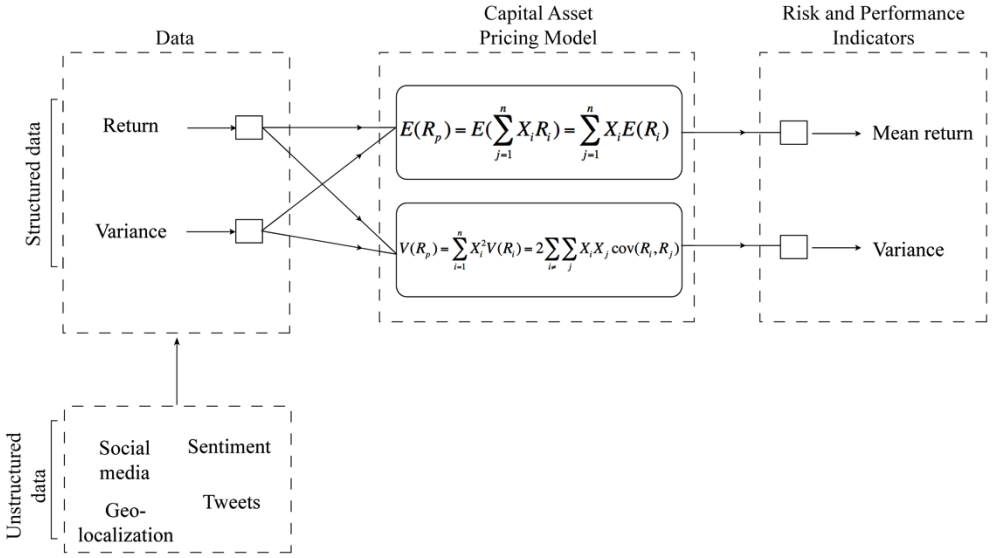

Figure 1.1: cadre d'analyse utilisant le CAPM pour le traitement de mégadonnées

La figure 1.1 représente le processus de structuration des données massives afin d'être utilisées dans le cadre d'analyse du MEDAF. Les auteurs ont ainsi cartographié le niveau de perception des risques de la population du Québec en 2012. Ils ont évalué neuf catégories de risques, notamment les risques reliés à l'économie, aux infrastructures, au système de la santé et à la gestion des projets publics. En se basant sur une méthodologie issue de l'ouvrage de de Marcellis-Warin & Peignier (2012), il est alors possible de modéliser à travers le temps (figure 1.2) le niveau de risque des thématiques étudiées. Ceci offre un nouveau niveau de lecture d'une notion difficilement perceptible, traditionnellement obtenue à la suite de sondages méthodologiquement lourds à réaliser.

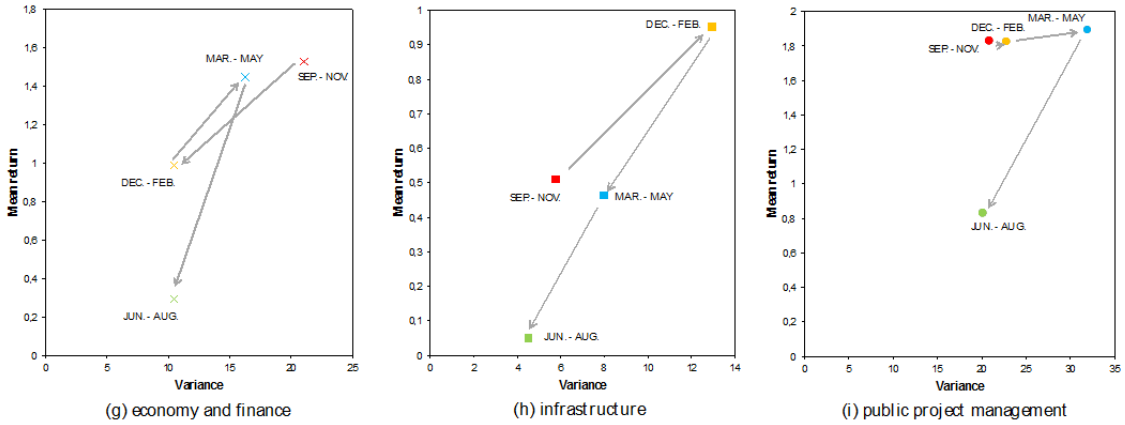

Figure 1.2 : évolution de la perception des niveaux de risques reliés à l'économie, reliés aux infrastructures et reliés à la gestion des projets publics

Ces modèles d'analyse repose sur les travaux fondateurs de la finance moderne. Cette dernière voit le jour dans les années 1950, avec notamment les fondements théoriques de Markowitz, puis sont adaptés une décennie plus tard à travers les travaux de Sharpe, de Lintner et de Mossin.

Markowitz (1952) élabore un modèle mathématique afin de prendre en compte les risques associés aux titres financiers. Ces derniers ne sont plus analysés séparément, mais plutôt au sein d'un portefeuille d'actions. Pour l'auteur, un investisseur se comporte de manière à maximiser ses rendements tout en minimisant la variance de son portefeuille : l'investisseur est un individu rationnel qui est averse au risque. Ce modèle repose donc sur des hypothèses fortes, notamment sur le fait que l'information doit être complète et disponible pour tout investisseur, puis que l'information est instantanément incorporée dans les prix du marché.

Finalement, ce modèle mathématique est par la suite transposé dans la pratique par Lintner (1965), par Mossin (1966), et par Sharpe (1963). Ils proposent d'appliquer les théories de Markowitz à un modèle pratique, le Modèle d'Évaluation Des Actifs Financiers (MEDAF). Les auteurs transposent les théories en évaluant ainsi les actifs financiers en termes de risques systémiques.

Afin de conclure cette section, méthodologies informatiques, physiques ou financières sont autant d'outils à la disposition des chercheurs pour dompter et structurer le flux de données que représentent les données massives. Efficacité, robustesse et prise en compte de l'aléatoire permettent de valoriser les données massives pour prendre en compte de nouveaux phénomènes économiques, et plus particulièrement en finance. La dernière partie de la revue de littérature traite

des opportunités de l'utilisation des données massives en finance, notamment en termes de réputation des entreprises ou des individus, et aussi afin d'établir des stratégies d'investissements boursiers.

1.3 Les opportunités des mégadonnées

Que faire avec autant de données ? Internet rompt toute notion de territorialité en rapprochant les individus et en proposant un contenu presqu'infini à portée de clic. Deux thématiques entourent l'utilisation des médias sociaux. En premier lieu, quel est l'impact des mégadonnées quant à la réputation des entreprises et des individus ? Ensuite, nous énumèrerons les différentes stratégies d'investissement mises en place utilisant les mégadonnées.

1.3.1 Réputation et influence

La réputation est un sujet de grande ampleur en sciences sociales et en gestion. Tenter de le résumer en quelques pages serait faire preuve de vanité. Nous ne nous y risquerons pas. Néanmoins, nous aborderons quelques articles faisant référence à la gestion de la réputation depuis la venue des médias sociaux, et l'implication sur l'influence des utilisateurs dans les réseaux.

Entre les années 1970 et 1980, la vision économique de la réputation apparaît comme un aspect stratégique indéniable permettant d'éviter l'entrée de nouveaux acteurs sur un marché. La présence d'asymétries d'information fait en sorte que les différents acteurs d'un même marché se basent sur les actions passées pour prendre leurs décisions, et anticipent celles des compétiteurs en fonction de leur réputation. Les travaux de Kreps & Wilson (1982), de Milgrom & Roberts (1982a, 1982b) et de Selten (1975) apparaissent comme fondateurs quant à la compréhension de la notion de réputation en adoptant un point de vue hérité de la théorie des jeux.

Quelques décennies plus tard, les frontières entre individus, consommateurs et investisseurs n'ont jamais été aussi ténues. Les firmes sont de plus en plus exposées à l'avis général : chaque utilisateur pouvant communiquer à sa guise, l'impact d'un simple message envoyé peut prendre des ampleur allant jusqu'aux répercussions boursières (Leavitt et al., 2009).

Warin et al. (2013) proposent un cadre d'analyse pour répondre au *buzz* auquel peut être soumise une compagnie. Lorsqu'une crise de réputation survient, l'impact sera mitigé par le niveau de réputation de la compagnie. Une compagnie jouissant d'une forte réputation verra l'impact sur son

cours de bourse plus faible que si elle n'avait pas investi dans sa réputation préalablement. En proposant un modèle mathématique d'interprétation de la réputation, les auteurs montrent qu'il est nécessaire de mettre en place des systèmes de communication efficaces afin de prévenir les crises de réputation, puis d'y répondre, limitant ainsi le temps de récupération face à celles-ci (voir figure 1.3).

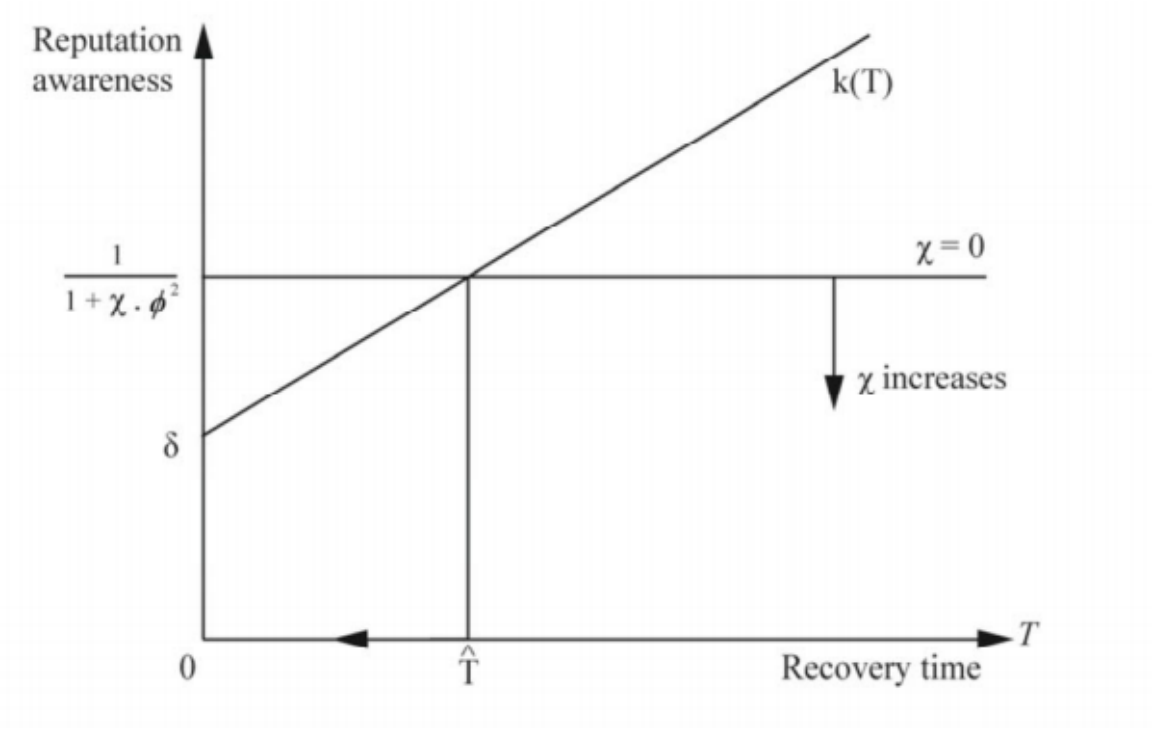

Figure 1.3: modélisation des conditions optimales pour répondre face à une crise, et répercussion sur la réputation des firmes

La nécessité d'identifier les influenceurs s'avère primordiale dans un monde aussi interconnecté que le nôtre. Cha et al. (2010) étudient les liens reliant les utilisateurs au sein des discussions publiées sur Twitter. Le groupe de recherche examine en particulier trois métriques, soit le nombre de followers, le nombre de fois où une personne est directement nommée et le nombre de fois où une personne voit son message retransmis (retweet).

Trois conclusions sont tirées depuis leur étude. Tout d'abord, le nombre de followers ne signifie pas nécessairement une facilité de retransmission de l'information. Les chanteurs populaires font partie des comptes d'utilisateurs les plus suivis, mais ils n'agissent pas comme relais de

l'information. Ensuite, les utilisateurs influents semblent acquérir une expertise reconnue au sein de plusieurs domaines. Leur dernier résultat concerne la manière de construire sa réputation sur Twitter : afin de gagner en crédibilité, un utilisateur se doit de se concentrer sur un sujet en particulier pour être reconnu expert dans son domaine.

Les retombées de cette recherche mettent en valeur la théorie de la propagation virale de l'information, propagation rendue possible non pas par des influenceurs traditionnels (nombre de followers élevé), mais par des personnes clefs au sein de certains réseaux. L'identification de ces personnes clefs reste une avenue de recherche prometteuse concernant les réseaux sociaux.

En mettant en valeur les lois de puissance dans les messages publiés sur Twitter, Weng et al. (2010) montrent que la forte réciprocité entre les relations des utilisateurs démontre un phénomène d'homophilie. Les individus tendent à suivre les utilisateurs leur ressemblant (cercle d'amis ou personnes partageant les mêmes idées). Cette notion permet de diviser en sous-réseaux les utilisateurs par domaine d'affinité, rendant ainsi l'identification de personnes clefs plus efficace.

Finalement, Bar-Haim et al. (2011) proposent une méthode d'identification d'experts financiers parmi les utilisateurs de Twitter. En utilisant la propension à prédire les cours de bourse, les auteurs soulignent l'importance de filtrer la masse de messages publiés afin de ne pas prendre en compte les avis des utilisateurs non experts participant au bruit.

1.3.2 Stratégies d'investissements technologiques

La dernière partie de cette revue de littérature se concentre sur les méthodes d'investissements tirées de l'utilisation des mégadonnées, et de Twitter en particulier.

La classification d'articles de presse en éléments modificateurs de tendance permet d'anticiper les fluctuations boursières (Fung et al., 2002). Ces auteurs ont mis sur pied un système de traitement de l'information permettant de classer tout nouvel article paru selon sa capacité à renverser une tendance boursière (voir figure 1.4). Lorsque leur système prédit qu'une action prendra de la valeur, alors ces actions sont immédiatement achetées, puis revendues soit au bout d'une heure, soit lorsque la valeur de l'action aura augmenté de 1%. À l'inverse, lorsque leur système prédit la baisse de valeur d'une action, celle-ci est vendue à découvert ; si la valeur de l'action est inférieure à 1% par rapport à la valeur de l'option, alors l'option est exercée, sinon elle le sera au bout d'une heure.

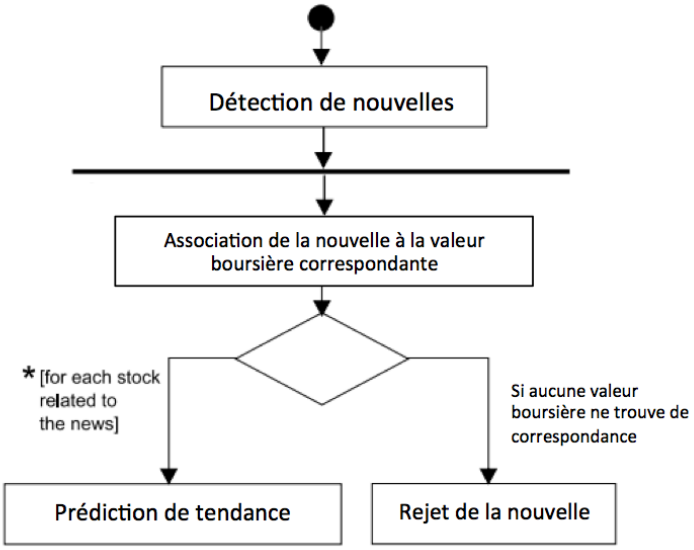

Figure 1.4 : schéma représentant le système de classification de nouvelles financières, adapté de Fung et al. 2002

De manière similaire, Schumaker & Chen (2010) ont mis au point un système de transactions automatiques basé sur la publication de nouvelles financières (figure 1.5). Lorsqu'une nouvelle est rendue publique, leur système achète (vend à découvert) puis vend (exerce l'option) au bout de vingt minutes lorsque l'action aura pris (perdu) de la valeur. Au bout de cinq semaines, leur système a réussi à obtenir un rendement de 8,50%, surpassant ainsi le S&P500 (+5,62 sur la même période).

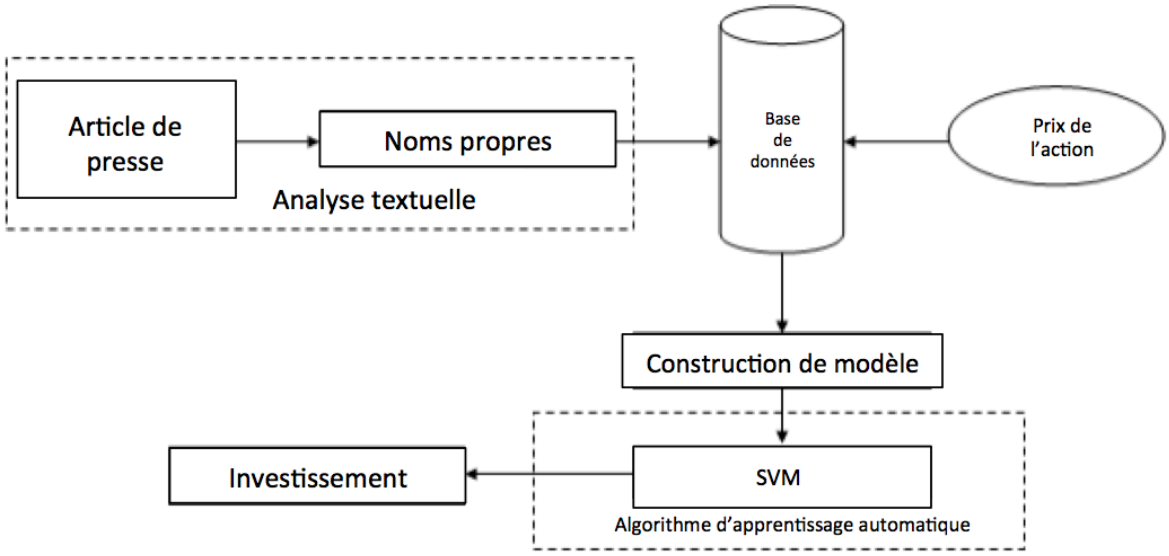

Figure 1.5 : schéma représentant le système de classification de l'information, adapté de Schumaker et Chen 2012

En se concentrant sur douze compagnies en particulier, Gidofalvi (2001) va plus loin en prouvant le pouvoir prédictif de certaines nouvelles, anticipant par vingt minutes les changements de tendances dans les prix des actions. Mittermayer (2004) propose quant à lui un algorithme d'investissement basé sur l'analyse sémantique des nouvelles financières (figure 1.6). Après avoir catégorisé le texte, son système effectue une recommandation d'achat puis revend ses positions après 58 minutes. Ses simulations montrent que le rendement moyen de chaque transaction s'élève à 0,11%, soit plus qu'un investissement aléatoire.

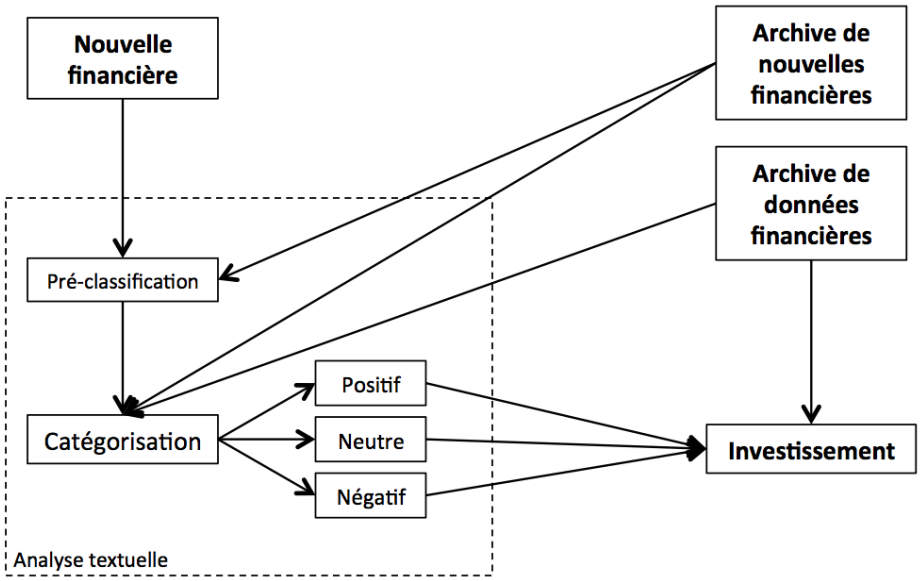

Figure 1.6 : schéma du système de traitement de l'information adapté de Mittermayer, 2004

En se basant sur le marché des devises, Vincent & Armstrong (2010) ont mis en évidence la présence d'une période où un message publié n'est pas pris en compte par le marché. Cette période dure entre deux et trois minutes pendant laquelle leurs algorithmes d'investissements peuvent profiter des asymétries d'information. Ils tentent de modéliser une valeur de volatilité associée au réseau social pour extraire des changements structurels dans les thèmes abordés. Lorsque survient une rupture structurelle, au lieu de continuer à échanger en bourse, leur algorithme cesse d'émettre des ordres afin de recalculer leurs modèles prédictifs. Sur une période de six mois, les rendements obtenus sont améliorés, passant de 0,56% à 1,27% par mois en prenant en compte les signaux obtenus par l'analyse de Twitter.

Zhang & Skiena (2010) utilisent aussi un algorithme prenant en compte les messages publiés sur Twitter. Ils classent les compagnies par valeur de sentiment, puis achètent les compagnies présentant des sentiments positifs à travers les messages publiés (achetant à découvert celles dont le sentiment s'y référant est négatif).

En simplifiant les précédents modèles, Chen et al. (2011) testent deux méthodes d'investissement. Lorsque leur système prédit que les actions prendront de la valeur, leur recommandation est d'investir la totalité de l'argent disponible puis de revendre le lendemain les actions. La seconde méthode d'investissement prend en compte la valeur prédite des actions : si le rendement de

l'action sera compris entre -0,1% et 0,05%, alors leur système investit 25% des liquidités disponibles ; si le rendement de l'action sera supérieur à 0,05%, alors leur système investit la totalité des liquidités disponibles ; si les rendements prévus sont inférieurs à -0,1%, alors il n'y aura pas d'investissements. Au cours d'une première simulation boursière, les stratégies d'investissement ont mené à des rendements de 3,19% et de 3,53% respectivement (le marché a effectué un rendement de 2,43%). Une seconde simulation boursière a permis d'obtenir des rendements de 5,32% avec la première méthode d'investissement et de 4,91% avec la seconde méthode d'investissement (contre 3,49% pour le marché).

Finalement, plusieurs fonds d'investissements technologiques se sont emparés des résultats des recherches académiques pour investir sur les marchés boursiers. Les résultats ne sont toutefois pas aussi prometteurs que le laisse suggérer l'avancée des recherches. En effet, Derwent Absolut Return, premier fonds d'investissement basé sur les messages publiés sur Twitter, a réussi à gérer 40 millions de dollars en 2012, obtenant un rendement de 1,86% en un mois. Néanmoins, ces résultats prometteurs au début, se sont avérés infondés (Mackintosh, 2012). Après fermeture et un *re-branding* du fonds en DCM Capital, la plateforme technologique a été vendue aux enchères pour seulement 186 000$ contre les 7,9 millions attendus (Malakian, 2013). D'autres fonds ont vu le jour, mais ont surtout évolué en fournisseurs de service, notamment en traitement de l'information sur Internet (MarketPsychData et Flyberry Capital).

1.4 Question de recherche et hypothèses

D'une vaste littérature scientifique, certes encore jeune mais variée, plusieurs questions de recherche émergent. Le but de ce mémoire est de se concentrer sur la valorisation de l'information pour les marchés financiers en utilisant les données issues des médias sociaux, et en particulier Twitter.

Afin de répondre à ce projet de recherche, plusieurs avenues sont envisagées.

D'abord, y a-t-il un type de message à prendre en compte en particulier, entre des messages mentionnant les noms de compagnies ou des messages financiers ? De quelle manière ces types de messages influencent-ils les performances boursières des compagnies citées ?

(H1) **Nous posons l'hypothèse que les messages financiers sont à privilégier afin de pouvoir anticiper les rendements des compagnies en se basant sur le volume de messages émis.**

De plus, comment utiliser ces résultats afin d'investir en bourse ?

(H2) **Nous posons l'hypothèse qu'il est possible de mettre sur pied des tableaux de bord afin de guider les investissements boursiers et ainsi maximiser les opportunités de gains**.

Une seconde avenue de recherche concerne les utilisateurs de Twitter. Avec 30000 messages financiers envoyés quotidiennement, peut-on se contenter de suivre certains utilisateurs en particulier ?

(H3) **L'approche utilisant le nombre de followers montre ses limites, nous posons alors l'hypothèse qu'une approche par réseau serait plus adaptée.**

CHAPITRE 2 S&P500 ET TWITTER… IMPACT DE 140 CARACTÈRES

« Breaking : Two Explosions in the White House and Barack Obama is injured »
_The Associated Press, 10h07, 23 avril 2013.

Soixante-douze caractères plus tard, le Dow Jones plonge, le S&P500 perd près de 121 milliards de dollars. Puis remonte après la confirmation que le compte de l'Associated Press ait bien été piraté. Soit un coût unitaire de 1,68 milliard de dollars par caractère écrit.

Cet évènement, bien qu'isolé, révèle l'imbrication des médias sociaux dans la sphère financière, et surtout leur influence par rapport au risque systémique des marchés. Jamais séparation entre attention publique, performances boursières et investissements n'a été plus fine.

L'objet de ce chapitre n'est pas de savoir comment utiliser spécifiquement l'information contenue dans un tweet, ni même d'étudier l'impact d'un message en particulier. Son objectif réside plutôt dans la compréhension et la caractérisation de la relation entre l'ensemble des messages envoyés sur Twitter et les performances boursières des compagnies citées. Avec un horizon temporel plus étendu (de l'ordre de la journée), nous explorons de quelle manière peuvent être anticipées les mesures de performances boursières avec les données non structurées, ou tout du moins comment ce type de données constitue (ou non) un apport additionnel d'information pertinente.

2.1 Méthodologie

Nous utilisons conjointement deux types de données afin de vérifier l'hypothèse deux (H2) de notre recherche, soit des données financières et des données non structurées (nombre de messages publiés par jour sur Twitter). De par l'étendue de la base de données utilisée (cinq cents compagnies avec des données quotidiennes sur une année), la méthodologie employée se base sur l'économétrie de panel. Après avoir récolté les données correspondant aux cinq cents compagnies du S&P500 au cours d'une année, nous raffinons notre échantillon à un ensemble de soixante et une compagnies. Ces compagnies ont été sélectionnées de manière à ne garder que les compagnies dont le nombre moyen de tweets financiers est d'au moins trente par jour.

Les régressions par la méthode des moindres carrés ordinaires (MCO) et les modèles probit constitueront les outils afin de caractériser l'influence des messages de Twitter sur les variables financières envisagées (quatre indicateurs de performance). Ces indicateurs correspondent à quatre types de rendements : (1) journalier, soit le rendement entre la valeur d'ouverture et la valeur de fermeture d'une même journée (*intraday return*) ; (2) nocturne, soit le rendement entre la valeur d'ouverture d'une journée et la valeur de fermeture de la veille (*overnight* return) ; (3) un rendement concernant les volumes d'échange des actions d'une compagnie (*volume return*) ; (4) anormaux, soit le rendement quotidien d'une compagnie par rapport à celui du S&P500 (*abnormal return*). Les modèles probit permettront de mesurer l'impact de l'augmentation d'une variable (nombre de tweets) sur la probabilité que la variable observée (rendements, alors rapportés sous la forme de variables binaires) puisse passer d'un état A à un état B, en l'occurrence être un rendement positif.

Dans une démarche exploratoire, les régressions par la méthode des MCO ne seront employées que pour comparer l'influence respectivement des deux types de tweets sur les quatre rendements étudiés (signe des relations et significativité des variables). Nous cherchons ainsi à différencier ces deux types de données non structurées (H1). Les modèles probit serviront à étudier quantitativement l'influence des variables relatives aux tweets sur les différents rendements (H2).

Ce second chapitre sera consacré à l'étude détaillée des deux premiers rendements (journalier et nocturne). Quant aux deux autres, les résultats seront résumés dans le chapitre 3 de ce mémoire avec l'étude de stratégies d'investissements employant les médias sociaux et l'impact d'annonces officielles.

Pour les modèles probit, cinq seuils de rendement seront testés, correspondant aux intervalles suivants :

- $1^{er}\ seuil = \begin{cases} 1\ si\ rendement\ positif \\ 0\ sinon \end{cases}$
- $2^{\text{è}}\ seuil = \begin{cases} 1\ si\ 0\% < rendement \leq 1\% \\ 0\ sinon \end{cases}$
- $3^{\text{è}}\ seuil = \begin{cases} 1\ si\ \ 1\% < rendement \leq 5\% \\ 0\ sinon \end{cases}$
- $4^{\text{è}}\ seuil = \begin{cases} 1\ si\ \ 5\% < rendement \leq 10\% \\ 0\ sinon \end{cases}$
- $5^{\text{è}}\ seuil = \begin{cases} 1\ si\ rendement > 10\% \\ 0\ sinon \end{cases}$

L'utilisation des différents seuils, et plus généralement des modèles probit au profit d'une méthodologie par MCO se justifie par l'interprétation des résultats. Un investisseur pourrait

rechercher des placements dont les rendements seront compris entre 1 et 5%. Afin de faciliter la mise en pratique des résultats économétriques, la méthodologie probit permet de d'identifier rapidement les résultats en observant les effets marginaux des variables observées.

De plus, deux types de variables binaires sont ajoutés aux différents modèles afin de prendre en compte les effets fixes. Ces deux variables binaires correspondent aux jours de la semaine (lundi au vendredi) pendant lesquels les actions sont échangées en bourse, les types d'industries dans lesquelles oeuvrent les compagnies du S&P500 et si les compagnies sont ou non dans une semaine d'annonces de résultats officiels. Pour les variables binaires dédiées aux jours de la semaine, la valeur du mardi a été omise pour être utilisée comme pivot. De la même manière, la valeur de l'industrie dédiée aux technologies de l'information sera le pivot des analyses. L'effet des annonces officielles sera étudié au chapitre quatre.

Finalement, nous avons contrôlé le décalage temporel de nos modèles en introduisant un décalage temporel d'une journée au niveau de nos variables correspondant au nombre de tweets. Le détail des variables utilisées est explicité dans la partie suivante.

L'ensemble des tests pour chaque variable sera effectué avec la méthodologie suivante :

- Modèles MCO (moindres carrés ordinaires) pour les valeurs absolues de rendement afin d'étudier la corrélation entre les variables et l'intensité des variables étudiées
- Modèles MCO pour les valeurs absolues de rendement avec décalage temporel de zéro à quatre jours
- Modèles MCO pour les valeurs positives de rendement
- Modèles probit avec seuils et variables de contrôles
- Modèles probit avec décalage temporel d'une journée
- Modèles probit avec interactions de variable

2.2 Données

Les données financières proviennent du site Yahoo! Finance. Elles correspondent au cours d'ouverture, au cours de fermeture et au volume d'action échangé par compagnie pour chaque jour entre le 1er mai 2012 et le 1er mai 2013 (soit 251 jours).

Les données non structurées sont obtenues grâce au site *People Browsr*[4]. Chaque compagnie fait l'objet de deux recherches, la première par rapport au nombre de fois où son nom est mentionné par jour (i.e. « netflix » pour Netflix) et la seconde par rapport au nombre de fois où le code mnémonique de la compagnie est mentionné (i.e. « $NFLX » pour Netflix). Cette dernière recherche utilise notamment le préfixe « $ », conventionnellement adopté sur Twitter afin de parler de compagnies inscrites en bourse. Ainsi, le premier type de recherche permet d'obtenir les taux de mention de la compagnie par un public « naïf » (sans connaissance particulière en finance) tandis que le second type de recherche archive les résultats produits par un échantillon de personnes dites « expertes ». Ces deux types de recherche seront nommées respectivement *Name* et *Ticker* dans la suite de ce mémoire. La liste complète des termes utilisés pour effectuer les requêtes sur l'agrégateur de données se trouve à l'annexe A. À nouveau, les données sont récoltées pour une période s'étalant entre le 1er mai 2012 et le 1er mai 2013.

Les trois autres variables binaires incorporées dans nos modèles concernent les types d'entreprise, les jours de la semaine et le fait d'être ou non dans une semaine d'annonce officielle :

- **Type d'entreprises** : le S&P500 est composé de dix industries distinctes. Chaque industrie fait l'objet d'une variable binaire (vente au détail, consommation de base, énergie, finance, santé, industrie, matériaux, télécommunication, services publiques, technologies de l'information)
- **Jour de la semaine** : chaque jour de la semaine fait l'objet d'une variable binaire, du lundi au vendredi
- **Semaine d'annonce officielle** : lorsque la compagnie annonce des résultats trimestriels ou annuels, cette variable prend la valeur 1 pour l'ensemble de la semaine pendant laquelle ces annonces sont prévues ; nous avons aussi ajouté une variable binaire concernant la semaine précédant les annonces officielles

Avant d'étudier plus en détail la structure des différentes variables, il est important de noter que la constitution du S&P500 représente de manière quasi-similaire chaque secteur industriel (sauf le secteur des télécommunications, voir figure 2.1).

[4] *www.gr.peoplebrowsr.com*

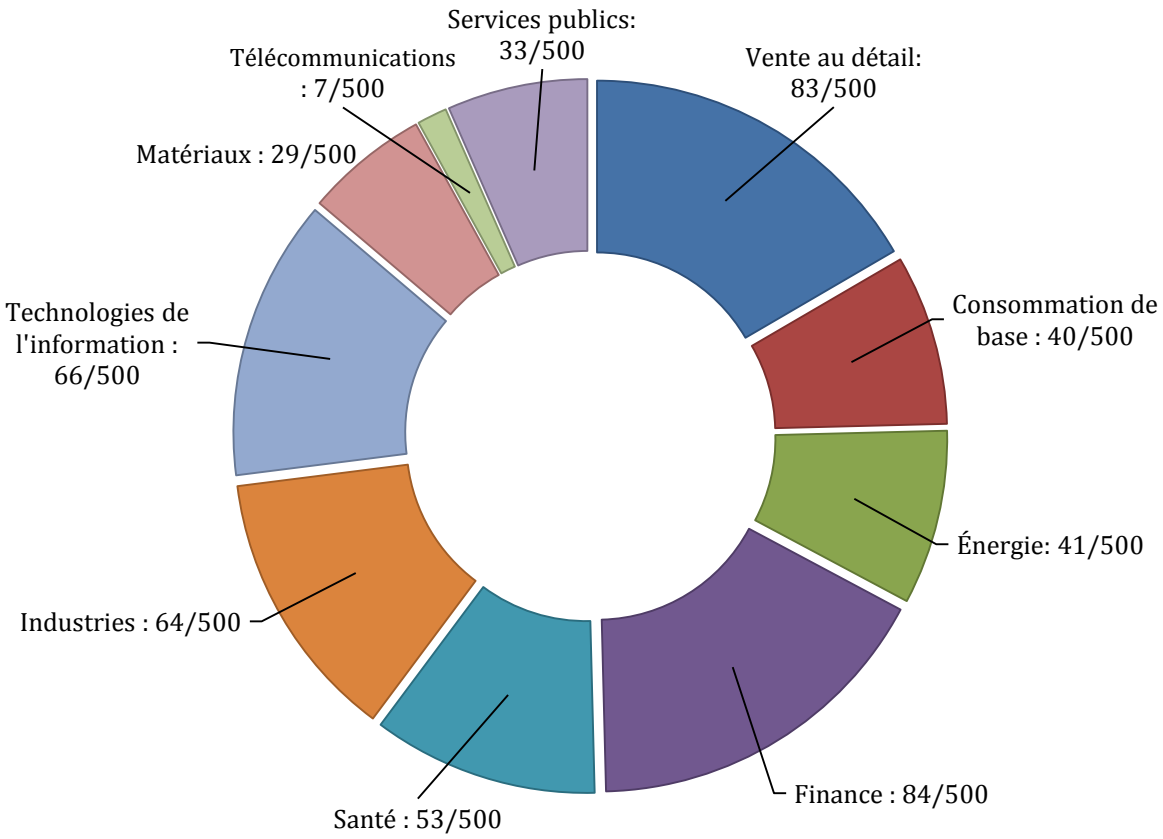

Figure 2.1 : répartition par industries des compagnies du S&P500

Cette répartition n'est toutefois pas semblable quand on la compare avec le nombre de tweets émis sur les compagnies. En se concentrant sur les messages de type *Name*, soit concernant le nom des compagnies, le secteur des technologies de l'information est prépondérant, avec 51% des messages publiés, puis par la suite le secteur de la vente au détail (31%) et de la consommation de base (11%) (voir figure 2.2).

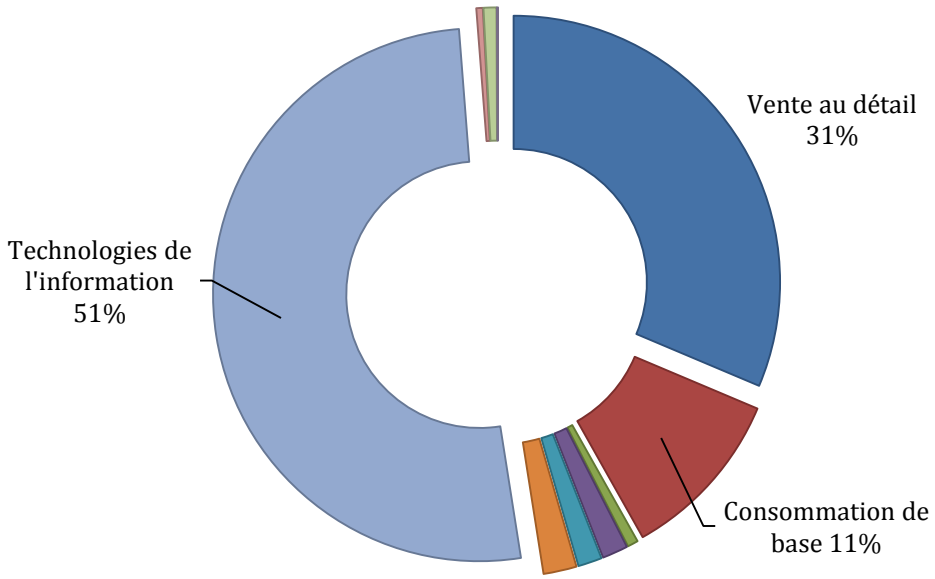

Figure 2.2 : répartition par secteur du nombre de tweets mentionnant le nom des compagnies (*Name*)

Cette répartition n'est toujours pas similaire si l'on considère le second type de messages publiés, soit les tweets *Ticker*. 64% des messages envoyés concernent des compagnies du secteur de la vente au détail, quand 17% concernent le secteur des technologies de l'information (voir figure 2.3).

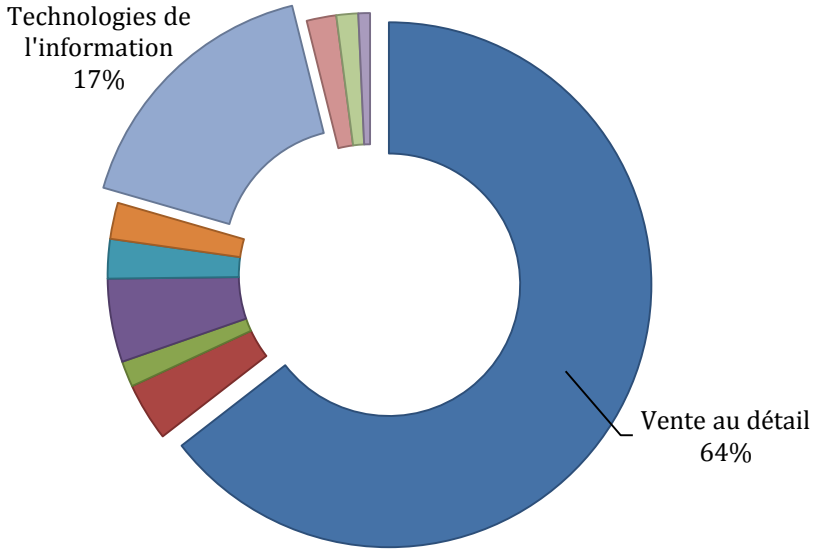

Figure 2.3 : répartition par secteur du nombre de tweets financiers (Ticker)

2.2.1 Échantillon sélectionné

Sur les cinq cents compagnies sélectionnées, nous décidons de restreindre notre échantillon aux compagnies mentionnées en moyenne au moins trente fois par jour sur Twitter (variable Ticker). Soixante-et-onze compagnies constitueront donc notre étude. Les requêtes utilisées pour ces compagnies sont résumées au tableau 2.1.

Tableau 2.1 : liste des compagnies sélectionnées et des requêtes utilisées

ID	Name	Ticker	Industrie	ID	Name	Ticker	Industrie
1	agilent	$A	Santé	37	jpmorgan	$JPM	Finance
2	alcoa	$AA	Matériaux	38	kellogg	$K	Consommation de base
3	amazon	$AMZN	Vente au détail	39	eli lilly	$LLY	Santé
4	american intl group	$AIG	Finance	40	loews	$L	Finance
5	apple	$AAPL	Technologies de l'information	41	lowe's cos	$LOW	Vente au détail
6	at&t	$T	Télécommunication	42	macy's	$M	Vente au détail
7	bank of america	$BAC	Finance	43	mastercard	$MA	Technologies de l'information
8	bestbuy	$BBY	Vente au détail	44	mcdonald	$MCD	Vente au détail
9	boeing	$BA	Industrials	45	merck	$MRK	Santé
10	caterpillar	$CAT	Industrials	46	microsoft	$MSFT	Technologies de l'information
11	celgene	$CELG	Santé	47	monster beverage	$MNST	Consommation de base
12	chesapeake	$CHK	Énergie	48	morgan stanley	$MS	Finance
13	chevron	$CVX	Énergie	49	netflix	$NFLX	Technologies de l'information
14	chipotle mexican	$CMG	Vente au détail	50	nike	$NKE	Vente au détail
15	cisco	$CSCO	Technologies de l'information	51	nvidia	$NVDA	Technologies de l'information
16	citigroup	$C	Finance	52	oracle	$ORCL	Technologies de l'information
17	cliffs natural	$CLF	Matériaux	53	penney	$JCP	Vente au détail
18	coach inc	$COH	Vente au détail	54	pfizer	$PFE	Santé
19	coca cola	$KO	Consommation de base	55	priceline.com	$PCLN	Vente au détail
20	dell	$DELL	Technologies de l'information	56	procter gamble	$PG	Consommation de base
21	dominion resources	$D	Services publics	57	qualcomm	$QCOM	Technologies de l'information
22	dow chemical	$DOW	Matériaux	58	ryder system	$R	Industrials
23	ebay	$EBAY	Technologies de l'information	59	salesforce	$CRM	Technologies de l'information
24	exxon	$XOM	Énergie	60	starbucks	$SBUX	Vente au détail
25	first solar	$FLSR	Industrials	61	target corp	$TGT	Vente au détail
26	ford	$F	Vente au détail	62	teco Énergie	$TE	Services publics
27	freeport mcmoran	$FCX	Matériaux	63	tripadvisor	$TRIP	Vente au détail
28	general electric	$GE	Industrials	64	united states steel corp	$X	Matériaux
29	general motors	$GM	Vente au détail	65	verizon	$VZ	Télécommunication
30	goldman sachs	$GS	Finance	66	visan inc	$V	Technologies de l'information
31	google	$GOOG	Technologies de l'information	67	walmart	$WMT	Consommation de base
32	hewlett	$HPQ	Technologies de l'information	68	disney company	$DIS	Vente au détail
33	home depot	$HD	Vente au détail	69	wells fargo	$WFC	Finance
34	intel	$INTC	Technologies de l'information	70	yahoo	$YHOO	Technologies de l'information
35	ibm	$IBM	Technologies de l'information	71	yum!	$YUM	Vente au détail
36	johnson & johnson	$JNJ	Santé				

2.2.2 Variables étudiées

Au cours de ce mémoire, nous tentons d'expliquer quatre variables financières pour étudier l'influence des médias sociaux sur les marchés financiers. Ces variables sont les suivantes :

- **Rendement journalier**, (chapitre deux) définit comme étant le rendement au cours d'une journée entre le cours de fermeture et le cours d'ouverture, tel que :

$$Rendements\ journaliers_t = \frac{Cours\ de\ fermeture_t}{Cours\ d'ouverture_t} - 1$$

- **Rendement nocturne**, (chapitre deux) définit comme étant le rendement entre le cours d'ouverture d'une journée et le cours de fermeture de la veille, tel que :

$$Rendements\ nocturnes_t = \frac{Cours\ d'ouverture_t}{Cours\ de\ fermeture_{t-1}} - 1$$

- **Rendement de volume**, (chapitre 4) le second étant définit comme étant la variation dans les volumes d'échange des actions d'une compagnie entre deux journées de transaction, tel que :

$$Rendements\ de\ volume_t = \frac{Volume_t}{Volume_{t-1}} - 1$$

- **Rendements anormaux**, (chapitre 4) définit comme étant les rendements considérés comme anormaux par rapport au marché (Zhang et al. (2013) et Choi et Varian (2012)), tel que :

$$Rendements\ anormaux_t = (\frac{Rendements\ journaliers_{compagnie,t}}{Rendements\ journaliers_{S\&P500,t}} - 1) * 100$$

Les statistiques descriptives des variables précédemment définies sont présentées dans le tableau 2.2.

Tableau 2.2 : statistiques descriptives des variables utilisées

	Ticker	Name	Volume	Rendement nocturne	Rendement journalier	Rendement anormal
Nombre d'observations	16637	17148	17821	17750	17821	17750
Moyenne	144,4	17121	14972100	0,0001991	0,0002613	0,08824
Médianne	54	1053	8211500	0,0004	0,0003301	0,1282
Mode	32	0	2417900	0	0	0,04271
Maximum	19546	1091212	463491000	0,3944	0,4499	45,6678
Minimum	0	0	107800	-0,754	-0,1594	-76,0579
Déviation standard	491,1	44886,6	23626950	0,01424	0,01597	2,274
Variance	241196,5	2014809445	5,58233E+14	0,0002028	0,000255	5,1713
1er quartile	31	60	3706100	-0,0034	-0,007017	-0,09169
3e quartile	112	8083,8	17402500	0,004	0,007668	1,1212

Plusieurs transformations de variables ont été effectuées. Afin d'éviter de trop grands écarts dans les variables correspondant aux données issues de Twitter, nous normalisons ces variables afin d'obtenir deux indices allant de zéro à cent, cent correspondant au maximum du nombre de tweet pour chaque compagnie. Deux indices pour le nombre de tweet financier (*TickerIndex*) et pour le nombre de tweet mentionnant le nom des compagnies (*NameIndex*) seront donc obtenus. Les différents types de rendements sont transformés de telle sorte que l'on obtient pour chaque rendement une variable correspondant à leurs valeurs absolues, une autre correspondant aux valeurs strictement positives de rendement, et une variable binaire égale à 1 si le rendement est

positif et 0 dans le cas contraire. Finalement, nous prenons le logarithme de toutes ces variables (différents rendements et indice correspondant aux tweets).

Après avoir vérifié la normalité des variables, nous testons leur indépendance. Les résultats de la matrice des corrélations sont confinés dans le tableau 2.3. Nous mettons en valeur les indices de corrélation supérieurs à 50%. On remarque que ces corrélations surviennent pour des variables correspondant à des rendements similaires (par exemple les valeurs absolues de rendements journaliers et les valeurs positives de rendements journaliers). Dans ces cas-ci, cette forte corrélation ne doit pas être prise en compte, car ces variables seront utilisées séparément. Il faut surtout noter la faible corrélation entre les variables de tweets, celles des jours de la semaine, celles des secteurs industriels avec les différents types de rendements, ce qui nous permettra d'effectuer nos analyses économétriques.

Tableau 2.3 : matrice des corrélations des variables utilisées

		rendements journaliers				rendements nocturnes				volume			rendements anormaux			tweets		jours de la semaine					secteurs industriels									
		rendements	log(rendements absolus)	log(rendements positifs)	probit rendements	rendements	log(rendements absolus)	log(rendements positifs)	probit rendements	volume	rendements	probit rendements	log(rendements absolus)	log(rendements positifs)	probit rendements	log(TickerIndex+1)	log(NameIndex+1)	lundi	mardi	mercredi	jeudi	vendredi	santé	énergie	consommation de base	vente au détail	industries	télécommunications	services publics	matériaux	finance	technologies de l'information
rendements journaliers	rendements	1,00	0,70	0,70	0,00	0,13	0,14	0,14	0,00	0,12	0,44	0,17	0,76	0,72	0,00	0,08	0,01	0,03	0,02	-0,01	-0,04	0,00	-0,07	-0,03	-0,05	0,03	0,10	-0,06	-0,06	0,08	0,00	0,00
	log(rendements absolus)	0,70	1,00	1,00	0,00	0,11	0,13	0,13	0,00	0,12	0,27	0,18	0,51	0,50	0,00	0,05	-0,01	0,02	0,02	0,01	-0,06	0,01	-0,07	-0,05	-0,08	0,05	0,05	-0,05	-0,06	0,09	0,02	0,01
	log(rendements positifs)	0,70	1,00	1,00	0,00	0,11	0,13	0,13	0,00	0,12	0,27	0,18	0,51	0,50	0,00	0,05	-0,01	0,02	0,02	0,01	-0,06	0,01	-0,07	-0,05	-0,08	0,05	0,05	-0,05	-0,06	0,09	0,02	0,01
	probit rendements	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
rendements nocturnes	rendements	0,13	0,11	0,11	0,00	1,00	0,59	0,59	0,00	0,12	0,22	0,11	0,65	0,65	0,00	0,11	0,05	0,06	0,02	0,01	-0,02	0,06	-0,05	-0,02	-0,06	0,00	0,05	-0,04	-0,05	0,06	0,03	0,02
	log(rendements absolus)	0,14	0,13	0,13	0,00	0,59	1,00	1,00	0,00	0,16	0,19	0,11	0,40	0,39	0,00	0,06	0,02	-0,06	0,01	0,02	-0,06	0,09	-0,09	-0,01	-0,10	-0,01	0,08	-0,06	-0,08	0,11	0,08	0,01
	log(rendements positifs)	0,14	0,13	0,13	0,00	0,59	1,00	1,00	0,00	0,16	0,19	0,11	0,40	0,39	0,00	0,06	0,02	-0,06	0,01	0,02	-0,06	0,09	-0,09	-0,01	-0,10	-0,01	0,08	-0,06	-0,08	0,11	0,08	0,01
	probit rendements	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
volume	volume	0,12	0,12	0,12	0,00	0,12	0,16	0,16	0,00	1,00	0,23	0,11	0,15	0,16	0,00	0,13	0,08	-0,01	0,01	0,02	-0,02	0,00	-0,05	-0,03	-0,09	-0,16	-0,03	0,03	-0,08	-0,03	0,35	0,05
	rendements	0,44	0,27	0,27	0,00	0,22	0,19	0,19	0,00	0,23	1,00	0,60	0,43	0,40	0,00	0,18	0,04	-0,07	0,06	0,00	0,04	-0,04	0,00	-0,04	-0,02	0,02	0,03	-0,02	0,01	0,00	0,02	-0,02
	probit rendements	0,17	0,18	0,18	0,00	0,11	0,11	0,11	0,00	0,11	0,60	1,00	0,17	0,17	0,00	0,12	0,02	-0,11	0,09	0,02	0,02	-0,04	0,02	-0,02	-0,01	-0,01	0,00	-0,01	0,01	0,00	0,01	0,00
rendements anormaux	log(rendements absolus)	0,76	0,51	0,51	0,00	0,65	0,40	0,40	0,00	0,15	0,43	0,17	1,00	0,97	0,00	0,12	0,03	0,02	0,01	-0,02	-0,02	0,01	-0,07	-0,03	-0,07	0,02	0,09	-0,06	-0,07	0,09	0,01	0,01
	log(rendements positifs)	0,72	0,50	0,50	0,00	0,65	0,39	0,39	0,00	0,16	0,40	0,17	0,97	1,00	0,00	0,12	0,03	0,02	0,01	-0,02	-0,02	0,01	-0,07	-0,03	-0,07	0,02	0,09	-0,06	-0,07	0,09	0,01	0,01
	probit rendements	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
tweets	log(TickerIndex+1)	0,08	0,05	0,05	0,00	0,11	0,06	0,06	0,00	0,13	0,18	0,12	0,12	0,12	0,00	1,00	0,03	0,00	0,00	0,03	0,00	0,00	-0,08	0,00	0,01	-0,07	-0,05	0,10	0,20	0,00	0,12	-0,05
	log(NameIndex+1)	0,01	-0,01	-0,01	0,00	0,05	0,02	0,02	0,00	0,08	0,04	0,02	0,03	0,03	0,00	0,03	1,00	-0,01	0,01	0,00	0,00	0,00	-0,01	0,05	-0,06	0,01	0,00	0,04	-0,18	-0,24	-0,13	0,29
jours de la semaine	lundi	0,03	0,02	0,02	0,00	-0,03	-0,06	-0,06	0,00	-0,01	-0,07	-0,11	0,02	0,02	0,00	0,00	-0,01	1,00	-0,20	-0,20	-0,18	-0,20	-0,21	-0,03	-0,02	0,00	0,01	0,01	-0,01	-0,02	-0,01	0,03
	mardi	0,02	0,02	0,02	0,00	-0,02	0,01	0,01	0,00	0,01	0,06	0,09	0,01	0,01	0,00	0,03	0,01	-0,20	1,00	-0,27	-0,29	-0,30	0,00	0,02	-0,02	-0,01	0,01	0,00	-0,03	0,02	0,01	0,00
	mercredi	-0,01	0,01	0,01	0,00	0,01	0,02	0,02	0,00	0,02	0,00	0,02	-0,02	-0,02	0,00	0,00	0,00	-0,18	-0,27	1,00	-0,26	-0,27	-0,01	-0,01	-0,01	0,01	0,01	0,00	0,00	0,02	0,02	0,00
	jeudi	-0,04	-0,06	-0,06	0,00	-0,02	-0,06	-0,06	0,00	-0,02	0,04	0,02	-0,02	-0,02	0,00	0,05	0,00	-0,20	-0,29	-0,26	1,00	-0,30	0,02	0,01	0,02	-0,01	0,01	0,04	-0,03	0,00	-0,05	
	vendredi	0,00	0,01	0,01	0,00	0,06	0,09	0,09	0,00	0,00	-0,04	-0,04	0,01	0,01	0,00	-0,08	0,00	-0,21	-0,30	-0,27	-0,30	1,00	0,00	-0,01	0,00	0,00	-0,01	0,00	0,00	-0,02	0,03	
secteurs industriels	santé	-0,07	-0,07	-0,07	0,00	-0,05	-0,09	-0,09	0,00	-0,05	0,00	0,02	-0,07	-0,07	0,00	-0,08	-0,01	-0,03	0,00	-0,01	0,02	0,00	1,00	-0,07	-0,08	-0,17	-0,08	-0,05	-0,05	-0,08	-0,11	-0,17
	énergie	-0,03	-0,05	-0,05	0,00	-0,02	-0,01	-0,01	0,00	-0,03	-0,04	-0,02	-0,03	-0,03	0,00	0,00	0,05	-0,02	0,02	-0,01	0,01	-0,01	-0,07	1,00	-0,06	-0,13	-0,06	-0,04	-0,04	-0,06	-0,09	-0,13
	consommation de base	-0,05	-0,08	-0,08	0,00	-0,06	-0,10	-0,10	0,00	-0,09	-0,02	-0,01	-0,07	-0,07	0,00	0,01	-0,06	0,00	-0,02	-0,01	0,02	0,00	-0,08	-0,06	1,00	-0,15	-0,08	-0,04	-0,04	-0,07	-0,10	-0,15
	vente au détail	0,03	0,05	0,05	0,00	0,00	-0,01	-0,01	0,00	-0,16	0,02	-0,01	0,02	0,02	0,00	-0,07	0,01	-0,06	0,01	0,02	0,00	0,00	-0,17	-0,13	-0,15	1,00	-0,16	-0,09	-0,09	0,15	-0,21	-0,32
	industries	0,10	0,05	0,05	0,00	0,05	0,08	0,08	0,00	-0,03	0,03	0,00	0,09	0,09	0,00	-0,05	0,00	0,01	0,01	0,01	-0,01	-0,01	-0,08	-0,06	-0,08	-0,16	1,00	-0,05	-0,04	-0,07	-0,10	-0,16
	télécommunications	-0,06	-0,05	-0,05	0,00	-0,04	-0,06	-0,06	0,00	0,03	-0,02	-0,01	-0,06	-0,06	0,00	0,10	0,04	-0,01	0,00	0,00	0,01	-0,01	-0,05	-0,04	-0,04	-0,09	-0,05	1,00	-0,03	-0,04	-0,06	-0,09
	services publics	-0,06	-0,06	-0,06	0,00	-0,05	-0,08	-0,08	0,00	-0,08	0,01	0,01	0,20	0,20	0,00	0,00	-0,24	-0,01	0,02	0,02	-0,03	0,00	-0,05	-0,04	-0,04	-0,09	-0,04	-0,03	1,00	-0,04	-0,06	-0,09
	matériaux	0,08	0,09	0,09	0,00	0,06	0,11	0,11	0,00	-0,03	0,00	0,00	0,09	0,09	0,00	0,00	-0,24	-0,01	0,02	0,02	-0,03	0,00	-0,08	-0,06	-0,07	-0,15	-0,07	-0,04	-0,04	1,00	-0,10	-0,15
	finance	0,00	0,02	0,02	0,00	0,03	0,08	0,08	0,00	0,35	0,02	0,01	0,01	0,01	0,00	0,12	-0,13	-0,01	0,01	0,02	0,00	-0,02	-0,11	-0,09	-0,10	-0,21	-0,10	-0,06	-0,06	-0,10	1,00	-0,21
	technologies de l'information	0,00	0,01	0,01	0,00	0,02	0,01	0,01	0,00	0,05	-0,02	0,00	0,00	0,01	0,00	-0,05	0,29	0,03	0,00	0,00	-0,05	0,03	-0,17	-0,13	-0,15	-0,32	-0,16	-0,09	-0,09	-0,15	-0,21	1,00

Les sections suivantes détaillent les résultats des modèles MCO et des modèles probit.

2.3 Rendements journaliers (Intraday return)

Afin d'alléger la lecture de l'étude, le détail des résultats concernant les rendements journaliers se trouve à l'annexe B.

2.3.1 Modèles MCO pour les valeurs absolues de rendement (tableau B.1)

Les trois modèles utilisés seront les suivants :

$$|Rendements\ journaliers_t| = \alpha.|Rendements\ journaliers_{t-1}| + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + constante$$

$$|Rendements\ journaliers_t| = \alpha.|Rendements\ journaliers_{t-1}| + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + \gamma.Jours_t + constante$$

$$|Rendements\ journaliers_t| = \alpha.|Rendements\ journaliers_{t-1}| + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + \delta.Industrie_t + constante$$

En comparant les résultats des analyses entre les modèles utilisant la variable IndiceTicker et ceux utilisant la variable IndiceName, on remarque que le R^2 des modèles est à chaque fois légèrement plus élevé dans le premier cas de figure que dans le second. Même s'il reste relativement bas (de 5,94% à 7,02% en utilisant les tweets financiers), les résultats sont dans les mêmes ordres de grandeur que ceux trouvés dans la littérature (Sprenger & Welpe, 2010). En d'autres mots, la seule utilisation de données non structurées ne permet pas d'expliquer l'ampleur des rendements journaliers, mais permet de capturer une fraction de la variance de la variable à expliquer.

La variable IndiceTicker est significative dans tous les modèles. De plus, la valeur des coefficients associés à la variable IndiceTicker est positive, suggérant une influence positive avec les valeurs absolues de rendements journaliers.

L'ajout d'effets fixes aux modèles améliore le pouvoir prédictif de ces derniers, augmentant de 32% la valeur du R^2 en ajoutant les effets fixes correspondant aux types d'industries. Le fait

d'appartenir aux industries de la santé, de l'énergie, de la consommation de base, des services publics et des télécommunications impacte négativement les valeurs absolues de rendements journaliers (comparativement aux entreprises du secteur des technologies de l'information). Au contraire, le fait d'appartenir au secteur de la vente au détail, au secteur industriel et au secteur des matériaux a un impact positif sur la variable étudiée.

La variable IndiceName n'est quant à elle jamais significative.

2.3.2 Modèles MCO avec décalage temporel (tableau B.2)

Les cinq modèles utilisés pour étudier l'influence des messages publiés dans le passé sont les suivants :

$$|Rendements\ journaliers_t| = \alpha.|Rendements\ journaliers_{t-1}| + \beta.\log(IndiceTicker[ou]IndiceName_{t-\theta} + 1) + constante$$

avec θ allant de 0 à 4, θ étant le retard ajouté dans les variables IndiceTicker et IndiceName.

Sans décalage temporel et avec un décalage d'une journée, la variable IndiceTicker est fortement significative dans les modèles proposés, tandis que la variable IndiceName ne l'est pas, mais le devient progressivement à partir d'un décalage temporel d'une journée. Ainsi, le jour-même, l'utilisation du nombre de tweets mentionnant le nom des compagnies ne semble pas propice à l'interprétation de la magnitude des rendements journaliers.

De plus, le coefficient associé à l'IndiceTicker est deux fois plus important en amplitude, suggérant un impact deux fois plus important sur la variable étudiée. Il devient négatif à partir de deux jours précédant les observations, ce qui indique que l'effet de l'augmentation (diminution) d'un point de l'indice Ticker impactera négativement (positivement) deux fois plus intensément que l'augmentation (la diminution) d'un point de l'indice Name sur les rendements journaliers futurs.

2.3.3 Modèles MCO centrés sur les rendements positifs (tableau B.3)

Afin de mesurer l'impact des variables sur les rendements strictement positifs, les trois modèles étudiés sont les suivants :

$$Rendements\ journaliers_{t,positif} = \alpha . Rendements\ journaliers_{t-1,positif} + \beta . \log(IndiceTicker[ou]IndiceName_{t-1} + 1) + constante$$

$$Rendements\ journaliers_{t,positif} = \alpha . Rendements\ journaliers_{t-1,positif} + \beta . \log(IndiceTicker[ou]IndiceName_{t-1} + 1) + \gamma . Jours_t + constante$$

$$Rendements\ journaliers_{t,positif} = \alpha . Rendements\ journaliers_{t-1,positif} + \beta . \log(IndiceTicker[ou]IndiceName_{t-1} + 1) + \delta . Industrie_t + constante$$

Cette fois-ci, la variable IndiceName est significative dans les trois modèles testés (sans effets fixes, avec effets fixes relatifs aux jours de la semaine, avec effets fixes relatifs aux types d'industrie).

Les coefficients de corrélation diminuent légèrement dans le cas des modèles utilisant la variable IndiceTicker par rapport aux modèles MCO en valeurs absolues. Néanmoins, le coefficient de la variable IndiceTicker est plus élevé que précédemment, ce qui suggère qu'une augmentation d'un pourcent de l'indice est corrélée positivement avec l'augmentation des les rendements positifs journaliers.

Finalement, les jours de la semaine s'avèrent non significatifs pour ces types de modèles, et seules quelques industries le sont. Ainsi, les industries de la santé, de la consommation de base, des télécommunications et des services publiques ont toutes une corrélation négative sur les rendements positifs des compagnies, alors que l'industrie des matériaux garde un effet positif sur ce type de rendement.

2.3.4 Modèles probit avec variables de contrôle (tableau B.4)

Les modèles probit utilisés dans cette sous-partie sont les suivants :

$$\Pr(Rendements\ journaliers_{t,\varphi}) = \alpha . Intraday\ return_{t-1,binaire} + \beta . \log(IndiceTicker[ou]IndiceName_t + 1) + constante$$

$$\Pr\left(Rendements\ journaliers_{t,\varphi}\right) = \alpha.Intraday\ return_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + \gamma.Jours_t + constante$$

$$\Pr\left(Rendements\ journaliers_{t,\varphi}\right) = \alpha.Intraday\ return_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + \delta.Industrie_t + constante$$

Avec φ étant les intervalles prédéfinies au début de la partie Méthodologie.

Cette série de modèles permet de raffiner les résultats précédents en introduisant des seuils de rendement. À nouveau, les modèles sont testés en utilisant séparément les variables IndiceTicker et IndiceName. L'interprétation des effets marginaux s'effectue par rapport à un pourcentage de chance de passer de l'état 0 à l'état 1, soit d'avoir des rendements compris dans les intervalles étudiés.

Il faut noter que nous avons ajouté la variable dépendante comprenant un retard d'une journée dans tous les modèles. Plus explicitement, cette variable ajoutée est une variable binaire qui prendra la valeur 1 quand le rendement de la veille a été positif, et 0 dans le cas contraire.

Plus les seuils sont élevés, plus le coefficient de corrélation des modèles augmente. Ceci est confirmé pour les trois types de modèles, que ce soit sans effets fixes, avec les effets fixes relatifs aux industries ou avec les effets fixes relatifs aux jours de la semaine. Ensuite, la variable IndiceTicker n'est pas significative pour un rendement journalier compris entre 0 et 1%, mais l'est pour tous les rendements supérieurs à 1% (sauf si on le contrôle par les types d'industrie ; IndiceTicker devient significatif). À l'image des coefficients de corrélation, les effets marginaux de la variable IndiceTicker augmentent avec l'augmentation de l'ampleur des seuils considérés. Dans le cas le plus prononcé, l'augmentation d'un pourcent d'IndiceTicker entraîne une augmentation de 21,42% des chances d'obtenir un rendement journalier supérieur à 10% (si la veille le rendement de la compagnie au cours de la journée était positif).

En observant les résultats des modèles probit comprenant les jours de la semaine, seul le jeudi s'avère significatif pour les rendements compris entre 1 et 5%. Le fait d'être jeudi est corrélé avec une diminution de 11,6% de chance d'obtenir de tels rendements par rapport au fait d'être mardi. Aux seuils supérieurs, les jours ne sont plus significatifs. Ils le sont néanmoins si on ne considère

que les rendements positifs, mais d'après la valeur des R^2, cette dernière observation est à relativiser.

Pour les modèles utilisant les types d'industrie, il faut noter que chaque seuil permet d'obtenir des modèles représentatifs. Le fait d'appartenir aux secteurs de la santé, de l'énergie, de la consommation de base, des télécommunications et des services publics augmente entre 10,54% et 13,14% les chances d'obtenir un rendement journalier compris entre 0 et 1% (par rapport aux compagnies du secteur des technologies de l'information). Seul le fait d'appartenir au secteur des matériaux augmente les chances d'obtenir un rendement compris entre 1 et 5%.

Concernant les modèles utilisant l'IndiceName, seul un modèle sur quinze s'avère significatif par rapport à cette variable (celui correspondant aux rendements compris entre 1 et 5) alors qu'avec l'indice Ticker, treize fois sur quinze la variable de tweet est significative.

Le signe de l'effet marginal du seul modèle significatif indique que la mention du nom des compagnies sur Twitter diminue les chances d'obtenir des rendements compris entre 1 et 5%. À l'inverse, l'effet marginal de l'IndiceTicker reste positif, suggérant que la mention de tweets financiers est corrélée avec une plus grande chance d'obtenir les rendements désirés.

2.3.5 Modèles probit avec décalage temporel d'une journée (tableau B.5)

À l'image des modèles précédemment abordés, nous introduisons un décalage temporel d'une journée dans la variable relative aux tweets, tel que :

$$Pr\left(Rendements\ journaliers_{t,\varphi}\right) = \alpha.Intraday\ return_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1} + 1) + constante$$

$$Pr\left(Rendements\ journaliers_{t,\varphi}\right) = \alpha.Intraday\ return_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1} + 1) + \gamma.Jours_t + constante$$

$$Pr\left(Rendements\ journaliers_{t,\varphi}\right) = \alpha.Intraday\ return_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1} + 1) + \delta.Industrie_t + constante$$

Les résultats du B.5 ont été élagués afin de ne garder que les modèles dont les variables étudiées sont significatives.

En introduisant un retard d'une journée dans les variables IndiceTicker et IndiceName, il est possible d'obtenir un modèle prédictif quant aux rendements journaliers des compagnies étudiées. Cette fois-ci, seuls les rendements compris entre 1 et 5% peuvent être expliqués par l'IndiceTicker. L'influence de cette variable est corrélée négativement avec les rendements journaliers, peu importe le type d'effets fixes que nous introduisons dans les modèles. Cette influence négative est similaire pour les modèles utilisant la variable IndiceName pour des rendements compris entre 1 et 5%. Il faut noter que pour des rendements supérieurs à 10%, seuls les modèles employant la variable IndiceName s'avèrent significatifs. L'influence de cette variable est positive.

La partie suivante permettra en détail d'étudier l'influence des variables de contrôle sur les modèles proposés avec l'introduction d'interaction entre les variables.

2.3.6 Modèles probit avec interaction de variables (tableau B.6)

Finalement, nous désirons interpréter adéquatement l'effet des variables indépendantes et des variables relatives aux messages publiés sur Twitter sur le rendement étudié. Pour cela, nous proposons un modèle avec interaction de variables.

Les quatre modèles étudiés sont les suivants :

$$Pr\left(Rendements\ journaliers_{t,\varphi}\right) = \alpha.intraday\ return_{t-1\ ;\ binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \gamma.Jours_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Jours_t] + constante$$

$$Pr\left(Rendements\ journaliers_{t,\varphi}\right) = \alpha.intraday\ return_{t-1\ ;\ binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \delta.Industrie_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Industrie_t] + constante$$

En prenant la différentielle des deux dernières équations, par rapport au jour de la semaine ou par rapport au type d'industrie, on obtient la réelle contribution de la variable concernant les messages publiés sur Twitter pour un jour de la semaine donnée, ou un type d'industrie donné, tel que :

$$\frac{\partial(Pr(Intraday\ return)_{t,\varphi})}{\partial Industrie_t} = \delta + \tau.\log(IndiceTicker[ou]IndiceName_{t-1} + 1)$$

Avec τ l'effet de l'augmentation la veille d'un point de l'IndiceTicker (ou IndiceName) sur le fait d'obtenir un rendement compris dans l'intervalle φ pour les compagnies de l'industrie spécifiée.

Le tableau suivant résume l'implication des variables de contrôles pour les deux types d'indice étudiés avec les seuils appropriés. Seules les valeurs significatives sont conservées. À nouveau, les pivots des modèles sont le mardi pour les jours de la semaine, et l'industrie des technologies de l'information pour les types d'industrie.

Tableau 2.4: tableau récapitulatif des effets marginaux pour les rendements journaliers

	Rendements journaliers									
	IndiceTicker (t-1)					IndiceName (t-1)				
Effets marginaux	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**
Pivot : mardi										
Lundi										
Mercredi			7,40%		61,50%					
Jeudi										
Vendredi									-17,37%	55,09%
Pivot : secteur des technologies de l'information										
Santé		-21,18%	19,59%				-9,78%	15,08%		
Énergie			-15,04%				-32,43%	41,38%		
Consommation de base	6,95%	15,85%	-18,38%				9,22%	-15,51%	-61%	
Vente au détail				33,78%					-53,92%	-56,82%
Industries		-15,80%	12,57%						-50,96%	
Télécommunications		-15,97%								
Services publics										
Matériaux	-8,97%	-12,30%		55,33%		-9,40%	-13,49%		-33,65%	
Finance	-7,19%	-8,26%		43,01%		-5,60%		-7,36%		

À la lumière de ces résultats, un investisseur peut effectuer de manière éclairée des recommandations d'achats basées sur des données non structurées. Il peut maximiser son opportunité de gain en ne sélectionnant que des entreprises faisant partie de certaines industries, ou en effectuant ses investissements à certains jours de la semaine. Si l'IndiceTicker augmente d'un point, alors le fait d'être mercredi apporte 7,40% de chance en plus d'obtenir un rendement compris entre 1 et 5% et 61,50% de chance en plus d'obtenir un rendement supérieur à 10% par rapport au mardi.

Les industries des matériaux (+55,33%), de la finance (+43,01%) et de la vente au détail (+33,78%) sont les industries à privilégier pour obtenir des rendements compris entre 5 et 10% par rapport à l'industrie des technologies de l'information lorsque la variable IndiceTicker augmente le jour précédent.

En utilisant la variable dédiée aux noms des compagnies cités sur Twitter, deux résultats sont à noter. Ensuite, concernant des rendements élevés (entre 5 et 10%, puis supérieurs à 10%), le fait d'appartenir aux secteur de la consommation de base, de la vente au détail et industriel diminue de moitié les chances d'obtenir de tels rendements par rapport à l'industrie des technologies de l'information quand l'IndiceName augmente. L'impact négatif de ces effets marginaux peut suggérer que l'augmentation du nombre de messages mentionnant le nom des compagnies n'est jamais favorable à l'obtention de hauts rendements, notamment lors de crises touchant une compagnie.

Dans la prochaine partie, nous étudierons l'effet des variables sur un autre type de rendement, le rendement nocturne. Nous utiliserons une méthodologie similaire, tout en rappelant les équations des modèles utilisés.

2.3.7 Rendements journaliers : conclusions

Afin de résumer les différents résultats obtenus, voici un récapitulatif des différentes observations préalablement obtenues :

- Le construit référant aux messages financiers sur Twitter est corrélé positivement avec l'intensité des rendements journaliers, alors que l'on note l'absence de corrélation pour l'IndiceName **(H1 confirmée)**
- L'effet de cette corrélation s'inverse en introduisant un retard temporel dans les modèles. Toutefois, l'indice Ticker reste significatif, et l'indice Name le devient avec un décalage temporel de deux jours
- À nouveau, l'indice Ticker est corrélé positivement avec les rendements strictement positifs. L'indice Name l'est aussi, mais dans une moindre mesure **(H1 confirmée)**
- Les modèles probit nous permettent de déterminer qu'à nouveau, l'IndiceTicker reste significatif pour la plupart des seuils étudiés, notamment avec l'ajout d'effets fixes dans les modèles, ce qui n'est pas le cas pour l'indice Name **(H1 confirmée)**
- En considérant les variables IndiceTicker et IndiceName comportant un retard temporel d'une journée, plusieurs relations sont obtenues par rapport aux rendements journaliers **(H1 non confirmée)**
- L'interaction des variables de contrôle avec les variables IndiceTicker et IndiceName comportant un retard temporel s'avère un véritable outil décisionnel pour les investisseurs,

présentant ainsi un panel de possibilités par rapport à deux références, soit par rapport à l'industrie des technologies de l'information et par rapport à la journée du mardi (mais cette méthode peut être déclinée pour tout autre industrie, à tout autre moment de la semaine) **(H2 confirmée)**

2.4 Rendements nocturnes (Overnight return)

Les rendements nocturnes modélisent les changements de prix des actions en dehors des heures d'ouverture des marchés. Maîtriser ces rendements permettrait de prévoir les soubresauts inhabituels occasionnés par des nouvelles publiées au courant d'une nuit. Le détail des résultats des différents modèles sont confinés dans les différentes parties de l'annexe C.

2.4.1 Modèles MCO pour les valeurs absolues de rendement (tableau C.1)

Les trois modèles utilisés seront les suivants :

$$|Rendements\ nocturnes_t| = \alpha.\log(|Rendements\ nocturnes_{t-1}|) + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + constante$$

$$|Rendements\ nocturnes_t| = \alpha.\log(|Rendements\ nocturnes_{t-1}|) + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + \gamma.Jours_t + constante$$

$$|Rendements\ nocturnes_t| = \alpha.\log(|Rendements\ nocturnes_{t-1}|) + \beta.\log(IndiceTicker[ou]IndiceName_t + 1) + \delta.Industrie_t + constante$$

Concernant les valeurs absolues des rendements nocturnes, il est intéressant de noter que la variable IndiceName est significative dans tous les modèles, tout comme la variable IndiceTicker. L'effet de ces variables est toutefois différent dans son ampleur, la variable IndiceTicker possédant un coefficient trois à quatre fois plus important que celui de la variable IndiceName.

Les jours de la semaine s'avèrent être une donnée pertinente pour ce genre de modèle : seul le lundi ne s'avère pas significatif par rapport au mardi. L'ajout de cette variable de contrôle permet

d'augmenter le coefficient de corrélation de 1,33% à 1,56% avec l'utilisation de la variable IndiceTicker, puis de 0,47% à 0,66% pour la variable IndiceName.

Finalement, toutes les types de secteurs industriels apparaissent significatif (sauf pour le secteur de la vente au détail avec l'IndiceTicker et le secteur de l'énergie avec l'IndiceName). Le signe devant cette variable binaire reste le même et pour le modèle utilisant les tweets financiers et pour le modèle utilisant le second type de tweet. Le fait d'appartenir aux secteurs de la santé, de l'énergie, de la consommation de base, des télécommunications ou des services publics est apparenté à une corrélation négative par rapport aux valeurs absolues du rendement nocturne. Au contraire, les secteurs industriels, des matériaux et de la finance possèdent évoluent avec la même direction que les rendements nocturnes.

2.4.2 Modèles MCO avec décalage temporel (tableau C.2)

Afin de mesurer l'impact des variables sur les rendements strictement positifs, les trois modèles comparés sont les suivants :

$$|Rendements\ nocturnes_t| = \alpha.\log(|Rendements\ nocturnes_{t-1}|) + \beta.\log(IndiceTicker[ou]IndiceName_{t-\theta} + 1) + constante$$

avec θ allant de 0 à 4, θ étant le retard ajouté dans la variable relative aux tweets dans les modèles.

En introduisant un décalage temporel dans les modèles précédemment étudiés, on remarque qu'avec une journée d'anticipation, la variable IndiceName perd sa significativité. Cette caractéristique est par contre retrouvée avec un décalage temporel plus élevé.

La variable IndiceTicker est quant à elle significative à toutes les périodes étudiées. Le coefficient associé aux rendements nocturnes est positif pour ce qui est de la journée d'observation, puis devient négatif à partir d'un retard supérieur à 1 journée. Ce coefficient est aussi deux à trois fois plus important en termes de magnitude pour la variable Ticker par rapport à la variable Name.

2.4.3 Modèles MCO centrés sur les rendements positifs (tableau C.3)

Afin de mesurer l'impact des variables sur les rendements strictement positifs, les trois modèles comparés sont les suivants :

$$Rendements\ nocturnes_{t,positif} = \alpha . Rendements\ nocturnes_{t-1,positif} + \beta . \log(IndiceTicker[ou]IndiceName_{t-1} + 1) + constante$$

$$Rendements\ nocturnes_{t,positif} = \alpha . Rendements\ nocturnes_{t-1,positif} + \beta . \log(IndiceTicker[ou]IndiceName_{t-1} + 1) + \gamma . Jours_t + constante$$

$$Rendements\ nocturnes_{t,positif} = \alpha . Rendements\ nocturnes_{t-1,positif} + \beta . \log(IndiceTicker[ou]IndiceName_{t-1} + 1) + \delta . Industrie_t + constante$$

Cette fois-ci, on observant seulement les rendements positifs, on remarque que la variable IndiceName est à chaque fois significative, ce qui est aussi le cas de la variable IndiceTicker.

Le coefficient associé à l'IndiceTicker est trois à quatre fois plus élevé que le coefficient associé à l'IndiceName. Finalement, ces coefficients sont de même signe, suggérant une influence positive sur les variables étudiées.

2.4.4 Modèles probit avec variables de contrôle (tableau C.4)

Les modèles probit utilisés dans cette sous-partie sont les suivants :

$$\Pr(Rendements\ nocturnes_{t,\varphi}) = \alpha . Rendements\ nocturnes_{t-1,binaire} + \beta . \log(IndiceTicker[ou]IndiceName_t + 1) + constante$$

$$\Pr(Rendements\ nocturnes_{t,\varphi}) = \alpha . Rendements\ nocturnes_{t-1,binaire} + \beta . \log(IndiceTicker[ou]IndiceName_t + 1) + \gamma . Jours_t + constante$$

$$\Pr(Rendements\ nocturnes_{t,\varphi}) = \alpha . Rendements\ nocturnes_{t-1,binaire} + \beta . \log(IndiceTicker[ou]IndiceName_t + 1) + \delta . Industrie_t + constante$$

Sur les quinze modèles étudiés, quatorze s'avèrent significatifs dans le cas de l'IndiceTicker, et seulement cinq le sont avec l'IndiceName.

Dans le modèle n'utilisant pas les variables de contrôle, la variable Ticker s'avère significative pour chacun des seuils, sauf pour celui concernant les rendements compris entre 0 et 1%. Pour les deux derniers seuils, l'ampleur des coefficients devant la variable indique qu'un pourcent gagné de la variable Ticker est corrélée avec une augmentation de 51% quant à la probabilité d'obtenir un rendement nocturne positif de 5% et plus. Les R^2 des modèles sont aussi élevés dans ces deux cas de figure, étant de 13,79% (5-10% de rendement) et 14,28% (10% et plus). Par rapport à l'indice Name, on remarque que les coefficients de corrélation sont plus faibles (respectivement 2,82% et 5,01%). De plus, le coefficient associé à la variable Name est 2,5 fois plus faible que celui associé à la variable Ticker.

Avec l'ajout de variables de contrôle associées aux jours de la semaine, on augmente légèrement le coefficient de corrélation des modèles. Que ce soit pour les modèles avec la variable Ticker ou ceux avec la variable Name, les mêmes jours s'avèrent significatifs pour certains seuils. Pour les rendements compris entre 1 et 5%, seuls le jeudi et le vendredi apparaissent pertinents, ayant un coefficient positif par rapport à la référence qu'est le mardi.

L'introduction de variable concernant les types d'entreprise améliore l'explication des modèles. En particulier, les rendements nocturnes compris entre 1 et 5% sont sensibles à sept type de secteurs industriels, notamment le secteur de la santé (influence négative), le secteur de la consommation de base (négative), le secteur des industries (positive), le secteur des télécommunication (négative), le secteur des services publics (négative), des matériaux (positive) et de la finance (positive). Ce type d'influence, caractérisée par le signe du coefficient associé aux secteurs industriels, est similaire pour les modèles utilisant les variables IndiceTicker et IndiceName.

2.4.5 Modèles probit avec décalage temporel d'une journée (tableau C.5)

À l'image des modèles précédemment abordés, nous introduisons un décalage temporel d'une journée dans la variable relative aux tweets, tel que :

$$\Pr\left(Rendements\ nocturnes_{t,\varphi}\right) = \alpha.Rendements\ nocturnes_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + constante$$

$$\Pr(Rendements\ nocturnes_{t,\varphi}) = \alpha.Rendements\ nocturnes_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \gamma.Jours_t + constante$$

$$\Pr(Rendements\ nocturnes_{t,\varphi}) = \alpha.Rendements\ nocturnes_{t-1,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \delta.Industrie_t + constante$$

En étudiant ces mêmes modèles, on remarque que l'ajout d'un décalage temporel d'une journée modifie fortement le nombre de modèles pouvant être expliqués. En effet, sur quinze modèles concernant chacune des deux variables étudiées (IndiceTicker et IndiceName), seuls neuf modèles présentent des variables significatives pour la variable IndiceTicker (contre quatorze sans décalage temporel) et trois pour la variable Name (contre sept sans décalage temporel).

L'influence de la variable Ticker s'avère positive dans les neuf modèles obtenus, tandis que celle de la variable Name est négative pour les rendements compris entre 1 et 5%. À nouveau, les industries apparaissent fortement significatives pour ce type de rendement en ce qui concernent les rendements nocturnes compris entre 1 et 5%.

2.4.6 Modèles probit avec interaction de variables (tableau C.6)

Les modèles étudiés pour cette dernière partie sont les suivants :

$$Pr(Rendements\ nocturnes)_{t\,;\,\varphi} = \alpha.Rendements\ nocturnes_{t-1\,;\,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \gamma.Jours_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Jours_t] + constante$$

$$Pr(Rendements\ nocturnes)_{t\,;\,\varphi} = \alpha.Rendements\ nocturnes_{t-1\,;\,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \delta.Industrie_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Industrie_t] + constante$$

En prenant la différentielle des deux équations précédentes, par rapport au jour de la semaine ou par rapport au type d'industrie, on obtient la réelle contribution de la variable concernant les

messages publiés sur Twitter pour un jour de la semaine donnée, ou un type d'industrie donné, tel que :

$$\frac{\partial(Pr(Rendements\ nocturnes)_{t\,;\,\varphi})}{\partial Industrie_t} = \delta + \tau.\log(IndiceTicker[ou]IndiceName_{t-1} + 1)$$

Avec τ l'effet de l'augmentation la veille d'un point de l'indice Ticker (ou Name) sur le fait d'obtenir un rendement compris dans l'intervalle φ pour les compagnies de l'industrie spécifiée.

Ce type de modèle permet, à l'image de l'étude pour les rendements journaliers, d'extraire l'impact spécifique d'un jour de la semaine ou d'un type d'industrie sur les rendements nocturnes. Quand l'IndiceTicker augmente d'un pourcent, les chances d'obtenir des rendements compris entre 5 et 10% diminuent les mercredi (-17,87%), jeudi (-14,78%) et vendredi (-13,30%) par rapport au mardi. Pour des rendements supérieurs (5 à 10%), le jeudi s'avère être la journal à favoriser, augmentant les chances d'obtenir de tels rendements par 42,18% par rapport au mardi.

L'IndiceName n'est pas significatif en faisant interagir la variable avec les jours de la semaine.

Le type d'industrie ne permet pas d'obtenir d'explication avec l'emploi de la variable IndiceTicker, sauf pour les rendements compris entre 1 et 5% pour l'industrie de la consommation de base (-29,43% par rapport au secteur des technologies de l'information) et des télécommunications (-24,22%). Toutefois, appartenir au secteur de la santé diminue les chances d'obtenir des rendements supérieurs (compris entre 5 et 10%) par 46,98% par rapport au secteur des technologies de l'information.

Par contre, ces analyses par secteur s'avèrent plus significatives en utilisant l'IndiceName. Les résultats les plus marqués correspondent aux rendements très élevés (+10% entre la fermeture des marchés et leur ouverture). Le fait d'appartenir aux secteurs de la vente au détail et de la consommation de base diminue les chances d'obtenir de tels rendements par 168% et 136% respectivement par rapport à l'industrie des technologies de l'information.

Le résumé par jour de semaine et par secteur industriel est confiné dans le tableau 2.5, et le détail de chaque modèle se retrouve au tableau C.6 de l'annexe C.

Tableau 2.5: tableau récapitulatif des effets marginaux pour les rendements nocturnes

	Rendements nocturnes									
	IndiceTicker (t-1)					IndiceName (t-1)				
Effets marginaux	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**
Pivot : mardi										
Lundi		8,29%		-34,71%						
Mercredi		8,85%	-17,87%							
Jeudi			-14,78%	42,18%						
Vendredi		5,93%	-13,30%							
Pivot : secteur des technologies de l'information										
Santé				-46,98%					-16,66%	
Énergie								25,87%		
Consommation de base			-29,43%			-6,37%		-29,64%		-168,52%
Vente au détail								-12,54%		-136,85%
Industries								-13,67%		
Télécommunications			-24,22%							
Services publics						-19,25%	-15,67%			
Matériaux								-10,40%		
Finance						-6,78%		-18,44%	-40,58%	

2.4.7 Rendements nocturnes : conclusions

Afin de résumer les différents résultats obtenus, voici un récapitulatif des différentes observations préalablement obtenues :

- Autant le construit référant aux messages financiers que celui référant aux noms des entreprises sont corrélés positivement avec l'intensité des rendements nocturnes. Néanmoins, l'effet de la variable dédiée aux messages financiers est deux fois plus important **(H1 non confirmée)**
- À l'image des rendements journaliers, l'effet des construits des Tweets s'avère négatif en considérant l'ajout d'un décalage temporel dans les modèles étudiés
- La variable IndiceTicker reste significative pour l'ensemble des modèles probit sans décalage temporel (quatorze modèles sur quinze) tandis que la variable Name ne l'est que pour sept modèles sur quinze **(H1 confirmée)**
- L'ajout d'un décalage temporel réduit le nombre de modèles explicatifs concluants pour les deux types de variables utilisées
- En utilisant l'interaction de variables entre les messages publiés sur Twitter et les variables de contrôle des jours de la semaine et des types d'industrie, il est possible d'extraire des relations prenant en compte les rendements nocturnes **(H2 confirmée)**

En résumant ce chapitre, l'emploi des médias sociaux permet d'affiner et d'optimiser les opportunités de rendements des compagnies inscrites en bourse. Ces rendements (journaliers et nocturnes) ne sont pas corrélés avec la même ampleur quand on considère des modèles comportant

les noms des compagnies (IndiceName) ou des messages financiers (IndiceTicker). En effet, les rendements journaliers semblent reliés plus intimement aux variations de la variable IndiceTicker (sans être sensible à aux variations de la variable IndiceName), tandis que pour les rendements nocturnes, les deux variables semblent significatives. **L'hypothèse 1 est ainsi confirmée dans le cas des rendements journaliers, mais ne peut l'être dans le cas des rendements nocturnes.**

À partir de ces résultats, il a été possible de déterminer les pourcentages de chance de dépasser différents seuils de rendements par rapport aux variables de référence, permettant d'esquisser des stratégies d'investissements. **L'hypothèse 2 est ainsi confirmée.**

Les autres types de rendements (anormaux et concernant les volumes) seront explicités au chapitre quatre avec la mise en place de stratégies d'investissements utilisant ces résultats.

De par l'ampleur des messages publiés sur Twitter (500 millions de messages par jour, dont 30000 financiers), une question s'impose : est-il possible de se concentrer sur une partie des messages en particulier ? Pour répondre à cette interrogation, il est nécessaire en premier lieu de comprendre la nature des messages publiés, et surtout les types d'utilisateurs présents sur Twitter. Cette problématique fera l'objet du troisième chapitre de ce mémoire.

CHAPITRE 3 CHUT! UN BRUIT…

Le deuxième chapitre de ce mémoire s'est penché sur l'analyse économétrique détaillée de la relation entre les rendements journaliers (ou nocturnes) avec le nombre de messages publiés sur Twitter, que ce soit des messages mentionnant le nom des compagnies ou l'indice d'identification sur les marchés financiers.

Ce troisième chapitre abordera quatre aspects. (1) Tout d'abord, nous continuerons l'analyse effectuée dans le second chapitre, mais cette fois-ci par rapport aux rendements anormaux puis concernant les volumes d'actions échangées. Sans entrer dans les détails méthodologiques, nous fournirons les tableaux de contrôles obtenus à partir des modèles probit avec interaction de variable. (2) Par la suite, nous étudierons une période fortement volatile pour une entreprise inscrite en bourse, celle où ses rapports financiers sont publiés. Nous avons établi des tableaux de contrôle pour les quatre types de rendements étudiés pendant la semaine où les résultats sont rendus publics, puis durant la semaine précédant leur parution. (3) Nous présenterons une stratégie d'investissement basée sur les volumes d'actions et sur les volumes de messages publiés. (4) Finalement, nous conclurons ce chapitre par des recommandations quant à l'utilisation des médias sociaux en finance, du point de vue des investisseurs, des compagnies et des organismes afin de valoriser l'information publique sur les médias sociaux.

3.1 Tableaux de contrôles des rendements (anormaux et volume)

La méthodologie utilisée est la même que celle du chapitre 2. Néanmoins, pour des soucis de clarté, nous n'aborderons que les résultats des modèles probit avec interaction de variables.

Nous avons définit un type de rendement concernant les volumes d'actions échangés dans nos modèles probit. La variable dépendante prend la valeur 1 lorsque les volumes échangés sont plus importants que ceux de la veille de l'ordre des seuils considérés (0%+ ; 0-1% ; 1-5% ; 5-10% ; 10%+), puis 0 dans le cas contraire.

Les quatre modèles utilisés pour les volumes d'actions échangées sont les suivants :

$$Pr(Volume\ return)_{t\,;\,\varphi} = \alpha.Volume\ return_{t-1\,;\,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \gamma.Jours_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Jours_t] + constante$$

$$Pr(Volume\ return)_{t\,;\,\varphi} = \alpha.Volume\ return_{t-1\,;\,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \delta.Industrie_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Industrie_t] + constante$$

Le tableau suivant résume les résultats obtenus, à nouveau en comparant les effets marginaux de IndiceTicker et de l'IndiceName. Lorsque l'IndiceTicker augmente d'un pourcent, le fait d'être jeudi diminue la probabilité d'avoir plus d'échanges par 19,57 % (par rapport au mardi). De la même façon, si l'IndiceName augmente d'un pourcent, la probabilité d'obtenir un rendement supérieur à 10% diminue de 5,60% par rapport au mardi.

En prenant en compte les secteurs industriels, lorsque l'IndiceTicker augmente, les probabilités d'obtenir des rendements compris entre 5 et 10% diminuent pour les secteurs des télécommunications et de la finance.

Tableau 3.1: tableau récapitulatif des effets marginaux pour les rendements concernant les volumes d'action échangés

	Volume									
	IndiceTicker (t-1)					IndiceName (t-1)				
Effets marginaux	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**
Pivot : mardi										
Lundi										
Mercredi										-5,60%
Jeudi		-19,57%								
Vendredi										
Pivot : secteur des technologies de l'information										
Santé										
Énergie										
Consommation de base						6,55%				
Vente au détail	-6,09%									
Industries										
Télécommunications				-22,06%		8,98%				
Services publics										
Matériaux										
Finance				-13,72%						

Nous avons définit un dernier type de rendement, les rendements anormaux. Ces rendements correspondent au ratio entre les rendements journaliers d'une compagnie par rapport aux rendements journaliers du S&P500.

À nouveau, les quatre équations des modèles utilisés sont les suivantes :

$$Pr(Rendements\ anormaux)_{t\,;\,\varphi} = \alpha.Rendements\ anormaux_{t-1\,;\,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \gamma.Jours_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Jours_t] + constante$$

$$Pr(Rendements\ anormaux)_{t\,;\,\varphi} = \alpha.Rendements\ anormaux_{t-1\,;\,binaire} + \beta.\log(IndiceTicker[ou]IndiceName_{t-1}+1) + \delta.Industrie_t + \tau.[\log(IndiceTicker[ou]IndiceName_{t-1}+1).Industrie_t] + constante$$

De ces résultats, nous pouvons tirer plusieurs observations. L'influence des messages financiers et des secteurs industriels est significative pour les rendements anormaux supérieurs à 5%, en particulier pour les secteurs de la vente au détail, des matériaux et de la finance.

Les rendements anormaux sont optimaux les vendredi lorsque l'indice Ticker augmente d'un pourcent, ayant 27,66% de chance en plus de survenir par rapport au mardi (pour des rendements supérieurs à 10%). Appartenir à l'industrie des matériau diminue les chances d'obtenir des rendements inférieurs à 5%, tandis que ces chances sont plus importantes pour des rendements plus subséquents (supérieur à 5%) par rapport au fait d'appartenir à l'industrie des technologies de l'information.

En se concentrant sur l'IndiceName, ces rendements supérieurs à 5% ont une probabilité plus élevée de survenir pour les compagnies de l'énergie (+68%) par rapport aux compagnies des technologies de l'information. Par contre, pour les compagnies des industries de la consommation de base, de la vente au détail, du secteur industriel et des matériaux, cette probabilité diminue.

Le tableau suivant récapitule les différents résultats obtenus grâce aux modélisations ; et le détail des simulations est disponible à l'annexe D.

Tableau 3.2: tableau récapitulatif des effets marginaux pour les rendements anormaux

	Rendements anormaux									
	IndiceTicker (t-1)					IndiceName (t-1)				
Effets marginaux	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**	**0% +**	**0-1%**	**1-5%**	**5-10%**	**10% +**
Pivot : mardi										
Lundi		-7,27%								
Mercredi										38%
Jeudi									-12,60%	
Vendredi										
Pivot : secteur des technologies de l'information										
Santé	-6,32%	-16,27%					-8,73%			
Énergie							-24,94%	15,98%	68,17%	
Consommation de base		-9,62%	-12,26%				7,78%	-8,64%	-36,76%	-71,88%
Vente au détail				19,37%	45,03%		6,85%		-20,95%	-70,39%
Industries									-29,94%	-46,80%
Télécommunications										
Services publics										
Matériaux	-11,58%	-8,29%	-12%	28,14%	74,19%				-24,94%	
Finance			-7,69%	22,90%			11,96%	-11,18%		

3.2 Période de résultats financiers

Les annonces officielles concernant les résultats financiers des compagnies constituent une période de forte volatilité. C'est l'heure de conforter les investisseurs, d'annoncer les mauvaises nouvelles, ou d'esquisser les plans futurs. Nous avons récolté les dates auxquelles les 71 compagnies étudiées ont publié leurs résultats trimestriels. De ces données, nous avons créé deux variables binaires. La première prend la valeur 1 quand la compagnie se trouve dans la semaine d'annonces officielles (et 0 sinon). La seconde prend la valeur 1 quand la compagnie se trouve la semaine précédant l'annonce de résultats trimestriels (et 0 sinon). Ces dates ont été obtenues à partir du site *finance.yahoo.fr* puis auprès des sites Internet des compagnies lorsque l'information n'était pas disponible.

Nous avons voulu tester les quatre types de rendements au cours de ces deux périodes, puis selon le type de messages utilisé. À nouveau, des tableaux récapitulatifs sont présentés afin d'illustrer les résultats des effets marginaux obtenus grâce aux simulations des modèles probit avec interaction de variables. Les seuils des rendements sont semblables à ceux utilisés précédemment. Dans le cas des semaines précédant les annonces officielles, nous utilisons le nombre de messages publiés cinq jours avant les rendements étudiés afin d'obtenir des interactions de variables interprétables.

Trois résultats peuvent être tirés de ces analyses (les tableaux suivants illustrent les résultats significatifs ; le détail se trouve à l'annexe E).

(1) Les rendements supérieurs à 1% s'avèrent maximaux pendant la semaine d'annonce officielle lorsque le nombre de tweets financiers augmente. C'est le cas pour les rendements journaliers, nocturnes et anormaux. L'augmentation de tweets financiers la semaine précédant les annonces officielles n'a pas d'implication significative sur les rendements journaliers et anormaux.

(2) Lorsque les tweets mentionnant le nom des compagnies augmentent la semaine précédant les annonces officielles, alors les probabilités d'obtenir des rendements nocturnes positifs sont plus faibles que lors de toute autre période.

(3) De faibles rendements (entre 0 et 1%) ont de plus faibles chances de survenir au cours des semaines de divulgation de résultats trimestriels (pour les rendements nocturnes et anormaux).

Tableau 3.3: effet de la divulgation des résultats trimestriels sur les rendements journaliers

	Rendements journaliers			
	IndiceTicker (t-1)		**IndiceName (t-1)**	
Seuils	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel
0% +				
0-1%				
1-5%				
5-10%		26,24%		
10% +				

Tableau 3.4: effet de la divulgation des résultats trimestriels sur les rendements nocturnes

	Rendements nocturnes			
	IndiceTicker (t-1)		**IndiceName (t-1)**	
Seuils	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel
0% +			-7,94%	
0-1%		-11,89%	-6%	
1-5%	11,29%	11,33%		
5-10%		40,60%		
10% +				

Tableau 3.5 : effets de la divulgation des résultats trimestriels sur les volumes d'échange

	Volume			
	IndiceTicker (t-1)		IndiceName (t-1)	
Seuils	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel
0% +		-6,63%		-5,74%
0-1%				
1-5%				-11,70%
5-10%				
10% +				

Tableau 3.6 : effet de la divulgation des résultats trimestriel sur les rendements anormaux

	Rendements anormaux			
	IndiceTicker (t-1)		IndiceName (t-1)	
Seuils	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel	Semaine **avant** rapport trimestriel	Semaine **pendant** rapport trimestriel
0% +				
0-1%		-9,70%		-7,69%
1-5%		6,59%		
5-10%		16,96%		
10% +				

En résumé, afin de maximiser les probabilités d'obtention de forts rendements, il est préférable d'investir lors de la semaine de résultats officiels plutôt la semaine les précédant. De plus, de faibles rendements ont peu de chance de survenir lors des semaines d'annonces officielles.

3.3 Investir avec le bruit

Présentées en dernière partie du chapitre un, les stratégies d'investissement sont multiples et font l'objet d'études approfondies. Ces stratégies dépendent des niveaux de risque auquel un investisseur accepte de s'exposer, et la méthodologie les quantifiant est issue des travaux de Markowitz et de Sharpe. À travers ce mémoire, nous avons vu que les médias sociaux constituent une nouvelle source de données qu'il faut prendre en compte. Plus qu'un simple émetteur de contenu informationnel, Twitter revêt aussi la fonction de miroir de l'opinion publique.

Cette information supplémentaire doit être valorisée afin d'être utilisée au sein de stratégies d'investissement. Dans cette optique, les tableaux de contrôle fournis aux chapitres 2 et 3 pourront servir de matériel de base à l'investisseur « 3.0 ».

Nous avons voulu tester l'hypothèse selon laquelle il est possible d'inférer une stratégie d'investissement à partir du volume de messages émis. Nous nous concentrons à nouveau sur les 71 compagnies constituant notre échantillon. Nous avons ainsi considéré cinq types de données : (1) le nombre de tweets financiers (Ticker) ; (2) le nombre de tweets mentionnant le nom des compagnies (Name) ; (3) le volume d'actions échangées ; (4) le fait que les rendements soient positifs ; (5) le fait que les rendements soient négatifs.

À partir de ces données nous introduisons trois variables binaires. Dans le cas de la variable Ticker, sa variable binaire correspondante prend la valeur 1 quand le nombre de tweets est anormalement élevé, c'est-à-dire supérieur à la moyenne plus une déviation standard sur l'ensemble de la période étudiée, puis 0 dans le cas contraire. On applique la même méthodologie pour la variable Name. Pour les volumes, lorsqu'au cours d'une journée a eu lieu une quantité anormale d'actions échangées, la variable prend la valeur 1, puis 0 dans le cas contraire (à nouveau, une quantité anormale correspond à une quantité supérieure à la somme de la moyenne plus une déviation standard sur la période étudiée).

Nous posons l'hypothèse que si à un jour ∂ une compagnie a été anormalement échangée en bourse, puis qu'elle a été anormalement citée sur les médias sociaux, deux cas de figure surviennent :

1. Soit une mauvaise nouvelle est anticipée, alors les rendements du jour $\partial + 1$ seront négatifs
2. Soit une bonne nouvelle est anticipée, alors les rendements du jour $\partial + 1$ seront positifs

Nous faisons interagir les variables binaires afin d'obtenir une série d'évènements au cours de la période étudiée. Le tableau suivant résume les cas de figure possibles. De plus, nous nous concentrons sur trois types de rendements : *daily_return*, soit les rendements entre les cours de fermeture de deux journées consécutives ; *intraday_return* (rendements journaliers), soit les rendements obtenus au cours d'une journée d'échange en bourse ; *overnight return* (rendements nocturnes), soit les rendements qui surviennent entre le cours d'ouverture d'une journée et le cours de fermeture de la veille.

Tableau 3.6: impact d'évènements anormaux sur les trois types de rendement

Quand la veille se produit un événement anormal (moyenne+ 1déviation standard)	Ces mesures sont impactées	Et ont X% de chance d'être **positive** ou **négative**	
Événement	**Mesures**	**Positive**	**Négative**
ticker	1	40,85%	59,15%
	2	40,85%	59,15%
	3	45,07%	54,93%
ticker x name	1	46,15%	53,85%
	2	41,54%	58,46%
	3	44,62%	55,38%
volume	1	47,89%	52,11%
	2	50,70%	49,30%
	3	43,66%	56,34%
volume x ticker	1	40,30%	59,70%
	2	40,30%	59,70%
	3	58,21%	41,79%
volume x name	1	43,08%	56,92%
	2	46,15%	53,85%
	3	44,62%	55,38%
volume x ticker x name	1	42,11%	57,89%
	2	42,11%	57,89%
	3	45,61%	54,39%
name	1	49,30%	50,70%
	2	42,25%	57,75%
	3	47,89%	52,11%
volume x ticker x positif	1	90,91%	9,09%
	2	100,00%	0,00%
	3	50,00%	50,00%
volume x ticker x negatif	1	52,11%	47,89%
	2	43,66%	56,34%
	3	52,11%	47,89%

1	daily_return
2	intraday_return
3	overnight return

Il est nécessaire de porter notre attention sur les 7 types d'évènements. À titre d'exemple, l'interprétation de ces simulations peut se lire ainsi (seconde ligne du tableau) : « quand la veille est publié un nombre anormalement élevé de tweets financiers, les rendements journaliers du lendemain seront impactés, et auront 40,85% de chance d'être positifs » [sur l'échantillon observé]. En observant la combinaison *volume/ticker/positif*, une stratégie d'investissement peut être mise en place. En effet, lorsque la veille les actions d'une compagnie sont anormalement échangées, que

de nombreux messages sont publiés sur Twitter, puis que l'action a terminé la journée en hausse, alors au cours de l'année d'observation l'action de la compagnie a terminé en hausse dans tous les cas de figure le lendemain (ligne 23 du tableau). Cette stratégie semble ainsi capter le *momentum* sur les marchés financiers.

Nous appellerons cette combinaison un signal. Lorsqu'un signal survient sur certaines actions, alors il est recommandé d'acheter à l'ouverture lesdites actions puis de les revendre avant la fin des marchés le jour même. La simulation effectuée entre mai 2012 et mai 2013 conforte ces hypothèses, avec l'obtention de rendements impressionnants. Nous avons calculé les rendements obtenus en utilisant les rendements journaliers (intraday) et quotidien (daily). La figure suivante illustre l'ampleur des bénéfices potentiels (+878% et +676% pour les rendements quotidiens et journaliers respectivement). Sur la même période, le S&P500 a augmenté de 14,3%.

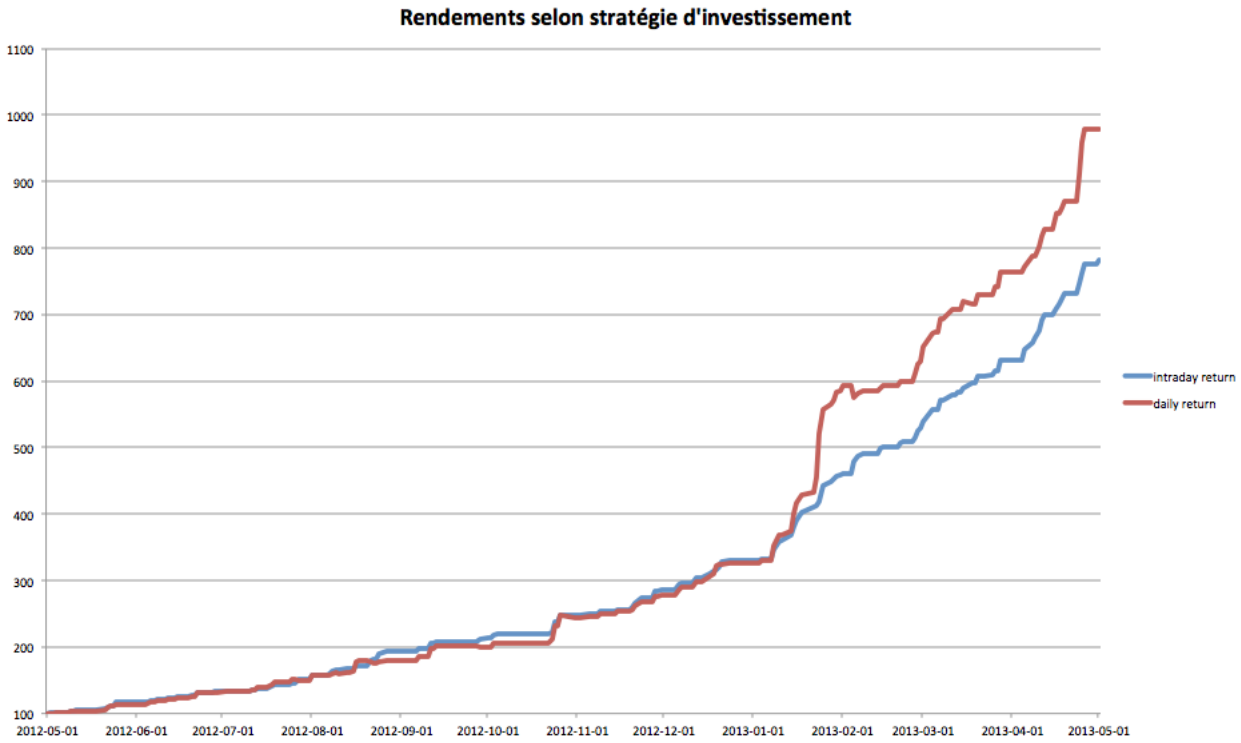

Figure 3.1 : rendements entre le 1er mai 2012 et le 1er mai 2013

Une critique pouvant être formulée serait que penser que l'explication d'un phénomène passé permette d'anticiper les évènements futurs. C'est pourquoi nous avons retenté l'expérience au cours d'une période de trois mois, allant du 1er mai 2013 au 1er août 2013 tout en comparant les

mêmes types de rendements. Toutes les données récoltées ne font pas partie de l'échantillon initial de nos simulations. La figure suivante fait état de résultats surpassant les rendements du S&P500 (+56% pour les rendements quotidiens et +57% pour les rendements journaliers). Sur la même période, le S&P500 a obtenu des rendements de 5,76%.

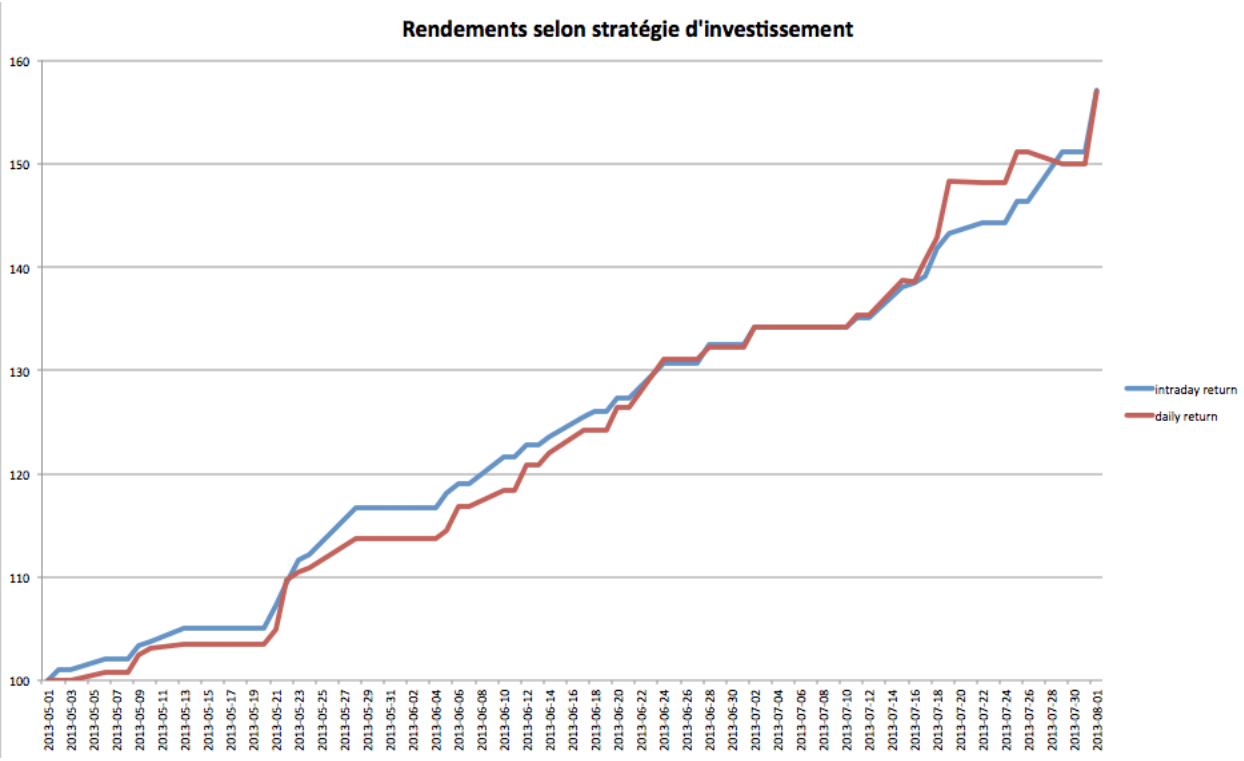

Figure 3.2 : rendements entre le 1er mai 2013 et le 1er août 2013

Ces résultats doivent être nuancés. L'ampleur des rendements obtenus n'est que le reflet d'une situation idéale ne prenant pas en compte de nombreux paramètres. Entre autres, un investisseur serait porté à vendre ses positions au milieu d'une journée si ses investissements s'avèrent rentables au lieu d'attendre la clôture des marchés. De plus, les coûts de transaction ne sont pas pris en compte. Finalement, les ordres d'achat et de vente sont effectués instantanément. Toutes ces nuances relativisent l'ampleur de la stratégie d'investissement abordée.

À la lumière des précédents résultats, bâtir des stratégies d'investissements basées sur l'utilisation des médias sociaux s'avère possible. Les tableaux de bord du chapitre deux puis du présent chapitre permettent ainsi de maximiser les opportunités de gain pour l'investisseur. **Ceci nous permet de confirmer notre deuxième hypothèse de recherche**.

CHAPITRE 4 PROLONGEMENT ET PISTES DE RECHERCHES FUTURES

500 millions de tweets par jour, dont 30 000 à caractère financier en moyenne. Qui sont les personnes communiquant sur les performances des entreprises ? Tous les messages sont-ils pertinents ? Et surtout, y-aurait ‘il un moyen de filtrer l’ensemble des messages afin de ne retirer que le signal essentiel ? Ce sont ces trois questions que nous aborderons au cours de ce second chapitre.

4.1 Méthodologie

L’introduction du chapitre 2 a montré que plusieurs secteurs et compagnies émergent en captant l’ensemble de l’attention sur Twitter. On pense notamment aux compagnies des technologies de l’information comme Google et Apple, mais aussi à certaines compagnies du secteur bancaire. Nous avons voulu observer quelles sont les dynamiques entourant ces compagnies, et plus précisément quelles sont les interactions entre les utilisateurs mentionnant ces compagnies. Ainsi, à partir de la liste du S&P500, nous avons sélectionné 400 compagnies parmi les plus discutées sur Twitter. Cette sélection est le résultat de la méthode d’acquisition de données issues de Twitter, limitant le nombre de requêtes différentes à 400. La liste des compagnies étudiées se trouve à l’annexe F.

En utilisant les packages d’extraction de données de R pour Twitter (*streamR* et *twitteR*), nous avons récolté tous les messages mentionnant ces compagnies entre le 18 novembre 2013 et le 28 février 2014, soit pendant 15 semaines. Cette collecte de données a eu lieu durant trois périodes distinctes, soit trois heures avant l’ouverture des marchés américains (de 6h30 à 9h30), 1h45 aux alentours de l’heure du repas (de 11h45 à 13h30) et trois heures après la fermeture des marchés (de 16h00 à 19h00).

4.2 Structure des messages publiés

Au cours de cette période, 489 342 messages ont été envoyés par 64 504 utilisateurs uniques. Le tableau suivant résume les principales statistiques de contenu associées aux messages.

Tableau 4.1 : caractéristiques des messages récoltés

Contenu des messages	Périodes temporelles			
	6:30 - 9:30	11:45 - 13:30	16:00 - 19:00	3 périodes
Pourcentage de messages contenant un hashtag (#)	25,02%	21,56%	22,67%	23,21%
Pourcentage de messages en réponse à un message (RT)	18,29%	16,41%	22,99%	19,64%
Pourcentage de messages contenant un URL (http)	77,65%	73,60%	75,44%	75,74%
Pourcentage de messages mentionnant un utilisateur (@)	5,06%	6,96%	6,35%	6,06%
Moyenne de caractères utilisés par message	96,8	95,05	99,3	97,3
Nombre de titres financiers mentionnés	2,01	1,95	1,96	1,97
Pourcentage de messages géolocalisés	0,49%	0,50%	0,49%	0,49%
Nombre moyen de messages par heure	954	1 392	812	1 012
Nombre d'utilisateurs uniques	24 134	20 594	36 522	64 504
Total de messages	171 657	126 670	191 015	489 342

Plusieurs caractéristiques des messages financiers sont mises en évidence. D'abord, les messages publiés en matinée avant l'ouverture des marchés comportent en moyenne plus de mots-clefs qu'aux autres périodes de la journée. C'est aussi à cette période de la journée que l'on retrouve le plus de liens internet inclus dans les messages, témoignant d'une volonté de partager l'information mais aussi de rediriger vers un contenu plus développé auprès d'autres sites Internet.

À mi-journée, les messages comportent un taux de mention des autres utilisateurs plus élevé qu'aux autres périodes. Ces messages s'adressent directement aux utilisateurs en les interpelant sur Twitter (i.e. « @WilliamSanger : $TWTR annonce ses résultats annuels cet après-midi »). Le volume de messages par heure est plus élevé à mi-journée par rapport aux autres périodes (1 392 messages par heure en moyenne contre 954 en matinée et 812 en soirée).

Le soir, c'est la proportion de messages retransmis qui se distingue par rapport aux autres périodes. Près de 23% des messages publiés sont des messages retransmis (contre 18,3% et 16,4% en matinée et à mi-journée respectivement), mettant en valeur les liens unissant les utilisateurs entre eux au sein des grappes d'émetteurs-receveurs.

De plus, les compagnies ne sont pas mentionnées dans les mêmes proportions. La popularité des compagnies du S&P500 observée au chapitre 1 est confirmée à travers l'échantillon étudié. En effet, les compagnies des technologies de l'information et du secteur bancaire sont les compagnies capturant la plupart des messages envoyés par jour, avec en moyenne 11%, 6,5%, 4,5% et 4% pour Apple, Facebook, Twitter et Google respectivement. Sept des dix firmes les plus tweetées mentionnées du secteur des technologies de l'information (voir figure 4.1).

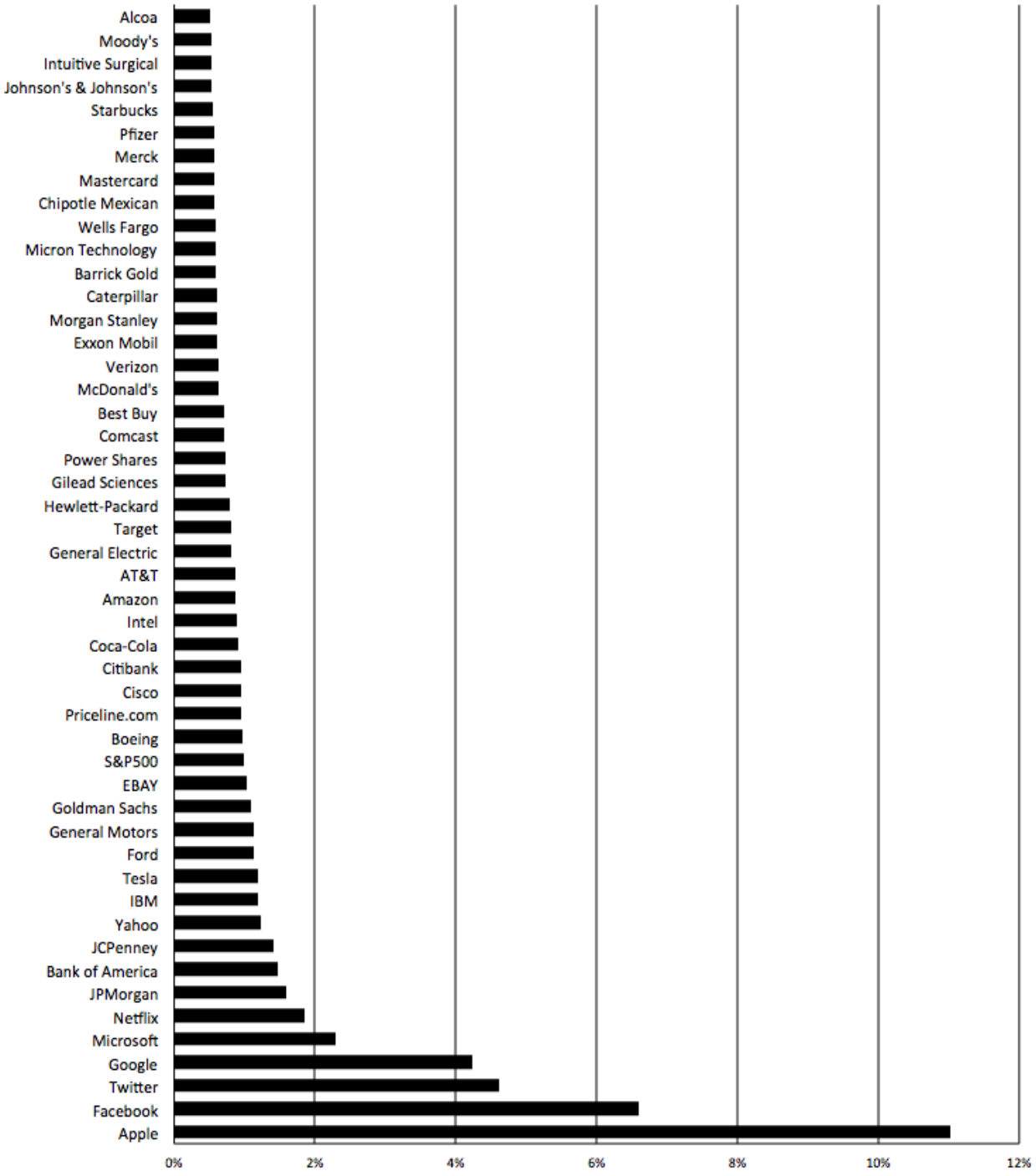

Figure 4.1 : place occupée par les compagnies au sein des discussion sur Twitter

4.3 Comment mesurer l'influence des utilisateurs ?

Cette question reste centrale pour l'interprétation des messages publiés sur Twitter. De quelle manière filtrer l'ensemble des conversations, afin de séparer le juste signal du bruit. Deux mesures

intuitives sont traditionnellement utilisées. La première consiste à évaluer le nombre de personnes suivant les utilisateurs (« followers »). Plus ce nombre est élevé, plus l'audience potentielle des messages envoyés sera importante. Néanmoins, une exposition potentielle ne correspond pas nécessairement à une véritable lecture du message. La seconde mesure envisagée est celle du nombre de messages envoyés. Plus l'espace virtuel est occupé, plus un individu sera présent. Cette dynamique participe grandement à la création de bruit, notamment si le signal envoyé par les utilisateurs n'est ni repris ni pertinent.

Ces deux dimensions seront étudiées à travers l'échantillon de tweets constitué afin de faire ressortir les profils d'utilisateurs correspondant à ces mesures. Elles seront par la suite confrontées à une autre mesure que nous proposons : filtrer les utilisateurs à suivre en fonction de leur impact au sein des réseaux d'utilisateurs, en utilisant notamment le nombre d'utilisateurs distincts retransmettant leurs messages. La question de la géolocalisation des messages sera finalement abordée.

4.3.1 Utilisateurs les plus « populaires »

Le tableau 4.2 représente les utilisateurs cumulant le plus grand nombre de followers à travers l'échantillon étudié. Parmi les 30 comptes considérés comme les plus populaires, la plupart des utilisateurs sont en fait les comptes officiels de médias spécialisés en finance ou en économie, comme le *Wall Street Journal*, *Bloomberg* ou le compte Twitter de *Yahoo! Finance*. Il faut noter que malgré le nombre élevé de personnes suivant ces utilisateurs, très peu de messages ont été envoyés au cours de la période où ont été récoltés les tweets.

Tableau 4.2 : utilisateurs les plus populaires

Utilisateur	Nb. Messages sur la période étudiée	Nb. total de messages	Followers sur la période étudiée	Description du compte
nytimes	4	134 000	10 984 206	Where the conversation begins. Follow for breaking news, special reports, http://NYTimes.com homepage links and RTs of our journalists.
TIME	1	83 000	5 404 794	Breaking news and current events from around the globe. Hosted by TIME staff. Tweet questions to our customer service team @TIMEmag_Service.
BBCWorld	1	168 000	5 334 015	News, features and analysis from BBC News (World edition). For UK edition, follow @BBCNews. For breaking news, follow@BBCBreaking. Latest sport news@BBCSport.
WSJ	71	79 900	4 120 489	In our 125th year, breaking news and features from the WSJ. Tweets by @rubinafillion @elanazak@allisonlichter @sarahmarshall@toddjolmstead and @mayaj.
Reuters	5	101 000	3 795 930	Top and breaking news, pictures, and videos from Reuters.
FoxNews	7	175 000	3 569 434	America's Strongest Primetime Lineup Anywhere! Follow America's #1 cable news network, delivering you breaking news, insightful analysis, and must-see videos.
CBSNews	1	56 600	2 955 683	The official twitter feed of CBS News. Follow for original reporting and trusted news
Forbes	43	59 000	2 737 655	Official Twitter account of http://Forbes.com , homepage for the world's business leaders.
Newsweek	3	21 300	2 142 744	Get smarter, faster.
TODAYshow	2	40 700	2 061 980	America's favorite morning show
timoreilly	1	29 700	1 738 781	Founder and CEO, O'Reilly Media. Watching the alpha geeks, sharing their stories, helping the future unfold.
CNBC	155	34 500	1 403 003	First in Business Worldwide
om	7	37 300	1 367 231	Partner, True Ventures. Founder, Gigaom. Lover of possibilities. Believer in people. Indulges in imagineering.
PirryTv	1	12 100	1 317 969	Siganme los buenos. (G STAR)
NBCNews	6	34 800	1 314 539	A leading source of global news and information for more than 75 years. Have a news tip or question? Ask @Rozzy, @Dubois, @JBaiata or@cdellaverson.
BloombergNews	7	69 300	1 311 125	Bloomberg News is the first word in business and finance.
SAI	39	69 000	1 288 527	The latest digital business news from @BusinessInsider
YourAnonNews	1	92 100	1 183 248	Supports digital and AFK activists.
Yahoo	4	37 400	1 153 164	Yahoo's official Twitter, sharing the best of our network. For email help:@YahooCare.
FortuneMagazine	20	26 900	1 149 549	FORTUNE's official Twitter feed. Here's what we're hearing, seeing, reading and writing. Visit us athttp://www.fortune.com
ForbesTech	8	42 800	1 097 594	Tech news and insights from Forbes.
FT	7	57 100	1 055 136	News stories, features and updates from the FT. For headlines follow @financialtimes. Register http://on.ft.com/socialregs for free access to 8 articles p/m.
FastCompany	1	75 700	972 990	Official Twitter feed for the Fast Company business media brand; inspiring readers to think beyond traditional boundaries & create the future of business.
MarketWatch	151	79 300	952 589	Tracking the pulse of the markets. Get business news, personal finance information & commentary. Tweets by @brianaguilar@SAFmedia & others at MarketWatch.
engadget	1	82 900	892 255	The definitive guide to this connected life.
USATODAY	4	85 600	870 286	The latest news and most interesting stories from USA TODAY. News that's meant to be shared.
LILBTHEBASEDGOD	1	127 000	812 169	Mogul, First Rapper Ever To Write And Publish A Book at 19, Film Score, Composer, Producer,Director/Photo/Branding/Marketing/Historical Online Figure #BASED
ReutersBiz	9	73 700	781 415	Top business news around the world. Join us @Reuters, @macroscope, @counterparties,@reutersinsider, @breakingviews
AppSame	1	31 300	776 202	A Conservative Political Marketing Firm helping to bring America back to its greatness https://www.facebook.com/AppSame
Change	5	14 900	767 947	http://Change.org is the world's largest petition platform. Our mission is to empower people everywhere to create the change they want to see.

4.3.2 Utilisateurs les plus « volubiles »

Une seconde mesure de réputation étudiée est celle correspondant au nombre de messages publiés. Le tableau 4.3 représente les utilisateurs ayant envoyé le plus de messages sur Twitter concernant les 400 compagnies étudiées. En se concentrant sur les 30 utilisateurs les plus actifs, on remarque qu'ils correspondent à des robots informatiques retransmettant des nouvelles financières. Bien que ces messages ne peuvent s'apparenter à des *spams*, l'aspect automatique de ces messages ajoute du bruit à l'ensemble. De plus, la plupart des comptes redirige vers des sites tiers proposant une expertise financière qui n'est pas possible de vérifier.

Tableau 4.3 : utilisateurs les plus volubiles

Utilisateur	Nb. Messages sur la période étudiée	Nb. total de messages	Followers	Description du compte	Site(s) de redirection	Création
TheStockHerald	6587	163 000	148	/	top10stocks2buy.com	09/09/2013
GavinGreenberg	6234	165 000	322	/	top10stocks2buy.com	01/10/2013
Pennystocks24	4958	181 000	18 500	Daily Stock News	stocks-news.net et pennystocks24.com	28/02/2014
stocknews247	4514	95 700	751	Stock Market News every Day	stocks-news.net et pennystocks24.com	28/02/2014
snn_team	4316	83 400	339	Breaking news and analysis from the SNN Team	stocks-news.net	18/05/2011
mrtoxo	4282	87 600	1 646	/	stocks-news.net	03/09/2009
The_Louie_Allan	4175	49 100	5	/	top10stocks2buy.com	05/11/2013
stock_newsnet	4146	/	/	Daily free Stock News	/	08/06/2010
stocknews99	4145	82 900	561	Stock Market News	stocks-news.net	10/11/2011
stocknews77	3966	86 000	507	Daily news stocks news	stocks-news.net	02/11/2011
stockwire24	3936	95 600	1 023	Wiring you the most recent and relevant news concerning US stocks. Stay ahead of the market with us, since there's no such thing as a free lunch in economics!	twitter.com/dowbands	19/07/2012
JoeDDSpencer	3818	66 600	133	/	top10stocks2buy.com	04/09/2013
CramerBuffett	3499	/	/	/	/	19/07/2013
MidNorl	3407	75 200	224	/	top10stocks2buy.com	15/03/2012
GreaterStocks	3373	64 800	3	/	top10stocks2buy.com	04/10/2013
EmmanuelGromer	3146	72 300	99	/	top10stocks2buy.com	04/09/2013
AlfredoGilham	3071	74 200	186	/	top10stocks2buy.com	04/09/2013
ForTraders	3005	452 000	4 154	/	4-traders.com	06/01/2010
AmericanBanking	2979	174 000	4 277	Banks, Credit Unions & Financial Institutions	americanbankingnews.com	31/07/2009
CordiaBranam	2900	75 100	177	/	top10stocks2buy.com	22/08/2013
MonteMose	2855	/	/	/	/	31/10/2013
ManInStocks	2665	56 100	6	/	top10stocks2buy.com	27/09/2013
TimeTheExits	2662	56 100	3	/	top10stocks2buy.com	07/10/2013
iHugoAlbert	2641	67 400	139	/	top10stocks2buy.com	08/09/2013
BullishNews	2488	90 700	253	/	top10stocks2buy.com	16/08/2013
WendellBianconi	2436	63 000	84	/	top10stocks2buy.com	04/09/2013
DamienSFaucher	2409	61 800	57	/	top10stocks2buy.com	08/09/2013
micenter	2379	195 000	1 366	The News, Insight, and Intelligence that can make a difference	marketintelligencecenter.com	28/04/2009
toryhollenberg	2314	70 600	123	/	top10stocks2buy.com	27/08/2013
LeonelCannavo	2302	81 900	201	/	top10stocks2buy.com	04/09/2013
mediasentiment	2248	55 300	4 812	Media Sentiment helps you to find the news that matters to you from accounts you follow on Twitter	mediasentiment.com	12/05/2009
SdockNews	2202	68 200	6	/	top10stocks2buy.com	01/10/2013
SeymourAlberta	2183	74 600	162	/	top10stocks2buy.com	04/09/2013
VernSonnenberg	2176	78 100	208	/	top10stocks2buy.com	04/09/2013
TickerReport	2145	57 500	230	Market moving news	tickerreport.com	21/10/2013
CliffatMIG	2115	50 800	108	/	top10stocks2buy.com	04/09/2013
StockNewsForYou	2058	72 400	176	/	top10stocks2buy.com	16/08/2013
VanMacher	1991	57 800	119	/	top10stocks2buy.com	04/09/2013
fin_vestor	1971	89 600	321	/	top10stocks2buy.com	13/02/2013

Les trois graphiques suivants illustrent le nombre de messages envoyés par utilisateur au cours des trois périodes de temps étudiées. Les échelles logarithmiques ont été utilisées afin de faire ressortir le fait qu'une poignée d'utilisateurs émettent la majorité des messages par rapport à l'ensemble des utilisateurs. En effet, 173 utilisateurs représentent 50% des messages envoyés entre 6h30 et 9h30 (soit 0,9%). De plus, 67,5% des utilisateurs n'ont émis qu'un seul message en matinée.

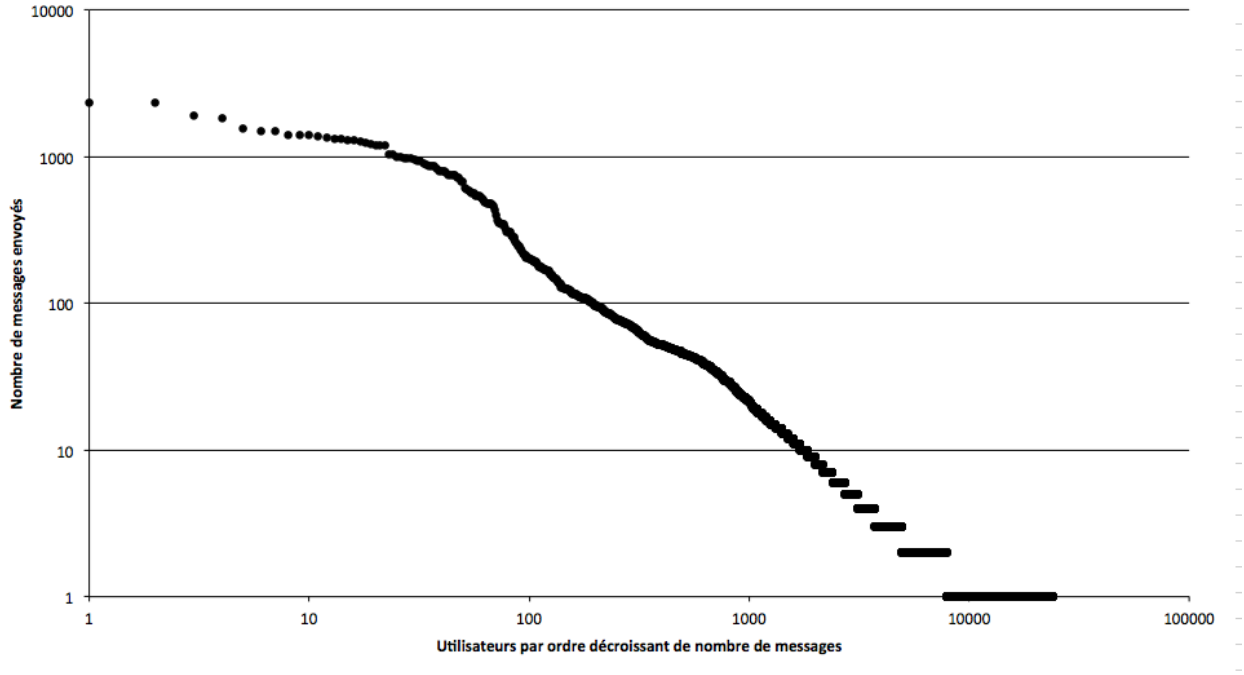

Figure 4.2 : nombre de messages envoyés par utilisateur entre 6h30 et 9h30

Ces proportions sont similaires sur l'heure du midi et le soir, où respectivement 204 utilisateurs (0,7%) et 297 utilisateurs (0,8%) génèrent 50% des messages envoyés. À mi-journée, ce sont 67,7% des utilisateurs qui n'ont émis qu'un seul message sur l'ensemble de la période étudiée et en soirée, ce sont 56,6% des utilisateurs qui n'ont émis qu'un seul message.

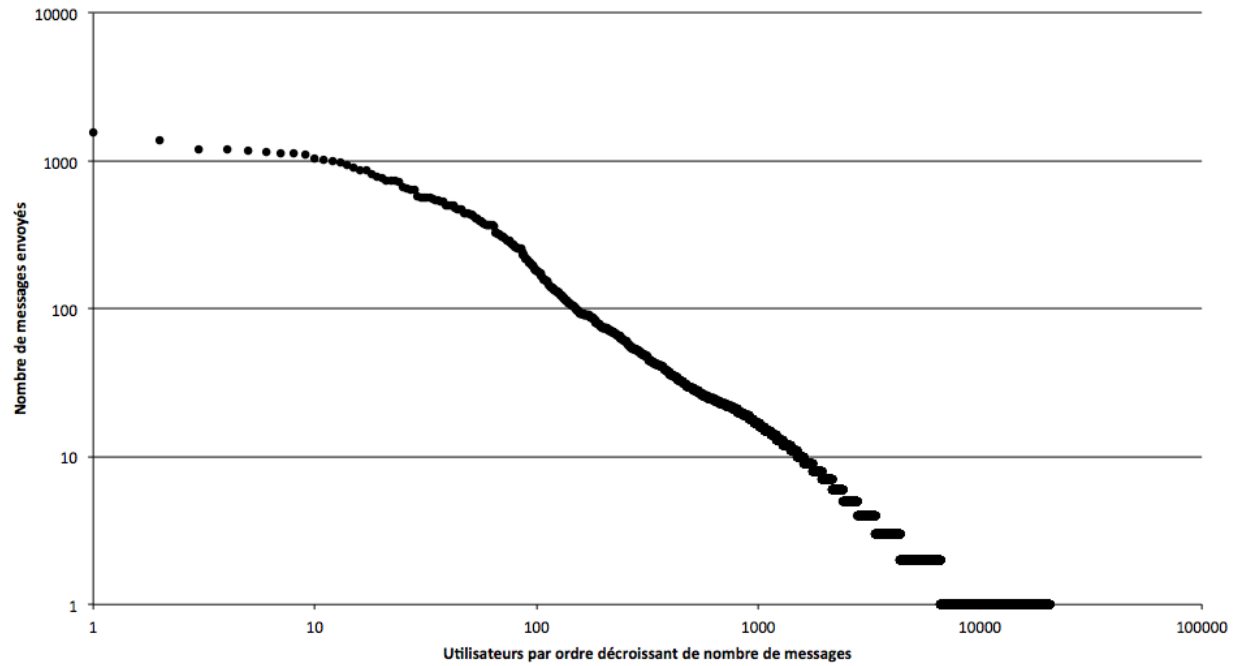

Figure 4.3 : nombre de messages envoyés par utilisateur entre 11h45 et 13h30

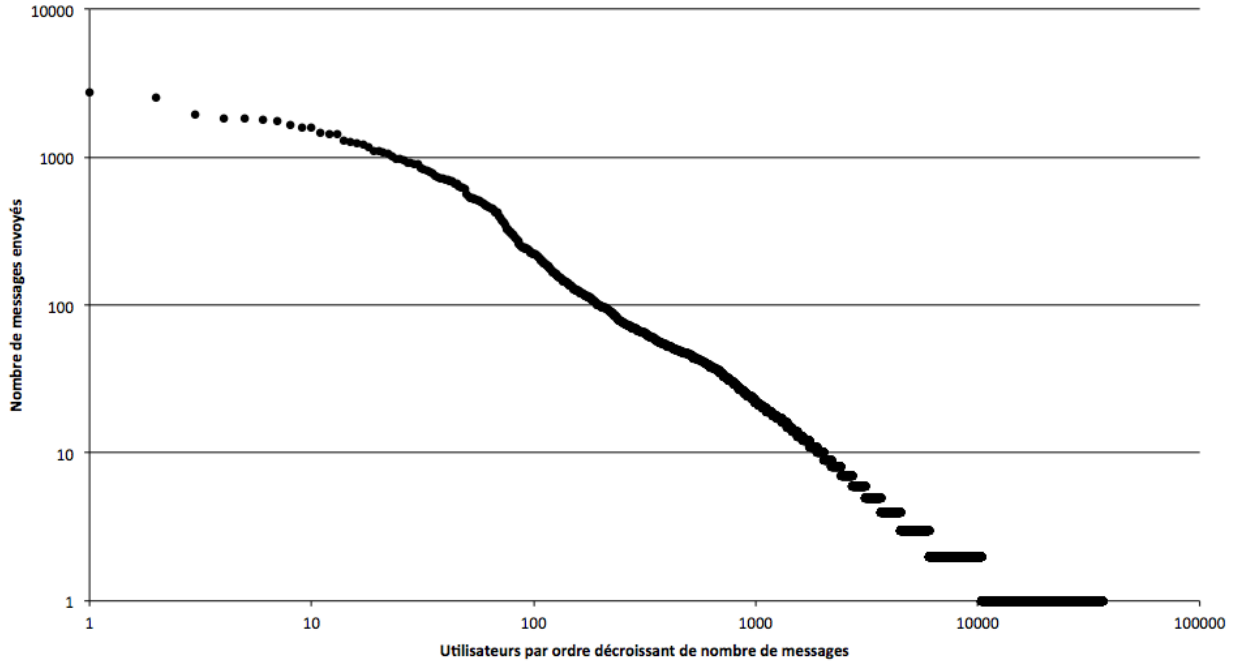

Figure 4.4 : nombre de messages envoyés par utilisateur entre 16h00 et 19h00

4.3.3 Une approche par réseau

La dernière approche envisagée est une approche par réseau. Un utilisateur possède un certain nombre de followers qui verront ses messages apparaître dans leur fil d'actualité. Néanmoins, apparaître dans un fil d'actualité ne signifie pas que les followers vont retransmettre l'information initiale à leur propre réseau. Nous pensons donc que l'action de republier un message est une mesure plus forte afin de mesurer l'impact de la mise en ligne d'un message. À titre d'exemple, le message envoyé par Barack Obama lors de sa réélection en 2012 a été retransmis par plus de 750 000 utilisateurs ; le *selfie* d'Ellen DeGeneres pris lors de la cérémonie des Oscars 2014 a été retransmis par plus de 3,4 millions d'utilisateurs. Ce type de message, appelé retweet (abrévié RT), possède la particularité de comporter le nom de la personne source du message (i.e. « RT @WilliamSanger : $TWTR en baisse de 15% après la vente d'actions des premiers actionnaires »).

En se basant sur cette caractéristique, nous avons filtré l'ensemble du contenu des messages publiés dans notre échantillon pour isoler deux types de données. Dans un premier temps, les émetteurs de messages, et dans un second temps le nom des personnes les retweetant. De cette manière, il est possible de relier une personne mettant en ligne un contenu original et le premier cercle de personnes retransmettant cette information. Ces couples d'utilisateurs permettront de tracer une cartographie du réseau de retransmission des messages.

De la même manière que pour les approches précédentes, nous isolons les 30 profils ayant été retweetés par le plus grand nombre d'utilisateurs différents (tableau 4.4). Cette mesure est basée sur la centralité de degré des utilisateurs, c'est à dire le nombre de liens les reliant entre eux. Les médias financiers font partie des profils d'utilisateur les plus retransmis, mais des profils individuels sont mis en valeur pour la première fois avec cette méthode, notamment Carl Icahn (investisseur), Paul La Monica (journaliste), Carl Quintanilla (journaliste) ou Jim Cramer (journaliste). L'avantage de cette méthode permet de filtrer le bruit occasionné par la publication de trop nombreux messages. Mis à part les comptes @SeekingAlpha, @Benzinga et @LaMonicaBuzz qui ont émis respectivement 1885, 743 et 494 messages pendant la période étudiée, tous les autres comptes ont publié moins de 200 messages.

Tableau 4.4: utilisateurs étant retweetés par le plus d'utilisateurs différents

Utilisateur	Nb. Messages sur la période étudiée	Nb. total de messages	Followers	Description du compte	Nombre de Retweets
philstockworld	17	4 347	480 000	High Finance For Real People - Fun and Profits! Order Stock World Weekly or our PSW Report @ http://tinyurl.com/TwitterOffer7	1 719
wsj	71	79 800	4 400 000	In our 125th year, breaking news and features from the WSJ. Tweets by @rubinafillion @elanazak@allisonlichter @sarahmarshall@toddjolmstead and @mayaj.	892
alperfrx	6	7 556	5 329	Kötü olanları aramızdan çıkardık.Şimdi O'nlara savaş açtık.İşin haricindeki insanlarla yakınlaşma. Güven diye bir şey yoktur.	471
cnbc	155	34 500	1 460 000	First in Business Worldwide	430
forbes	43	58 900	3 020 000	Official Twitter account of http://Forbes.com , homepage for the world's business leaders.	264
cnbcsocial	19	5 350	6 200	Welcome to @CNBC's social media feed -- a forum for all things social. Our newsroom team uses this feed to engage & post social-related content. Tweet at us!	239
marketwatch	151	79 100	1 010 000	Tracking the pulse of the markets. Get business news, personal finance information & commentary. Tweets by @brianaguilar@SAFmedia & others at MarketWatch.	202
cnnmoney	39	58 800	715 000	The world's leading business and finance website and the online home to @FortuneMagazineand @Money Magazine.	186
seekingalpha	1 885	237 000	56 000	Seeking Alpha is the #1 crowd sourced equity research platform. Follow analysis by investors for investors from Seeking Alpha.	180
yahoofinance	221	50 700	328 000	Yahoo Finance is your go-to place for financial news, data, and more! 100% feed-free tweets. Retweets are not necessarily endorsements.	162
carl_c_icahn	4	129	164 000	Chairman of Icahn Enterprises L.P.; etc., etc. Some people get rich studying artificial intelligence. Me, I make money studying natural stupidity.	156
foxbusiness	161	61 300	201 000	The official twitter page of FOX Business Network: Capitalism lives here. Ask your cable provider for FOX Business in your neighborhood.	147
cnbcnow	69	12 500	87 100	CNBC is the recognized world leader in business news, providing real-time financial market coverage & business information. Follow for what's trending.	130
stocktwits	75	41 000	353 000	We created the $ prefix for stocks (e.g. $MSFT). Follow @StockTwits and go to http://stocktwits.com for real-time ideas and stock conversations.	130
lamonicabuzz	494	55 600	35 500	Paul R. La Monica's The Buzz on@CNNMoney. All stocks and economy. All the time. Except when I mix in references to sports and pop culture too.	129
pepsico	4	20 900	137 000	The official home for PepsiCo news, product information and event coverage on Twitter. Tweet us and say hello!	119
gm	29	10 300	239 000	Bringing GM information to Twitter one tweet at a time from@maryhenige (^MH), @philcolley(^PC), @Ternespt (^PT),@psullivan85 (^PS) and@julie_halsey (^JH).	104
carlquintanilla	129	22 700	70 300	Putting it together .. bit by bit.	100
jimcramer	179	41 500	750 000	I am founder of TheStreet and I run the charitable trust portfolio, Action Alerts PLUS. I also host CNBC's Mad Money and blog daily on RealMoney.com. Booyah!	91
srussolillo	77	8 984	7 016	@WSJ reporter; Morning MoneyBeat scribe. Part-time MBA student at Baruch College. Blue Hen for life.	89
mahmut_cebi	1	3 570	22 100	Güzel söz insanı ve hayatı güzelleştirir	86
mistermainst	54	1 018	647	Mr. Main Street | Finance & economic news for the average joe | Dividends Rule Everything Around Me #DREAM | Views are own | RT do not = endorsement	85
bloombergtv	108	44 100	226 000	The official Bloomberg Television Twitter page. Offering the latest global business and markets news, as well as exclusive insights from newsmakers.	84
benzinga	743	62 200	36 100	Stay up to date on breaking news & trading ideas throughout the day. Get real time alerts. Start your free trial of Benzinga Pro: http://bit.ly/TgWmfa	74
andriacheng	26	1 663	4 870	Retail reporter with MarketWatch/WSJ Digital Network, telling not only hard business stories, but also on style, trends and what makes a brand/store tick.	71
businessinsider	37	151 000	537 000	The latest business news and analysis.	69
reformedbroker	59	73 300	74 900	Chairman of the Twitter Federal Reserve, author of the forthcoming book 'Clash of the Financial Pundits'	69
ibdinvestors	152	28 700	55 600	Investor's Business Daily provides leading stock market news and analysis, powerful investing products and education to help investors make money in the market.	64
jboorstin	69	6 169	20 800	Media Reporter. Journalist	63

En utilisant le logiciel de cartographie de réseau Gephi, il est possible de visualiser les relations entre les différents utilisateurs (Bastian et al., 2009). Les graphiques ont été générés en utilisant les algorithmes *ForceAtlas2* (Jacomy et al., 2011). Ces algorithmes repoussent les utilisateurs qui ne possèdent pas de liens entre eux et attirent ceux qui sont reliés, mettant en évidence les regroupements d'utilisateurs.

La figure 4.5 permet de dresser un portrait d'ensemble du maillage des utilisateurs actifs sur la période étudiée. Trois entités se forment au sein de ce réseau avec chacune des propriétés distinctes.

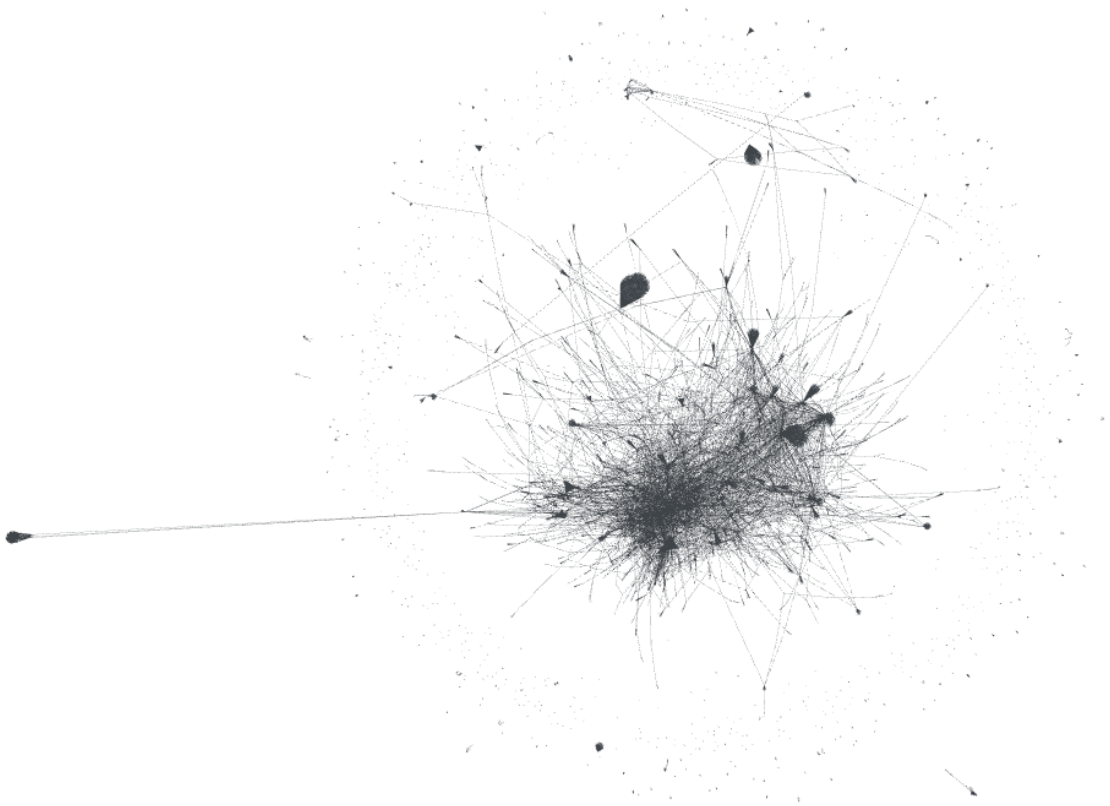

Figure 4.5 : cartographie globale du réseau des messages retransmis

En se basant sur la figure 4.6, on remarque que l'utilisateur le plus retweeté (@PhilStockWorld, cercle en haut de la figure) n'est quasiment pas rattaché aux autres groupes et reste en dehors de la plupart des interactions. Cette dynamique suggère que *PhilStockWorld* émet du contenu qui est retweeté par des utilisateurs, mais ces utilisateurs restent en vase clos par rapport à l'ensemble du réseau.

Le second groupe (cercle à droite de la figure) qui émerge est celui formé par les chaines d'information financières, notamment *CNBC*, avec au centre le compte du *Wall Street Journal* (*@WSJ*). Ce réseau met en évidence des points centraux à travers ce réseau que l'on pourrait qualifié de structuré (*@cnnmoney*, *@cnbcsocail*, *@cnbcnow*, *@cnbc*, avec au centre *@WSJ*).

Finalement, le troisième groupe (cercle en bas de la figure 4.6) est composé par la plupart des utilisateurs, et se caractérise par une agglutination autour de dizaines d'utilisateurs plus connectés. Cette structure suggère une forte interconnectivité entre les utilisateurs, mais ne permet pas

d'identifier de points centraux comme pour le précédent groupe. En d'autres termes, le retrait d'un des points de ce dernier réseau n'empêche pas la propagation de l'information, tandis que le retrait d'un élément des deux premiers groupes modifierait profondément la structure des sous-réseaux.

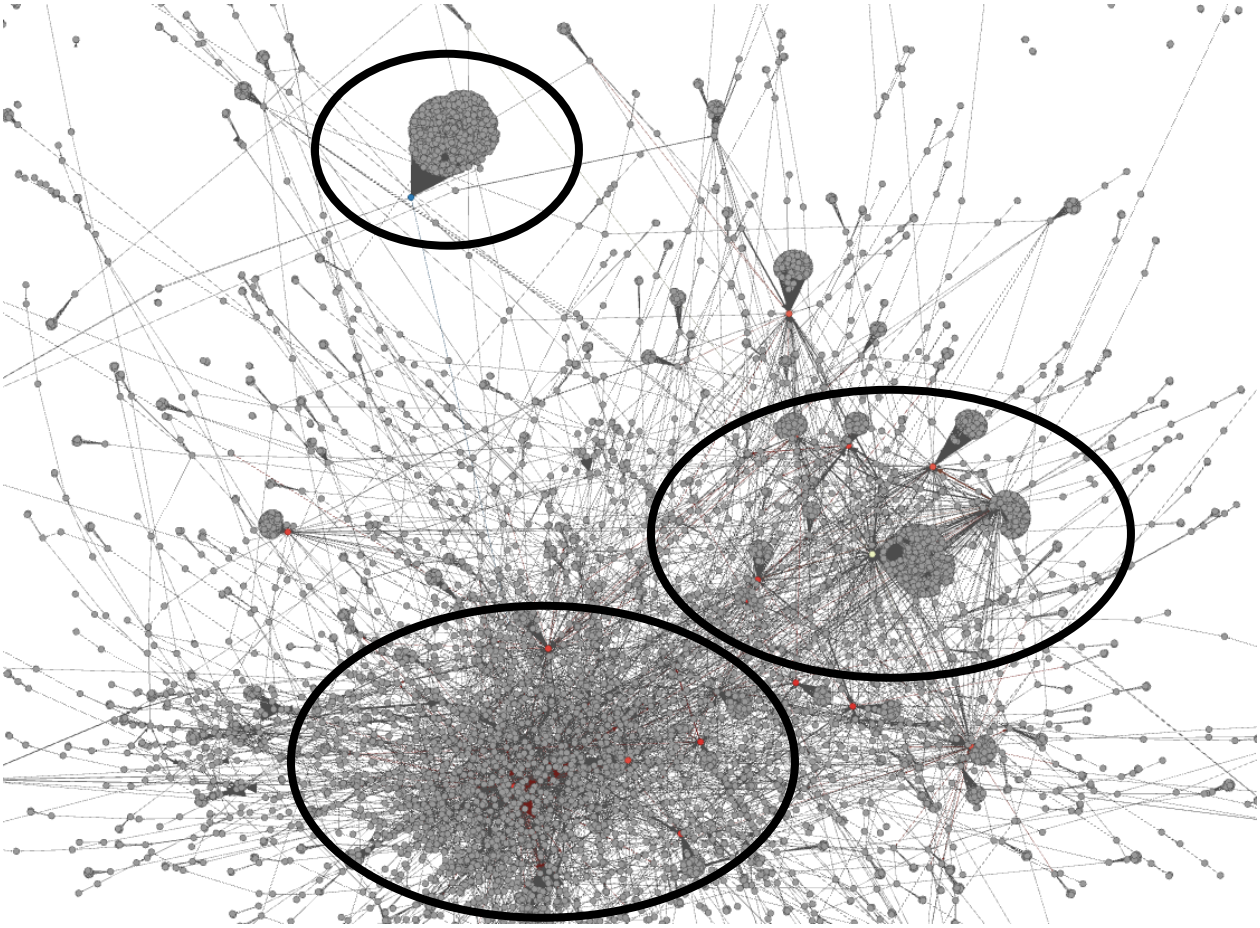

Figure 4.6 : détail du réseau et identification des sous-réseaux

Outre la centralité de degré des utilisateurs, une seconde mesure peut être utilisée, soit la centralité d'intermédiarité. Celle-ci capture l'occurrence d'un nœud à se retrouver sur les chemins les plus courts entre les différents nœuds d'un réseau. Après avoir normalisé le nombre d'occurrence afin d'obtenir des valeurs comparables entre 0 et 1, le tableau 4.5 recense les utilisateurs détenant la centralité d'intermédiarité la plus élevée.

En analysant le type de profils obtenus, on remarque la présence de journalistes financiers, de dirigeants d'entreprises, d'investisseurs et d'analystes, mais aussi de médias financiers. De plus, le faible nombre de followers de certains utilisateurs suggère que ces personnes ne sont pas encore suivies par l'ensemble des utilisateurs de Twitter. Cette dernière caractéristique permet d'obtenir un signal privilégié pour toute personne les suivant.

Tableau 4.5 : utilisateurs ayant une centralité d'intermédiarité la plus élevée

Utilisateurs	Centralité d'intermédiarité	Nombre de followers	Description
lamonicabuzz	1,0000	35 500	Paul R. La Monica's The Buzz on@CNNMoney. All stocks and economy. All the time. Except when I mix in references to sports and pop culture too.
wallstjesus	0,8691	32 700	Son of the Market Gods
bgallo	0,8064	10 100	keep calm and risk on...
rayno	0,6642	5 289	Technology analysis, digital marketing, and markets. http://CMSWire.com and Business Insider contributor. Ex Light Reading, Red Herring.
cnnmoney	0,2291	715 000	The world's leading business and finance website and the online home to @FortuneMagazineand @Money Magazine.
hovastocks	0,2284	323	Investor/Trader
fastmoneylydia	0,1975	6 608	Lydia Thew - Executive Producer Fast Money Halftime Report on CNBC.
tlmontana	0,1973	10 100	Manage funds for clients & self. Interested in quality stocks with liquidity. Enjoy ideas from $$ oriented &serious traders. CHAT ACCESS: madieyeonspy@gmail.com
stocktwitsjohn	0,1680	9 537	StockTwits CEO. Former exec producer of CNBC's Fast Money/Halftime Report & stocks team leader at Bloomberg News. Opinions are my own, but everyone's got one.
stocktwits	0,1648	353 000	We created the $ prefix for stocks (e.g. $MSFT). Follow @StockTwits and go to http://stocktwits.com for real-time ideas and stock conversations.
herbgreenberg	0,1249	61 100	TheStreet, CNBC, journalist, hype-buster, creator/proprietor of the ORIGINAL Hostile React-o-Meter.
yahoofinance	0,1183	328 000	Yahoo Finance is your go-to place for financial news, data, and more! 100% feed-free tweets. Retweets are not necessarily endorsements.
sconsetcapital	0,1152	5 214	Ye Olde Hedge Fund Shoppe 1 & 20. SI HAEC INSOLITA RES VERA EST, QUID EXINDE VERUM EST?
ppearlman	0,1130	21 000	Interactive Editor, Yahoo! Finance
murphyrosecliff	0,1117	4 274	Founder / CEO Rosecliff Capital Husband & Father of 5. CNBC Fast Money Contributor
lifesciencesmkt	0,1114	2 202	Healthcare professionals working in the clinical and non-clinical settings including biotechnology, pharmaceutical, medical devices and life sciences industries
traderstewie	0,1071	23 900	7 Day FREE Trial http://artoftrading.net/ Blog : http://theimpatienttrader.blogspot.com/
mistermainst	0,1046	647	Mr. Main Street | Finance & economic news for the average joe | Dividends Rule Everything Around Me #DREAM | Views are own | RT do not = endorsement
howardlindzon	0,1013	244 000	Chairman/Co-Founder of Stocktwits..GP of Social Leverage (Angel),Wallstrip creator (purchased by $CBS),Momentum, Stocks, Stock Market & LOL hunter..Love Popcorn
swake183	0,1013	60	Active trader who follows the market daily and has a passion for stocks
philstockworld	0,0983	480 000	High Finance For Real People - Fun and Profits! Order Stock World Weekly or our PSW Report @ http://tinyurl.com/TwitterOffer7
louisvillewhale	0,0977	902	Swing and position trading strategies. Stocks, ETFs, mutual funds, options, currencies. Macro & technical focus.
sassy_spy	0,0929	6 151	Trader - Sassy - UCLA and USC Graduate - Contributer to The Street (http://www.thestreet.com/author/1525293/RachelShasha/all.html ...)
optionmonster	0,0902	49 200	Talent hits a target no one else can hit; Genius hits a target no one else can see. I am the Co-founder of optionMONSTER & http://tradeMONSTER.com
paulwoll	0,0724	10 700	Stocks/options trader, investor, & mentor. I work hard to post high quality setups for free. Site in development.
ampressman	0,0669	4 420	Yahoo Finance tech reporter, gadget geek, suburban dad and luckiest husband around. One of Time Magazine's 140 Best Twitter Feeds in 2014.
cnbcfastmoney	0,0657	80 700	Fast Money Halftime Report with @ScottWapnerCNBC at 12P ET. Fast Money with@MelissaLeeCNBC at 5P ET.
mnycx	0,0512	3 752	Stocks/ Options, Sex and Rock & Roll
swissscheese	0,0507	269	use twitter for entertainment. No advice here. Follow me at your own risk. I wouldn't.

4.3.4 Géolocalisation des Tweets

La géolocalisation des messages reste un phénomène rare pour les Tweets à caractère financier. Sur l'ensemble des messages étudiés, seuls 0,5% des messages sont géolocalisés. Les États-Unis, la Finlande, la Hongrie et la République tchèque sont les pays émettant majoritairement ce type de messages (figure 4.7).

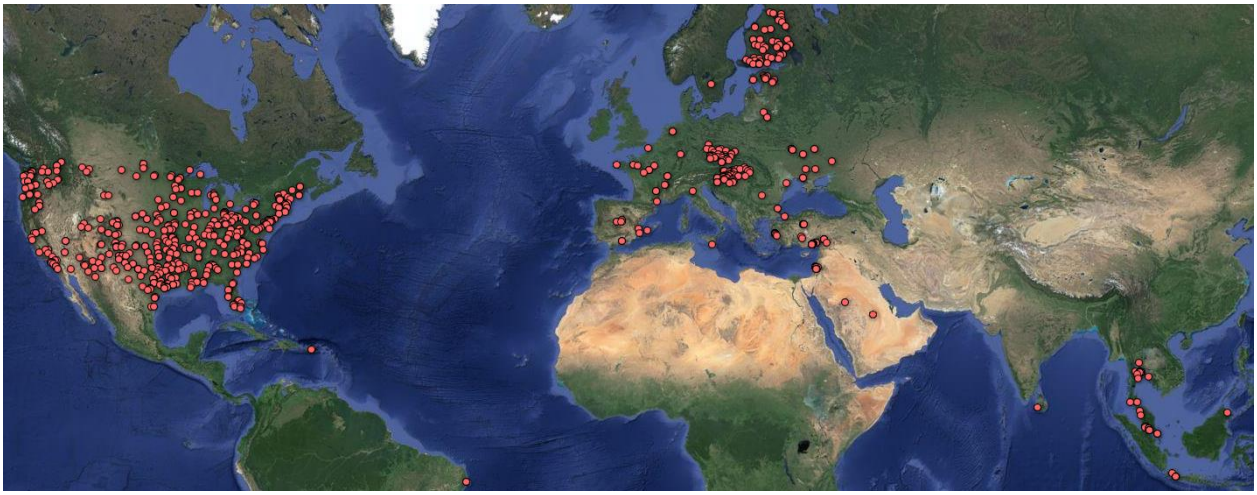

Figure 4.7 : cartographie des messages géolocalisés

Cette approche reste néanmoins prometteuse, car la géolocalisation des messages permettrait d'identifier et de cibler des émetteurs à fort potentiel pour la finance, notamment les personnes travaillant dans les quartiers financiers. Le point faible de cette approche réside toutefois dans le caractère volontaire de la géolocalisation des messages, cette fonction n'étant pas activée par défaut sur les comptes des utilisateurs.

4.4 Pertinence des utilisateurs et recommandations

Les conclusions de cette dernière partie peuvent être résumées par les atouts et les lacunes des différentes approches étudiées précédemment.

1. La première, basée sur le nombre total de followers des comptes Twitter, met en valeur l'audience potentielle d'un message envoyé. En se basant sur cette approche, on peut qualifier Twitter de média financier, permettant la propagation de nouvelles reliées à l'économie.
2. La seconde, basée sur le nombre de messages envoyés, illustre la notion de bruit sur le réseau social, notamment par le comportement automatique des robots. Très peu d'utilisateurs génèrent un nombre élevé de messages, tandis que la grande majorité des utilisateurs ne se contente d'écrire qu'un seul message.
3. La troisième approche modélise les réseaux de propagation des messages et quantifie la centralité des utilisateurs. Ces réseaux fortement tissés mettent en valeur l'interconnectivité des utilisateurs, et expliquent notamment la propagation virale de certains évènements.

4. Finalement, l'approche basée sur la géolocalisation des messages permet de raffiner l'analyse mais reste très marginale, de par son adoption très peu répandue auprès des utilisateurs.

Utiliser la troisième approche apporte ainsi un niveau de lecture supplémentaire par rapport à des mesures de réputation traditionnelles. **Nous pouvons ainsi confirmer l'hypothèse 3.**

Afin d'utiliser efficacement ces conclusions à des fins d'investissements, un investisseur doit (1) suivre les utilisateurs possédant le plus de followers pour connaitre l'information connue par tous et (2) filtrer les messages émis pour éviter de lire les trop nombreux messages envoyés automatiquement. (3) L'avantage comparatif de cet investisseur sera de compléter ces approches en suivant les pivots centraux des réseaux, en se basant sur la centralité d'intermédiarité des utilisateurs et sur le nombre d'utilisateurs transmettant effectivement les messages émis.

4.5 Vers une valorisation des mégadonnées dans l'industrie financière

La dernière partie de ce mémoire propose des pistes afin de maximiser le potentiel des données massives pour l'industrie financière. De quelle manière profiter de cette révolution technologique que constituent les mégadonnées ?

Pour l'investisseur, il est essentiel de ne pas tomber dans le piège naïf qui consiste à ne prendre en compte que les messages publiés sur Twitter. Tout d'abord, la parole de certains utilisateurs occupe une place disproportionnée par rapport à leur apport réel d'information objective (chapitre 4). De plus, les compagnies inscrites en bourse ne sont pas discutées avec la même ampleur sur Twitter. Par une mauvaise méthodologie et en se concentrant sur quelques compagnies, il est facile de retomber dans une spirale de comportements moutonniers. De plus, le but de ce mémoire a été de montrer que les comportements moutonniers peuvent être anticipés, exprimés à travers les médias sociaux. Un investisseur se doit donc de suivre l'ensemble des messages publiés sur la totalité des compagnies. Une analyse sémantique des messages émis complète l'information nécessaire pour traiter rapidement le volume important de messages publiés. Les résultats de l'analyse des médias sociaux doivent être incorporés aux modèles de gestion de risque déjà existant.

Pour les compagnies cotées en bourse, leur réputation devient indéniablement l'actif le plus précieux (De Marcellis-Warin & Teodoresco, 2012). Écouter ce qui se dit sur soi-même doit

devenir une pratique courante au sein des compagnies, car une rumeur, même infondée, peut être dommageable à long terme sur le cours de bourse. Il est nécessaire aussi d'assurer une présence numérique et sociale à travers les comptes Facebook et Twitter (et d'autres), puis de dédier des ressources afin d'alimenter en contenu ces représentations numériques des compagnies. Peu de compagnies communiquent sur leurs performances boursières ni même ne répondent aux messages publiés, pratique qu'il faudrait (re)penser dans un univers de plus en plus connecté. Si des modèles d'investissements sont basés sur les messages décrivant les compagnies, alors ces compagnies auraient avantage à communiquer de manière assidue.

Pour les organismes régulateurs, identifier les nœuds principaux des réseaux afin de soulever les faiblesses potentielles deviendra un objectif à l'avenir. En effet, si 140 caractères constituent un tweet, lorsqu'émis par une personne médiatique l'impact peut être désastreux sur les compagnies. Le tweet falsifié de l'Associated Press en est un parfait exemple. À l'inverse, les prises de position de l'investisseur Carl Icahn sont amplifiées par l'effet des médias sociaux. Lorsque ce dernier acquiert des parts de la compagnie Apple et envoie un tweet confirmant cet achat, les actions passent de 475$ à 494$. Un tel impact peut avoir un effet dévastateur sur une compagnie, notamment lorsque les décisions stratégiques sont remises en cause, et ce en seulement 140 caractères. Cette pratique d'identifier certains nœuds sensibles du réseau peut entrer en contradiction avec la liberté régnant sur les marchés, mais peut aussi éviter de regrettables débordements.

Pour les compagnies d'informations financières, prendre en compte les médias sociaux. Proposer aux clients (individuels ou institutionnels) des outils d'aide à la décision interprétant efficacement les mégadonnées, simples d'utilisation et incorporer dans les modèles d'évaluation de risque, constitue la prochaine étape pour maximiser le potentiel des mégadonnées.

CONCLUSION

La valorisation des mégadonnées apparaît comme un véritable défi pour la finance, et constitue une source d'opportunités sans précédent. Comportements moutonniers, rumeurs, données massives, instantanéité de l'information, stratégies d'investissements… tous les ingrédients sont réunis pour offrir de nouveaux outils venant compléter un arsenal financier déjà ultraperformant.

Ce mémoire s'insère dans un contexte où la recherche scientifique en finance est en (r)évolution. La finance comportementale souligne les faiblesses du postulat de Fama, et les hypothèses de bases du modèle théorique mis au point par Markowitz sont dans la pratique réfutées par le comportement des investisseurs. Les apports technologiques d'Internet et la puissance de calcul performante et abordable font en sorte que de nouvelles approches voient le jour. Parmi celles-ci, l'utilisation des média sociaux occupe une place de plus en plus importante dans la littérature scientifique.

Nous avons décidé de ne pas opter pour des techniques d'apprentissage automatique au profit d'une méthodologie économétrique encore peu employée à travers les travaux de recherche. Ce mémoire s'est donc penché sur l'étude en détails de l'impact financier des messages publiés sur Twitter.

Au premier chapitre, nous avons souligné les travaux effectués en finance utilisant Internet comme source principale de données (moteurs de recherche, forums Internet et médias sociaux). Nous avons présenté trois techniques pouvant être utilisées pour la structuration des mégadonnées (informatique, physique et financière). Ce chapitre s'est conclu sur les opportunités offertes par l'utilisation de ce nouveau type de données, notamment en termes de mise en place de stratégies d'investissement.

Au chapitre deux, nous avons étudié en détail deux types de rendements, les rendements journaliers et les rendements nocturnes, puis nous avons obtenus deux tableaux de bord permettant de maximiser les opportunités de gain suivant le type d'industrie étudié et suivant le jour de la semaine considéré.

Le chapitre trois résume deux autres types de rendements, les rendements anormaux et la variation des volumes d'actions échangées. Nous avons étudié l'impact des rapports annuels sur les quatre rendements abordés, puis nous avons esquissé une stratégie d'investissements potentielle. Finalement, ce chapitre s'est conclu sur une série de recommandations auprès des différents acteurs des marchés financiers.

Le dernier chapitre a pris une trajectoire différente en se concentrant sur le contenu des messages financiers publiés, et surtout sur les interactions entre les utilisateurs. Il apparait que sur les 30000 messages financiers publiés quotidiennement, une fraction de compagnies capte l'ensemble des messages. De plus, une fraction d'utilisateurs volubiles occupe l'espace numérique. Néanmoins, une approche par cartographie de réseau permet de contourner ces biais de sélection en identifiant les nœuds centraux, et par conséquent les personnes dont les messages émis auront le plus grand impact.

Des trois hypothèses de départ, nous pouvons en déduire les observations suivantes :

- **Hypothèse 1 confirmée** dans le cas des rendements journaliers, **non confirmée** dans le cas des rendements nocturnes. Les variables IndiceTicker et IndiceName impactent différemment les rendements étudiés à travers ce mémoire.
- **Hypothèse 2 confirmée**. Les tableaux de bord peuvent servir d'outils d'aide à la décision en vue de maximiser les opportunités de gain des investisseurs.
- **Hypothèse 3 confirmée**. Une approche utilisant les réseaux d'utilisateur permet de mettre en valeur les nœuds sensibles d'un réseau, une mesure plus fiable que celles impliquant le nombre de followers ou le nombre de messages publiés.

Toutefois, plusieurs limitations sont inhérentes à notre étude. En premier lieu, le faible R^2 de nos simulations montre que les rendements boursiers sont dans l'absolu peu influencés par les messages publiés sur Twitter. Néanmoins, la compréhension des mécanismes entre médias sociaux et finance apporte un complément d'information pouvant faire la différence lorsqu'incorporée dans les modèles d'évaluation des risques actuellement utilisés. La seconde limitation que nous notons concerne notre échantillon, constitué de 71 compagnies. Nous n'avons étudié que l'impact sur les compagnies les plus discutées sur Twitter (possédant en moyenne 30 messages pas jour). Finalement, la stratégie d'investissement que nous proposons reste simple et pourrait être raffinée afin de prendre en compte les éléments traditionnels des stratégies d'investissements.

Sans constituer une recette miracle à suivre, la contribution majeure de cette étude réside avant tout dans son caractère méthodologique. Les techniques économétriques mises en avant ont permis d'appréhender rationnellement les médias sociaux dans un cadre financier. Une des forces de cette approche est la possibilité de modifier et d'adapter les points focaux des modèles (journée du mardi

et secteur des technologies de l'information au cours de cette étude). Ainsi, craindre que de tels résultats ne deviennent obsolètes par leur adoption de la part d'autres acteurs de la finance n'est plus un risque, mais plutôt une donnée supplémentaire à prendre en compte pour modéliser les comportements.

Afin de compléter cette étude et pour répondre aux limitations mentionnées, plusieurs pistes de recherche sont possibles. Tout d'abord, agrandir l'échantillon de compagnies à l'ensemble du S&P500, ou du moins refaire cette étude avec des compagnies peu discutées sur Twitter. La compréhension de ces mécanismes apporterait un supplément d'information déterminant pour les compagnies ne se trouvant pas sur le devant la scène médiatique. Ensuite, prendre en compte la notion de bris structurel dans les données des médias sociaux. Si après modélisation les compagnies adoptent un comportement différent sur les médias sociaux, alors les modèles prédictifs deviendront obsolètes et doit être actualisés. Finalement, il serait intéressant de se pencher sur d'autres marchés, notamment en Europe, ou de manière plus ambitieuse en Chine avec l'utilisation de Weibo.

L'investisseur 2.0 (ou 3.0) devra prendre en compte les médias sociaux, c'est indéniable. La valorisation de l'information par le PRISM des mégadonnées changera le visage de la finance, et cette nouvelle ressource, ce nouvel or noir numérique, l'*or-bits*, constituera l'innovation radicale de processus du début du XXI^{è} siècle.

BIBLIOGRAPHIE

Akerlof, G. A., & Shiller, R. J. (2009). *Animal Spirits: How Human Psychoology Drives the Economy, and Why It Matters for Global Capitalism*. États-Unis: Princeton University Press.

Altshuler, Y., Pan, W., & Pentland, A. (Sandy). (2012). Trends Prediction Using Social Diffusion Models. In S. J. Yang, A. M. Greenberg, & M. Endsley (Eds.), *Social Computing, Behavioral - Cultural Modeling and Prediction* (pp. 97–104). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-29047-3_12

Antonina Kloptchenko, T. E. (2004). Combining data and text mining techniques for analysing financial reports. *Int. Syst. in Accounting, Finance and Management*, *12*, 29–41.

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise ? The Information Content of Internet Stock Message Boards. *Journal of Finance*, 1259–1294.

Bahng, J. S. (2004). Structural Breaks and the Normality of Stock Returns. *Swiss Journal of Economics and Statistics (SJES)*, *140*(II), 207–227.

Bank, M., Larch, M., & Peter, G. (2010). *Google Search Volume and its Influence on Liquidity and Returns of German Stocks* (SSRN Scholarly Paper No. ID 1666763). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1666763

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1310–1319). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145569

Barbera, P. (2014). streamR: Access to Twitter Streaming API via R (Version 0.2.1). Retrieved from http://cran.r-project.org/web/packages/streamR/index.html

Barbosa, L., & Feng, J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 36–44). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1944566.1944571

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *ICWSM*. Retrieved from https://gephi.org/users/publications/

Bifet, A., & Frank, E. (2010). Sentiment Knowledge Discovery in Twitter Streaming Data. In *Proceedings of the 13th International Conference on Discovery Science* (pp. 1–15). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=1927300.1927301

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, Pages 1–8. doi:10.1016/j.jocs.2010.12.007

Bollen, J., Pepe, A., & Mao, H. (2009). *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena* (arXiv e-print No. 0911.1583). Retrieved from http://arxiv.org/abs/0911.1583

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web Search Queries Can Predict Stock Market Volumes. *PLoS ONE*, *7*(7), e40014. doi:10.1371/journal.pone.0040014

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Presented at the Seventh International World-Wide Web Conference (WWW 1998), Brisbane, Australia. Retrieved from http://ilpubs.stanford.edu:8090/361/

Brink, R. V. D., Rusinowska, A., & Steffen, F. (2011). *Measuring Power and Satisfaction in Societies with Opinion Leaders : An Axiomatization* (Université Paris1 Panthéon-Sorbonne (Post-Print and Working Papers) No. halshs-00587726). HAL. Retrieved from http://ideas.repec.org/p/hal/cesptp/halshs-00587726.html

Brown, E. (2012). Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market. *SAIS 2012 Proceedings*. Retrieved from http://aisel.aisnet.org/sais2012/7

Buechel, B., Hellmann, T., & Kloessner, S. (2012). *Opinion Dynamics under Conformity* (Working Paper No. 469). Bielefeld University, Center for Mathematical Economics. Retrieved from http://ideas.repec.org/p/bie/wpaper/469.html

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. In *in ICWSM '10: Proceedings of international AAAI Conference on Weblogs and Social*.

Chancharat, S., Kamalian, A. R., & Valadkhani, A. (2009). *Random Walk and Multiple Structural Breaks In Thai Stock Market* (MPRA Paper No. 50395). University Library of Munich, Germany. Retrieved from http://ideas.repec.org/p/pra/mprapa/50395.html

Chen, R., & Lazer, M. (2011). *Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement*. Stanford.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*, 2–9. doi:10.1111/j.1475-4932.2012.00809.x

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. doi:10.1023/A:1022627411411

Da, Z., Engelberg, J., & Gao, P. (2011). *In Search of Fundamentals* (SSRN Scholarly Paper No. ID 1589805). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1589805

Das, S. R., & Chen, M. Y. (2001). *Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web* (SSRN Scholarly Paper No. ID 276189). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=276189

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, *53*(9), 1375–1388. doi:10.1287/mnsc.1070.0704

De Choudhury, M., Sundaram, H., John, A., & Seligmann, D. D. (2008). Can blog communication dynamics be correlated with stock market activity? In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia* (pp. 55–60). New York, NY, USA: ACM. doi:10.1145/1379092.1379106

De Marcellis-Warin, N., & Teodoresco, S. (2012). Corporate Reputation: Is Your Most Strategic Asset at Risk? *CIRANO Burgundy Report, 2012RB-02*.

Devitt, A., & Ahmad, K. (2007). Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Dimpfl, T., & Jank, S. (2011). *Can internet search queries help to predict stock market volatility?* (CFR Working Paper No. 11-15). University of Cologne, Centre for Financial Research (CFR). Retrieved from http://ideas.repec.org/p/zbw/cfrwps/1115.html

Domingos, P., & Richardson, M. (2001). Mining the Network Value of Customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 57–66). New York, NY, USA: ACM. doi:10.1145/502512.502525

Engle, R. F., & Ng, V. K. (1991). *Measuring and Testing the Impact of News on Volatility* (Working Paper No. 3681). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w3681

Facebook and the Brave New World of Social Research using Big Data | The Policy and Internet Blog. (n.d.). Retrieved from http://blogs.oii.ox.ac.uk/policy/facebook-and-the-brave-new-world-of-social-research-using-big-data/

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work*. *The Journal of Finance*, *25*(2), 383–417. doi:10.1111/j.1540-6261.1970.tb00518.x

Fung, G. P. C., Yu, J. X., & Lam, W. (2002). News Sensitive Stock Trend Prediction. In M.-S. Chen, P. S. Yu, & B. Liu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 481–493). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/3-540-47887-6_48

Gabaix, X. (2008). *Power Laws in Economics and Finance* (Working Paper No. 14299). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14299

Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature*, *423*(6937), 267–270. doi:10.1038/nature01624

Gentry, J. (2013). twitteR: R based Twitter client (Version 1.1.7). Retrieved from http://cran.r-project.org/web/packages/twitteR/index.html

Gidofalvi, G., & Gidófalvi, G. (2001). *Using News Articles to Predict Stock Price Movements*.

Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report*.

Gopikrishnan, P., Plerou, V., Amaral, L. A. N., Meyer, M., & Stanley, H. E. (1999). Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, *60*(5), 5305–5316. doi:10.1103/PhysRevE.60.5305

Gopikrishnan, P., Plerou, V., Gabaix, X., & Stanley, H. E. (2000). Statistical Properties of Share Volume Traded in Financial Markets. *Physical Review E*, *62*(4), R4493–R4496. doi:10.1103/PhysRevE.62.R4493

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 78–87). New York, NY, USA: ACM. doi:10.1145/1081870.1081883

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, *11*(1), 10–18. doi:10.1145/1656274.1656278

Jacomy, M., Heymann, S., Venturini, T., & Bastian, M. (2011). ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization. Retrieved from http://www.medialab.sciences-po.fr/fr/blog/forceatlas2-a-graph-layout-algorithm-for-handy-network-visualization/

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98* (pp. 137–142). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/BFb0026683

Karabulut, Y. (2011). *Can Facebook Predict Stock Market Activity?* (SSRN Scholarly Paper No. ID 1919008). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1919008

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the Spread of Influence Through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 137–146). New York, NY, USA: ACM. doi:10.1145/956750.956769

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 201320040. doi:10.1073/pnas.1320040111

Kreps, D., & Wilson, R. (1982). Reputation and Imperfect Information. *Journal of Economic Theory*, *27*, 253–279.

Leavitt, A., Burchard, E., Fisher, D., & Gilbert, S. (2009). *The Influentials: new approaches for analyzing influence on Twitter*. Retrieved March 19, 2013, from http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, *47*(1), 13. doi:10.2307/1924119

Lux, T. (1996). The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks. *Applied Financial Economics*, *6*(6), 463–475. doi:10.1080/096031096333917

Mackintosh, J., & Editor, I. (2012, May 24). Last tweet for Derwent's Absolute Return. *Financial Times*. Retrieved from http://www.ft.com/intl/cms/s/0/d5d9c3f8-a5bf-11e1-b77a-00144feabdc0.html#axzz36t8RqRsp

Malakian, A. (2013). *Sentiment Analysis Still Has a Long Way to Go on Wall Street*. *www.waterstechnology.com*. Retrieved July 8, 2014, from http://www.waterstechnology.com/buy-side-technology/opinion/2250200/sentiment-analysis-still-has-a-long-way-to-go-on-wall-street

Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives*, *17*(1), 59–82. doi:10.1257/089533003321164958

Mao, H., Counts, S., & Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *arXiv:1112.1051*. Retrieved from http://arxiv.org/abs/1112.1051

Marcellis-Warin, N. de, & Peignier, I. (2012). *Perception des risques au Québec – Baromètre CIRANO 2012* (CIRANO Monographs). CIRANO. Retrieved from http://ideas.repec.org/b/cir/cirmon/2012mo-02.html

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011).

Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 227–236). New York, NY, USA: ACM. doi:10.1145/1978942.1978975

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance, 7*(1), 77–91.

Milgrom, P., & Roberts, J. (1982a). Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis. *Econometrica, 50,* 443–460.

Milgrom, P., & Roberts, J. (1982b). Predation, Reputation and Entry Deterrence. *Journal of Economic Theory, 27*.

Mittal, A., & Goel, A. (n.d.). Stock Prediction Using Twitter Sentiment Analysis. Retrieved from http://tomx.inf.elte.hu/twiki/pub/Tudas_Labor/2012Summer/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

Mittermayer, M.-A. (2004). Forecasting Intraday stock price trends with text mining techniques. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004* (p. 10 pp.–). doi:10.1109/HICSS.2004.1265201

Moon, G.-H., & Yu, W.-C. (2010). Volatility Spillovers between the US and China Stock Markets: Structural Break Test with Symmetric and Asymmetric GARCH Approaches. *Global Economic Review*, *39*(2), 129–149.

Mossin, J. (1966). Equilibrium In A Capital Asset Market. *Econometrica, 34*(4), 768–783.

Nakajima, S., Tatemura, J., Hino, Y., Hara, Y., & Tanaka, K. (2005). Discovering Important Bloggers based on Analyzing Blog Threads.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance, 6*(3), 144–160.

Pan, W., Altshuler, Y., & Pentland, A. (Sandy). (2012). Decoding Social Influence and the Wisdom of the Crowd in Financial Trading Network (pp. 203–209). IEEE. doi:10.1109/SocialCom-PASSAT.2012.133

Pan, W., Cebrian, M., Dong, W., Kim, T., Fowler, J., & Pentland, A. (2010). Modeling Dynamical Influence in Human Interaction Patterns. *arXiv:1009.0240 [physics]*. Retrieved from http://arxiv.org/abs/1009.0240

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *In Proceedings of the ACL* (pp. 271–278).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10* (pp. 79–86). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1118693.1118704

Pentland, A. (2014). *Social Physics - How Good Ideas Spread - The Lessons From A New Science*. The Penguin Press. Retrieved from http://www.penguin.ca/nf/Book/BookDisplay/0,,9781594205651,00.html

Plerou, V., Gopikrishnan, P., Gabaix, X., & Stanley, H. E. (2001). *Quantifying Stock Price Response to Demand Fluctuations* (arXiv e-print No. cond-mat/0106657). Retrieved from http://arxiv.org/abs/cond-mat/0106657

Plerou, V., Gopikrishnan, P., Nunes Amaral, L. A., Gabaix, X., & Eugene Stanley, H. (2000). Economic fluctuations and anomalous diffusion. *Physical Review E*, *62*(3), R3023–R3026. doi:10.1103/PhysRevE.62.R3023

Porshnev, A., Redkin, I., & Shevchenko, A. (2013). *Improving Prediction of Stock Market Indices by Analyzing the Psychological States of Twitter Users* (SSRN Scholarly Paper No. ID 2368151). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2368151

Preis, T., Reith, D., & Stanley, E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data | Explore | Research | WBS. *Philosophical Transactions of the Royal Society*, pp. 5707–5719.

Preis, T., Schneider, J. J., & Stanley, H. E. (2011). Switching processes in financial markets. *Proceedings of the National Academy of Sciences*, *108*(19), 7674–7678. doi:10.1073/pnas.1019484108

Ramos, S. B., Veiga, H., & Latoeiro, P. (2013). *Predictability of stock market activity using Google search queries* (Statistics and Econometrics Working Paper No. ws130605). Universidad Carlos III, Departamento de Estadística y Econometría. Retrieved from http://ideas.repec.org/p/cte/wsrepe/ws130605.html

Rubin, A., & Rubin, E. (2009). *Informed Investors and the Internet* (SSRN Scholarly Paper No. ID 1677703). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1677703

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 513–522). New York, NY, USA: ACM. doi:10.1145/2124295.2124358

Schumaker, R. P., & Chen, H. (2010). *A Discrete Stock Price Prediction Engine Based on Financial News*.

Selten, R. (1975). Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory*, *4*, 25–55.

Seo, Y., Giampapa, J., & Sycara, K. (2002). *Text Classification for Intelligent Portfolio Management*. Retrieved from http://citeseer.ist.psu.edu/seo02text.html

Sharpe, W. F. (1963). A Simplified Model for Portfolio Analysis. *Management Science*, *9*(2), 277–293. doi:10.1287/mnsc.9.2.277

Smailovic, J., Grcar, M., & Znidarsic, M. (n.d.). Sentiment analysis on tweets in a financial domain. Retrieved from http://ipssc.mps.si/papers/smailovic-paper.pdf

Song, X., Chi, Y., Hino, K., & Tseng, B. (2007). Identifying Opinion Leaders in the Blogosphere. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (pp. 971–974). New York, NY, USA: ACM. doi:10.1145/1321440.1321588

Sprenger, T., & Welpe, I. (2010). *Tweets and Trades: The Information Content of Stock Microblogs* (SSRN Scholarly Paper No. ID 1702854). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1702854

Stanley, H. E., Buldyrev, S. V., Franzese, G., Havlin, S., Mallamace, F., Kumar, P., … Preis, T. (2010). Correlated randomness and switching phenomena. *Physica A: Statistical Mechanics and Its Applications*, *389*(15), 2880–2893. doi:10.1016/j.physa.2010.02.023

Tapiero, C. S. (2013). *The Future of Financial Engineering* (SSRN Scholarly Paper No. ID 2259232). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2259232

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, *63*(3), 1437–1467. doi:10.1111/j.1540-6261.2008.01362.x

Timmermann, A. (2001). *Structural Breaks, Incomplete Information and Stock Prices* (University of California at San Diego, Economics Working Paper Series No. qt1sn269d7). Department of Economics, UC San Diego. Retrieved from http://ideas.repec.org/p/cdl/ucsdec/qt1sn269d7.html

Tumarkin, R., & Whitelaw, R. (2001, May). News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal*, pp. 41–50.

Varian, H. R. (2014a). Beyond Big Data. *Business Economics*, *49*(1), 27–31. doi:10.1057/be.2014.1

Varian, H. R. (2014b). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. doi:10.1257/jep.28.2.3

Vincent, A., & Armstrong, M. (2010). *Predicting Break-Points in Trading Strategies with Twitter* (SSRN Scholarly Paper No. ID 1685150). Rochester, NY: Social Science Research Network. Retrieved from https://docs.google.com/viewer?url=http://www.finanzaonline.com/forum/attachments/forex/1499783d1319814776-oibo-ii-la-disfatta-twitter-ts.pdf&chrome=true

Warin, T., De Marcellis-Warin, N., Troadec, A., Sanger, W., & Nembot, B. (2014). *Un état des lieux sur les données massives* (No. 2014RB-01). Montréal, Canada: CIRANO. Retrieved from http://www.cirano.qc.ca/pdf/publication/2014RB-01.pdf

Warin, T., Marcellis-Warin, N. de, Sanger, W., Nembot, B., & Mirza, V. H. (2013). *Corporate Reputation and Social Media: A Game Theory Approach* (CIRANO Working Paper No. 2013s-18). CIRANO. Retrieved from http://ideas.repec.org/p/cir/cirwor/2013s-18.html

Warin, T., & Sanger, W. (2014). Structuring Big Data : How Financial Models may Help. CIRANO.

Warin, T., Sanger, W., & Troadec, A. (n.d.). Élections au Québec sur Twitter. CIRANO: Élections. Retrieved from www.elections.cirano.qc.ca

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 261–270). New York, NY, USA: ACM. doi:10.1145/1718487.1718520

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1220575.1220619

Wysocki, P. (1998, November). Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards by Peter D. Wysocki :: SSRN. *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=160170

Yang, Y., & Liu, X. (1999). A Re-examination of Text Categorization Methods. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 42–49). New York, NY, USA: ACM. doi:10.1145/312624.312647

Zhang, D., & Zhou, L. (Nov.). Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *34*(4), 513–522. doi:10.1109/TSMCC.2004.829279

Zhang, W., Shen, D., Zhang, Y., & Xiong, X. (2013). Open source information, investor attention, and asset pricing. *Economic Modelling*, *33*(C), 613–619.

Zhang, W., & Skiena, S. (n.d.). *Trading Strategies To Exploit News Sentiment*.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear.” *Procedia - Social and Behavioral Sciences*, *26*, 55–62. doi:10.1016/j.sbspro.2011.10.562

ANNEXE A : LISTE DES REQUÊTES UTILISÉES POUR LES COMPAGNIES DU S&P500

Tableau A.1 : liste des requêtes utilisées pour les compagnies du S&P500

ID	Compagnie	Ticker	Name	Secteur industriel
1	3M Co.	$MMM	3m	Industrials
2	Abbott Laboratories	$ABT	abbott	Health Care
3	AbbVie	$ABBV	abbvie	Health Care
4	Abercrombie & Fitch Company A	$ANF	abercrombie	Consumer Discretionary
5	ACE Limited	$ACE	ace limited	Financials
6	Accenture plc	$ACN	accenture	Information Technology
7	Actavis Inc	$ACT	actavis	Health Care
8	Adobe Systems Inc	$ADBE	adobe	Information Technology
9	ADT Corp	$ADT	adt corp	Industrials
10	Advanced Micro Devices	$AMD	amd	Information Technology
11	AES Corp	$AES	aes corp	Utilities
12	Aetna Inc	$AET	aetna	Health Care
13	AFLAC Inc	$AFL	aflac	Financials
14	Agilent Technologies Inc	$A	agilent	Health Care
15	AGL Resources Inc.	$GAS	agl resources	Utilities
16	Air Products & Chemicals Inc	$APD	air products	Materials
17	Airgas Inc	$ARG	airgas	Materials
18	Akamai Technologies Inc	$AKAM	akamai	Information Technology
19	Alcoa Inc	$AA	alcoa	Materials
20	Alexion Pharmaceuticals	$ALXN	alexion	Health Care
21	Allegheny Technologies Inc	$ATI	allegheny	Materials
22	Allergan Inc	$AGN	allergan	Health Care
23	Allstate Corp	$ALL	allstate	Financials
24	Altera Corp	$ALTR	altera	Information Technology
25	Altria Group Inc	$MO	altria	Consumer Staples
26	Amazon.com Inc	$AMZN	amazon	Consumer Discretionary
27	Ameren Corp	$AEE	ameren	Utilities
28	American Electric Power	$AEP	american electric	Utilities
29	American Express Co	$AXP	american express	Financials
30	American Intl Group Inc	$AIG	american intl	Financials
31	American Tower Corp A	$AMT	american tower	Financials
32	Ameriprise Financial	$AMP	ameriprise	Financials
33	AmerisourceBergen Corp	$ABC	amerisourcebergen	Health Care
34	Amgen Inc	$AMGN	amgen	Health Care
35	Amphenol Corp A	$APH	amphenol	Industrials
36	Anadarko Petroleum Corp	$APC	anadarko	Energy
37	Analog Devices, Inc.	$ADI	analog devices	Information Technology
38	Aon plc	$AON	aon	Financials
39	Apache Corporation	$APA	apache	Energy
40	Apartment Investment & Mgmt	$AIV	apartment investment	Financials
41	Apollo Group Inc	$APOL	apollo group	Consumer Discretionary
42	Apple Inc.	$AAPL	apple	Information Technology
43	Applied Materials Inc	$AMAT	applied materials	Information Technology
44	Archer-Daniels-Midland Co	$ADM	adm	Consumer Staples
45	Assurant Inc	$AIZ	assurant inc	Financials
46	AT&T Inc	$T	at&t	Telecommunications Services
47	Autodesk Inc	$ADSK	autodesk	Information Technology
48	Automatic Data Processing	$ADP	automatic data processing	Information Technology
49	AutoNation Inc	$AN	autonation	Consumer Discretionary
50	AutoZone Inc	$AZO	autozone	Consumer Discretionary

ID	Compagnie	Ticker	Name	Secteur industriel
51	AvalonBay Communities, Inc.	$AVB	avalonbay	Financials
52	Avery Dennison Corp	$AVY	avery dennison	Industrials
53	Avon Products	$AVP	avon	Consumer Staples
54	Baker Hughes Inc	$BHI	baker hughes	Energy
55	Ball Corp	$BLL	ball corp	Materials
56	Bank of America Corp	$BAC	bank of america	Financials
57	The Bank of New York Mellon Corp.	$BK	bny mellon	Financials
58	Bard (C.R.) Inc.	$BCR	bard	Health Care
59	Baxter International Inc.	$BAX	baxter	Health Care
60	BB&T Corporation	$BBT	bb&t	Financials
61	Beam Inc.	$BEAM	beam inc	Consumer Staples
62	Becton Dickinson	$BDX	becton	Health Care
63	Bed Bath & Beyond	$BBBY	bed bath beyond	Consumer Discretionary
64	Bemis Company	$BMS	bemis	Materials
65	Berkshire Hathaway	$BRK.B	berkshire	Financials
66	Best Buy Co. Inc.	$BBY	bestbuy	Consumer Discretionary
67	BIOGEN IDEC Inc.	$BIIB	biogen	Health Care
68	BlackRock	$BLK	blackrock	Financials
69	Block H&R	$HRB	block h&r	Financials
70	BMC Software	$BMC	bmc software	Information Technology
71	Boeing Company	$BA	boeing	Industrials
72	BorgWarner	$BWA	borgwarner	Consumer Discretionary
73	Boston Properties	$BXP	boston properties	Financials
74	Boston Scientific	$BSX	boston scientific	Health Care
75	Bristol-Myers Squibb	$BMY	bristol myers	Health Care
76	Broadcom Corporation	$BRCM	broadcom	Information Technology
77	Brown-Forman Corporation	$BF.B	brown forman	Consumer Staples
78	C. H. Robinson Worldwide	$CHRW	robinson worldwide	Industrials
79	CA, Inc.	$CA	ca	Information Technology
80	Cablevision Systems Corp.	$CVC	cablevision systems	Consumer Discretionary
81	Cabot Oil & Gas	$COG	cabot	Energy
82	Cameron International Corp.	$CAM	cameron international	Energy
83	Campbell Soup	$CPB	campbell	Consumer Staples
84	Capital One Financial	$COF	capital one	Financials
85	Cardinal Health Inc.	$CAH	cardinal health	Health Care
86	Carefusion	$CFN	carefusion	Health Care
87	Carmax Inc	$KMX	carmax	Consumer Discretionary
88	Carnival Corp.	$CCL	carnival corp	Consumer Discretionary
89	Caterpillar Inc.	$CAT	caterpillar	Industrials
90	CBRE Group	$CBG	cbre group	Financials
91	CBS Corp.	$CBS	cbs	Consumer Discretionary
92	Celgene Corp.	$CELG	celgene	Health Care
93	CenterPoint Energy	$CNP	centerpoint	Utilities
94	CenturyLink Inc	$CTL	centurylink	Telecommunications Services
95	Cerner	$CERN	cerner	Health Care
96	CF Industries Holdings Inc	$CF	cf industries	Materials
97	Charles Schwab	$SCHW	charles schwab	Financials
98	Chesapeake Energy	$CHK	chesapeake	Energy
99	Chevron Corp.	$CVX	chevron	Energy
100	Chipotle Mexican Grill	$CMG	chipotle mexican	Consumer Discretionary

ID	Compagnie	Ticker	Name	Secteur industriel
101	Chubb Corp.	$CB	chubb	Financials
102	CIGNA Corp.	$CI	cigna	Health Care
103	Cincinnati Financial	$CINF	cincinnati financial	Financials
104	Cintas Corporation	$CTAS	cintas	Industrials
105	Cisco Systems	$CSCO	cisco	Information Technology
106	Citigroup Inc.	$C	citigroup	Financials
107	Citrix Systems	$CTXS	citrix	Information Technology
108	Cliffs Natural Resources	$CLF	cliffs natural	Materials
109	The Clorox Company	$CLX	clorox	Consumer Staples
110	CME Group Inc.	$CME	cme group	Financials
111	CMS Energy	$CMS	cms energy	Utilities
112	Coach Inc.	$COH	coach inc	Consumer Discretionary
113	The Coca Cola Company	$KO	coca cola	Consumer Staples
114	Coca-Cola Enterprises	$CCE	coca cola	Consumer Staples
115	Cognizant Technology Solutions	$CTSH	cognizant	Information Technology
116	Colgate-Palmolive	$CL	colgate	Consumer Staples
117	Comcast Corp.	$CMCSA	comcast	Consumer Discretionary
118	Comerica Inc.	$CMA	comerica	Financials
119	Computer Sciences Corp.	$CSC	computer sciences corp	Information Technology
120	ConAgra Foods Inc.	$CAG	conagra	Consumer Staples
121	ConocoPhillips	$COP	conoco	Energy
122	CONSOL Energy Inc.	$CNX	consol	Energy
123	Consolidated Edison	$ED	consolidated edison	Utilities
124	Constellation Brands	$STZ	constellation brands	Consumer Staples
125	Corning Inc.	$GLW	corning	Industrials
126	Costco Co.	$COST	costco	Consumer Staples
127	Covidien plc	$COV	covidien	Health Care
128	Crown Castle International Corp.	$CCI	cown castle	Telecommunications Services
129	CSX Corp.	$CSX	csx	Industrials
130	Cummins Inc.	$CMI	cummins	Industrials
131	CVS Caremark Corp.	$CVS	cvs caremark	Consumer Staples
132	D. R. Horton	$DHI	drhorton	Consumer Discretionary
133	Danaher Corp.	$DHR	danaher	Industrials
134	Darden Restaurants	$DRI	darden	Consumer Discretionary
135	DaVita Inc.	$DVA	davita	Health Care
136	Deere & Co.	$DE	deere	Industrials
137	Dell Inc.	$DELL	dell	Information Technology
138	Delphi Automotive	$DLPH	delphi	Consumer Discretionary
139	Denbury Resources Inc.	$DNR	denbury	Energy
140	Dentsply International	$XRAY	dentsply	Health Care
141	Devon Energy Corp.	$DVN	devon energy	Energy
142	Diamond Offshore Drilling	$DO	diamond offshore	Energy
143	DirecTV	$DTV	directv	Consumer Discretionary
144	Discover Financial Services	$DFS	discover financial	Financials
145	Discovery Communications	$DISCA	discovery communications	Consumer Discretionary
146	Dollar General	$DG	dollar general	Consumer Discretionary
147	Dollar Tree	$DLTR	dollar tree	Consumer Discretionary
148	Dominion Resources	$D	dominion resources	Utilities
149	Dover Corp.	$DOV	dover	Industrials
150	Dow Chemical	$DOW	dow chemical	Materials

ID	Compagnie	Ticker	Name	Secteur industriel
151	Dr Pepper Snapple Group	$DPS	dr pepper	Consumer Staples
152	DTE Energy Co.	$DTE	dtw energy	Utilities
153	Du Pont (E.I.)	$DD	dupont	Materials
154	Duke Energy	$DUK	duke energy	Utilities
155	Dun & Bradstreet	$DNB	dun badstreet	Industrials
156	E-Trade	$ETFC	e-trade	Financials
157	Eastman Chemical	$EMN	eastman chemical	Materials
158	Eaton Corp.	$ETN	eaton corp	Industrials
159	eBay Inc.	$EBAY	ebay	Information Technology
160	Ecolab Inc.	$ECL	ecolab inc	Materials
161	Edison Int'l	$EIX	edison intl	Utilities
162	Edwards Lifesciences	$EW	edwards lifesciences	Health Care
163	Electronic Arts	$EA	ea	Information Technology
164	EMC Corp.	$EMC	emc	Information Technology
165	Emerson Electric	$EMR	emerson electric	Industrials
166	Ensco plc	$ESV	ensco	Energy
167	Entergy Corp.	$ETR	entergy	Utilities
168	EOG Resources	$EOG	eog	Energy
169	EQT Corporation	$EQT	eqt	Utilities
170	Equifax Inc.	$EFX	equifax	Financials
171	Equity Residential	$EQR	equity residential	Financials
172	Estee Lauder Cos.	$EL	estee lauder	Consumer Staples
173	Exelon Corp.	$EXC	exelon	Utilities
174	Expedia Inc.	$EXPE	expedia	Consumer Discretionary
175	Expeditors Int'l	$EXPD	expeditors	Industrials
176	Express Scripts	$ESRX	express scripts	Health Care
177	Exxon Mobil Corp.	$XOM	exxon	Energy
178	F5 Networks	$FFIV	f5 networks	Information Technology
179	Family Dollar Stores	$FDO	family dollar	Consumer Discretionary
180	Fastenal Co	$FAST	fastenal	Industrials
181	FedEx Corporation	$FDX	fedex	Industrials
182	Fidelity National Information Services	$FIS	fidelity	Information Technology
183	Fifth Third Bancorp	$FITB	fifth third bancorp	Financials
184	First Horizon National	$FHN	first horizon national	Financials
185	First Solar Inc	$FSLR	first solar	Industrials
186	FirstEnergy Corp	$FE	firstenergy	Utilities
187	Fiserv Inc	$FISV	fiserv	Information Technology
188	FLIR Systems	$FLIR	flir	Industrials
189	Flowserve Corporation	$FLS	flowserve	Industrials
190	Fluor Corp.	$FLR	fluor	Industrials
191	FMC Corporation	$FMC	fmc	Materials
192	FMC Technologies Inc.	$FTI	fmc technologies	Energy
193	Ford Motor	$F	ford motor	Consumer Discretionary
194	Forest Laboratories	$FRX	forest laboratories	Health Care
195	Fossil, Inc.	$FOSL	fossil	Consumer Discretionary
196	Franklin Resources	$BEN	franklin ressources	Financials
197	Freeport-McMoran Cp & Gld	$FCX	freeport mcmoran	Materials
198	Frontier Communications	$FTR	frontier communications	Telecommunications Services
199	GameStop Corp.	$GME	gamestop	Consumer Discretionary
200	Gannett Co.	$GCI	gannett	Consumer Discretionary

ID	Compagnie	Ticker	Name	Secteur industriel
201	Gap (The)	$GPS	gap	Consumer Discretionary
202	Garmin Ltd.	$GRMN	garmin	Consumer Discretionary
203	General Dynamics	$GD	general dynamics	Industrials
204	General Electric	$GE	general electric	Industrials
205	General Mills	$GIS	general mills	Consumer Staples
206	General Motors	$GM	general motors	Consumer Discretionary
207	Genuine Parts	$GPC	genuine parts	Consumer Discretionary
208	Genworth Financial Inc.	$GNW	genworth financial	Financials
209	Gilead Sciences	$GILD	gilead sciences	Health Care
210	Goldman Sachs Group	$GS	goldman sachs	Financials
211	Goodyear Tire & Rubber	$GT	goodyear	Consumer Discretionary
212	Google Inc.	$GOOG	google	Information Technology
213	Grainger (W.W.) Inc.	$GWW	grainger	Industrials
214	Halliburton Co.	$HAL	halliburton	Energy
215	Harley-Davidson	$HOG	harley davidson	Consumer Discretionary
216	Harman Int'l Industries	$HAR	harman	Consumer Discretionary
217	Harris Corporation	$HRS	harris	Information Technology
218	Hartford Financial Svc.Gp.	$HIG	hartford financial	Financials
219	Hasbro Inc.	$HAS	hasbro	Consumer Discretionary
220	HCP Inc.	$HCP	hcp	Financials
221	Health Care REIT	$HCN	health care REIT	Financials
222	Helmerich & Payne	$HP	helmerich payne	Energy
223	Hess Corporation	$HES	hess	Energy
224	Hewlett-Packard	$HPQ	hewlett	Information Technology
225	Home Depot	$HD	home depot	Consumer Discretionary
226	Honeywell Int'l Inc.	$HON	honeywell	Industrials
227	Hormel Foods Corp.	$HRL	hormel foods	Consumer Staples
228	Hospira Inc.	$HSP	hospira	Health Care
229	Host Hotels & Resorts	$HST	host hotels & resorts	Financials
230	Hudson City Bancorp	$HCBK	hudson city bancorp	Financials
231	Humana Inc.	$HUM	humana	Health Care
232	Huntington Bancshares	$HBAN	huntington bancshares	Financials
233	Illinois Tool Works	$ITW	illinois tool works	Industrials
234	Ingersoll-Rand PLC	$IR	ingersol rand plc	Industrials
235	Integrys Energy Group Inc.	$TEG	integrys energy	Utilities
236	Intel Corp.	$INTC	intel	Information Technology
237	IntercontinentalExchange Inc.	$ICE	intercontinentalexchange	Financials
238	International Bus. Machines	$IBM	international bus. machines	Information Technology
239	International Flav/Frag	$IFF	international flav frag	Materials
240	International Game Technology	$IGT	international game technology	Consumer Discretionary
241	International Paper	$IP	international paper	Materials
242	Interpublic Group	$IPG	interpublic group	Consumer Discretionary
243	Intuit Inc.	$INTU	intuit	Information Technology
244	Intuitive Surgical Inc.	$ISRG	intuitive surgical	Health Care
245	Invesco Ltd.	$IVZ	invesco	Financials
246	Iron Mountain Incorporated	$IRM	iron mountain	Industrials
247	Jabil Circuit	$JBL	jabil circuit	Information Technology
248	Jacobs Engineering Group	$JEC	jacobs engineering	Industrials
249	JDS Uniphase Corp.	$JDSU	jds uniphase	Information Technology
250	Johnson & Johnson	$JNJ	johnson & johnson	Health Care

ID	Compagnie	Ticker	Name	Secteur industriel
251	Johnson Controls	$JCI	johnson controls	Consumer Discretionary
252	Joy Global Inc.	$JOY	joy global	Industrials
253	JPMorgan Chase & Co.	$JPM	jpmorgan	Financials
254	Juniper Networks	$JNPR	juniper networks	Information Technology
255	Kansas City Southern	$KSU	kansas city southern	Industrials
256	Kellogg Co.	$K	kellogg	Consumer Staples
257	KeyCorp	$KEY	keycorp	Financials
258	Kimberly-Clark	$KMB	kimberly clark	Consumer Staples
259	Kimco Realty	$KIM	kimco realty	Financials
260	Kinder Morgan	$KMI	kinder morgan	Energy
261	KLA-Tencor Corp.	$KLAC	kla tencor	Information Technology
262	Kohl's Corp.	$KSS	kohl's	Consumer Discretionary
263	Kraft Foods Group	$KRFT	kraft foods	Consumer Staples
264	Kroger Co.	$KR	kroger	Consumer Staples
265	L Brands Inc.	$LTD	L brands	Consumer Discretionary
266	L-3 Communications Holdings	$LLL	L-3 communications	Industrials
267	Laboratory Corp. of America Holding	$LH	laboratory of america	Health Care
268	Lam Research	$LRCX	lam research	Information Technology
269	Legg Mason	$LM	legg mason	Financials
270	Leggett & Platt	$LEG	legget platt	Industrials
271	Lennar Corp.	$LEN	lennar	Consumer Discretionary
272	Leucadia National Corp.	$LUK	leucadia national	Financials
273	Life Technologies	$LIFE	life technologies	Health Care
274	Lilly (Eli) & Co.	$LLY	eli lilly	Health Care
275	Lincoln National	$LNC	lincoln national	Financials
276	Linear Technology Corp.	$LLTC	linear technology	Information Technology
277	Lockheed Martin Corp.	$LMT	lockheed	Industrials
278	Loews Corp.	$L	loews	Financials
279	Lorillard Inc.	$LO	lorillard	Consumer Staples
280	Lowe's Cos.	$LOW	lowe's cos	Consumer Discretionary
281	LSI Corporation	$LSI	LSI	Information Technology
282	LyondellBasell	$LYB	lyondellbasell	Materials
283	M&T Bank Corp.	$MTB	M&T bank	Financials
284	Macerich	$MAC	macerich	Financials
285	Macy's Inc.	$M	macy's	Consumer Discretionary
286	Marathon Oil Corp.	$MRO	marathon oil	Energy
287	Marathon Petroleum	$MPC	marathon petroleum	Energy
288	Marriott Int'l.	$MAR	marriott	Consumer Discretionary
289	Marsh & McLennan	$MMC	marsh mclennan	Financials
290	Masco Corp.	$MAS	masco	Industrials
291	Mastercard Inc.	$MA	mastercard	Information Technology
292	Mattel Inc.	$MAT	matter	Consumer Discretionary
293	McCormick & Co.	$MKC	mccormick	Consumer Staples
294	McDonald's Corp.	$MCD	mcdonald	Consumer Discretionary
295	McGraw-Hill	$MHFI	mcgraw hill	Financials
296	McKesson Corp.	$MCK	mckesson	Health Care
297	Mead Johnson	$MJN	mead johnson	Consumer Staples
298	MeadWestvaco Corporation	$MWV	meadwestvaco	Materials
299	Medtronic Inc.	$MDT	medtronic	Health Care
300	Merck & Co.	$MRK	merck	Health Care

ID	Compagnie	Ticker	Name	Secteur industriel
301	MetLife Inc.	$MET	metlife	Financials
302	Microchip Technology	$MCHP	microchip	Information Technology
303	Micron Technology	$MU	micron	Information Technology
304	Microsoft Corp.	$MSFT	microsoft	Information Technology
305	Molex Inc.	$MOLX	molex	Information Technology
306	Molson Coors Brewing Company	$TAP	molson coors	Consumer Staples
307	Mondelez International	$MDLZ	mondelez	Consumer Staples
308	Monsanto Co.	$MON	monsanto	Materials
309	Monster Beverage	$MNST	monster beverage	Consumer Staples
310	Moody's Corp	$MCO	moody	Financials
311	Morgan Stanley	$MS	morgan stanley	Financials
312	The Mosaic Company	$MOS	mosaic company	Materials
313	Motorola Solutions Inc.	$MSI	motorola solutions	Information Technology
314	Murphy Oil	$MUR	murphy oil	Energy
315	Mylan Inc.	$MYL	mylan	Health Care
316	Nabors Industries Ltd.	$NBR	nabord	Energy
317	NASDAQ OMX Group	$NDAQ	nasdaq omx	Financials
318	National Oilwell Varco Inc.	$NOV	national oilwell varco	Energy
319	NetApp	$NTAP	netapp	Information Technology
320	NetFlix Inc.	$NFLX	netflix	Information Technology
321	Newell Rubbermaid Co.	$NWL	newell rubbermaid	Consumer Discretionary
322	Newfield Exploration Co	$NFX	newfield exploration	Energy
323	Newmont Mining Corp. (Hldg. Co.)	$NEM	newmont mining	Materials
324	News Corporation	$NWSA	news corporation	Consumer Discretionary
325	NextEra Energy Resources	$NEE	nextera energy resources	Utilities
326	NIKE Inc.	$NKE	nike	Consumer Discretionary
327	NiSource Inc.	$NI	nisource	Utilities
328	Noble Corp	$NE	noble corp	Energy
329	Noble Energy Inc	$NBL	noble energy	Energy
330	Nordstrom	$JWN	nordstrom	Consumer Discretionary
331	Norfolk Southern Corp.	$NSC	norfolk southern	Industrials
332	Northern Trust Corp.	$NTRS	northern trust	Financials
333	Northrop Grumman Corp.	$NOC	northrop grumman	Industrials
334	Northeast Utilities	$NU	northeast utilities	Utilities
335	NRG Energy	$NRG	nrg energy	Utilities
336	Nucor Corp.	$NUE	nucor	Materials
337	Nvidia Corporation	$NVDA	nvidia	Information Technology
338	NYSE Euronext	$NYX	nyse euronext	Financials
339	O'Reilly Automotive	$ORLY	o'reilly automotive	Consumer Discretionary
340	Occidental Petroleum	$OXY	occidental petroleum	Energy
341	Omnicom Group	$OMC	omnicom	Consumer Discretionary
342	ONEOK	$OKE	oneok	Utilities
343	Oracle Corp.	$ORCL	oracle	Information Technology
344	Owens-Illinois Inc	$OI	owens illinois	Materials
345	PACCAR Inc.	$PCAR	paccar	Industrials
346	Pall Corp.	$PLL	pall	Industrials
347	Parker-Hannifin	$PH	parker hannifin	Industrials
348	Patterson Companies	$PDCO	patterson	Health Care
349	Paychex Inc.	$PAYX	paychex	Information Technology
350	Peabody Energy	$BTU	peabody energy	Energy

ID	Compagnie	Ticker	Name	Secteur industriel
351	Penney (J.C.)	$JCP	penney	Consumer Discretionary
352	Pentair Ltd.	$PNR	pentair	Industrials
353	People's United Bank	$PBCT	people united bank	Financials
354	Pepco Holdings Inc.	$POM	pepco holdings	Utilities
355	PepsiCo Inc.	$PEP	pepsico	Consumer Staples
356	PerkinElmer	$PKI	perkinelmer	Health Care
357	Perrigo	$PRGO	perrogo	Health Care
358	PetSmart, Inc.	$PETM	petsmart	Consumer Discretionary
359	Pfizer Inc.	$PFE	pfizer	Health Care
360	PG&E Corp.	$PCG	pg&e	Utilities
361	Philip Morris International	$PM	philip morris	Consumer Staples
362	Phillips 66	$PSX	philips 66	Energy
363	Pinnacle West Capital	$PNW	pinnacle west capital	Utilities
364	Pioneer Natural Resources	$PXD	pioneet natural ressources	Energy
365	Pitney-Bowes	$PBI	pitney bowes	Industrials
366	Plum Creek Timber Co.	$PCL	plum creek timber	Financials
367	PNC Financial Services	$PNC	pnc financial	Financials
368	Polo Ralph Lauren Corp.	$RL	polo ralph lauren	Consumer Discretionary
369	PPG Industries	$PPG	ppg industries	Materials
370	PPL Corp.	$PPL	ppl	Utilities
371	Praxair Inc.	$PX	praxair	Materials
372	Precision Castparts	$PCP	precision castparts	Industrials
373	Priceline.com Inc	$PCLN	priceline	Consumer Discretionary
374	Principal Financial Group	$PFG	principal financial group	Financials
375	Procter & Gamble	$PG	procter	Consumer Staples
376	Progressive Corp.	$PGR	progressive	Financials
377	Prologis	$PLD	prologis	Financials
378	Prudential Financial	$PRU	prudential financial	Financials
379	Public Serv. Enterprise Inc.	$PEG	public service enterprise	Utilities
380	Public Storage	$PSA	public storage	Financials
381	Pulte Homes Inc.	$PHM	pulte homes	Consumer Discretionary
382	PVH Corp.	$PVH	pvh	Consumer Discretionary
383	QEP Resources	$QEP	qep resources	Utilities
384	Quanta Services Inc.	$PWR	quanta services	Industrials
385	QUALCOMM Inc.	$QCOM	qualcomm	Information Technology
386	Quest Diagnostics	$DGX	quest diagnostics	Health Care
387	Range Resources Corp.	$RRC	range resources	Energy
388	Raytheon Co.	$RTN	raytheon	Industrials
389	Red Hat Inc.	$RHT	red hat	Information Technology
390	Regeneron	$REGN	regeneron	Health Care
391	Regions Financial Corp.	$RF	regions financial	Financials
392	Republic Services Inc	$RSG	republic services	Industrials
393	Reynolds American Inc.	$RAI	reynolds american	Consumer Staples
394	Robert Half International	$RHI	robert half international	Industrials
395	Rockwell Automation Inc.	$ROK	rockwell automation	Industrials
396	Rockwell Collins	$COL	rockwell collins	Industrials
397	Roper Industries	$ROP	roper industries	Industrials
398	Ross Stores	$ROST	ross stores	Consumer Discretionary
399	Rowan Cos.	$RDC	rowan cos	Energy
400	Ryder System	$R	ryder system	Industrials

ID	Compagnie	Ticker	Name	Secteur industriel
401	Safeway Inc.	$SWY	safeway	Consumer Staples
402	SAIC	$SAI	saic	Industrials
403	Salesforce.com	$CRM	salesforce	Information Technology
404	SanDisk Corporation	$SNDK	sandisk	Information Technology
405	SCANA Corp	$SCG	scana	Utilities
406	Schlumberger Ltd.	$SLB	schlumberger	Energy
407	Scripps Networks Interactive Inc.	$SNI	scripps networks	Consumer Discretionary
408	Seagate Technology	$STX	seagate technology	Information Technology
409	Sealed Air Corp.(New)	$SEE	sealed air	Materials
410	Sempra Energy	$SRE	sempra energy	Utilities
411	Sherwin-Williams	$SHW	sherwin williams	Consumer Discretionary
412	Sigma-Aldrich	$SIAL	sigma aldrich	Materials
413	Simon Property Group Inc	$SPG	simon property	Financials
414	SLM Corporation	$SLM	slm corporation	Financials
415	Smucker (J.M.)	$SJM	smucker	Consumer Staples
416	Snap-On Inc.	$SNA	snap on	Consumer Discretionary
417	Southern Co.	$SO	southern	Utilities
418	Southwest Airlines	$LUV	southwest airlines	Industrials
419	Southwestern Energy	$SWN	southwestern energy	Energy
420	Spectra Energy Corp.	$SE	spectra energy	Energy
421	Sprint Nextel Corp.	$S	sprint nextel	Telecommunications Services
422	St Jude Medical	$STJ	jude medical	Health Care
423	Stanley Black & Decker	$SWK	black & decker	Consumer Discretionary
424	Staples Inc.	$SPLS	staples	Consumer Discretionary
425	Starbucks Corp.	$SBUX	starbucks	Consumer Discretionary
426	Starwood Hotels & Resorts	$HOT	starwood hotels	Consumer Discretionary
427	State Street Corp.	$STT	state street corp	Financials
428	Stericycle Inc	$SRCL	stericycle	Industrials
429	Stryker Corp.	$SYK	stryker	Health Care
430	SunTrust Banks	$STI	suntrust banks	Financials
431	Symantec Corp.	$SYMC	symantec	Information Technology
432	Sysco Corp.	$SYY	sysco	Consumer Staples
433	T. Rowe Price Group	$TROW	rowe price	Financials
434	Target Corp.	$TGT	target	Consumer Discretionary
435	TE Connectivity Ltd.	$TEL	te conectivity	Information Technology
436	TECO Energy	$TE	teco energy	Utilities
437	Tenet Healthcare Corp.	$THC	tenet heathcare	Health Care
438	Teradata Corp.	$TDC	teradara	Information Technology
439	Teradyne Inc.	$TER	teradyne	Information Technology
440	Tesoro Petroleum Co.	$TSO	tesoro petroleum	Energy
441	Texas Instruments	$TXN	texas instruments	Information Technology
442	Textron Inc.	$TXT	textron	Industrials
443	The Hershey Company	$HSY	hershey	Consumer Staples
444	The Travelers Companies Inc.	$TRV	travelers companies	Financials
445	Thermo Fisher Scientific	$TMO	thermo fisher scientific	Health Care
446	Tiffany & Co.	$TIF	tiffany	Consumer Discretionary
447	Time Warner Inc.	$TWX	time warner	Consumer Discretionary
448	Time Warner Cable Inc.	$TWC	time warner cable	Consumer Discretionary
449	TJX Companies Inc.	$TJX	tjx	Consumer Discretionary
450	Torchmark Corp.	$TMK	torchmark	Financials

ID	Compagnie	Ticker	Name	Secteur industriel
451	Total System Services	$TSS	total system services	Information Technology
452	TripAdvisor	$TRIP	tripadvisor	Consumer Discretionary
453	Tyson Foods	$TSN	tyson foods	Consumer Staples
454	Tyco International	$TYC	tyco international	Industrials
455	U.S. Bancorp	$USB	us bancorp	Financials
456	Union Pacific	$UNP	union pacific	Industrials
457	United Health Group Inc.	$UNH	united health	Health Care
458	United Parcel Service	$UPS	united parcel service	Industrials
459	United States Steel Corp.	$X	united states steel	Materials
460	United Technologies	$UTX	united technologies	Industrials
461	Unum Group	$UNM	unum	Financials
462	Urban Outfitters	$URBN	urban outfitters	Consumer Discretionary
463	V.F. Corp.	$VFC	vf corp	Consumer Discretionary
464	Valero Energy	$VLO	valero energy	Energy
465	Varian Medical Systems	$VAR	varian medical systems	Health Care
466	Ventas Inc	$VTR	ventas	Financials
467	Verisign Inc.	$VRSN	verisign	Information Technology
468	Verizon Communications	$VZ	verizon communications	Telecommunications Services
469	Viacom Inc.	$VIAB	viacom	Consumer Discretionary
470	Visa Inc.	$V	visa	Information Technology
471	Vornado Realty Trust	$VNO	vornado realty trust	Financials
472	Vulcan Materials	$VMC	vulcan materials	Materials
473	Wal-Mart Stores	$WMT	walmart	Consumer Staples
474	Walgreen Co.	$WAG	walgreen	Consumer Staples
475	The Walt Disney Company	$DIS	walt disney	Consumer Discretionary
476	Washington Post Co B	$WPO	washington post	Consumer Discretionary
477	Waste Management Inc.	$WM	waste management	Industrials
478	Waters Corporation	$WAT	waters corporation	Health Care
479	WellPoint Inc.	$WLP	wellpoint	Health Care
480	Wells Fargo	$WFC	wells fargo	Financials
481	Western Digital	$WDC	western digital	Information Technology
482	Western Union Co	$WU	western union	Information Technology
483	Weyerhaeuser Corp.	$WY	weyerhaeuser	Financials
484	Whirlpool Corp.	$WHR	whirlpool	Consumer Discretionary
485	Whole Foods Market	$WFM	whole foods	Consumer Staples
486	Williams Cos.	$WMB	williams cos	Energy
487	Windstream Communications	$WIN	windstream	Telecommunications Services
488	Wisconsin Energy Corporation	$WEC	wisconsin energy	Utilities
489	WPX Energy, Inc.	$WPX	wpx	Energy
490	Wyndham Worldwide	$WYN	wyndham	Consumer Discretionary
491	Wynn Resorts Ltd	$WYNN	wynn	Consumer Discretionary
492	Xcel Energy Inc	$XEL	xcel	Utilities
493	Xerox Corp.	$XRX	xerox	Information Technology
494	Xilinx Inc	$XLNX	xilinx	Information Technology
495	XL Capital	$XL	xl capital	Financials
496	Xylem Inc.	$XYL	xylem	Industrials
497	Yahoo Inc.	$YHOO	yahoo	Information Technology
498	Yum! Brands Inc	$YUM	yum!	Consumer Discretionary
499	Zimmer Holdings	$ZMH	zimmer	Health Care
500	Zions Bancorp	$ZION	zions bancorp	Financials

ANNEXE B : RÉSULTATS DES MODÉLISATIONS POUR LES RENDEMENTS JOURNALIERS

Tableau B.1: résultats des modèles MCO pour les valeurs absolues de rendements journaliers.

*** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1.

Variable dépendante	**Valeurs absolues de rendements journaliers**											
Variable dépendante (t-1)	0,2377	***	0,2316	***	0,2383	***	0,2321	***	0,2119	***	0,2106	***
IndiceTicker	0,0006	***			0,0006	***			0,0009	***		
IndiceName			0,00001				0,00001				0,0001	
Jours de la semaine												
Lundi					0,0008	***	0,0006	**				
Mardi (omis)												
Mercredi					0,0004		0,0003					
Jeudi					0,0004		0,0005	*				
Vendredi					-0,0003		-0,0005					
Secteurs industriels												
Santé									-0,0023	***	-0,0026	***
Énergie									-0,0008	*	-0,0007	
Consommation de base									-0,0019	***	-0,0018	***
Vente au détail									0,0005	**	0,0006	**
Industries									0,0021	***	0,0018	***
Télécommunications									-0,0042	***	-0,0036	***
Services publics									-0,0052	***	-0,0038	***
Matériaux									0,003	***	0,0031	***
Finance									-0,0003		0,0001	
Technologies de l'information (omis)												
Constante	0,0068	***	0,0082	***	0,0065	***	0,008	***	0,0091	***	0,0082	***
Nombre d'observations	16462		16210		16462		16210		16462		16210	
R^2	0,0594		0,0539		0,0605		0,0549		0,0787		0,0702	

Tableau B.2 : résultats des modèles MCO avec décalage temporel pour les valeurs absolues de rendements journaliers. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Valeurs absolues de rendements journaliers**					
Variable dépendante (t-1)	0,2377 ***	0,2316 ***	0,2333 ***	0,2298 ***	0,2207 ***	0,2251 ***
IndiceTicker	0,0006 ***					
IndiceName		0,00001				
Délai d'un jour						
IndiceTicker			-0,0005 ***			
IndiceName				-0,0002 ***		
Délai de deux jours						
IndiceTicker					-0,0006 ***	
IndiceName						-0,0003 ***
Délai de trois jours						
IndiceTicker						
IndiceName						
Délai de quatre jours						
IndiceTicker						
IndiceName						
Constante	0,0068 ***	0,0082 ***	0,0092 ***	0,0087 ***	0,0095 ***	0,009 ***
Nombre d'observations	16462	16210	16462	16210	16461	16209
R^2	0,0594	0,0539	0,0557	0,0541	0,0536	0,0528

Variable dépendante (t-1)	0,2253 ***	0,1503 ***	0,222 ***	0,2247 ***		
IndiceTicker						
IndiceName						
Délai d'un jour						
IndiceTicker						
IndiceName						
Délai de deux jours						
IndiceTicker						
IndiceName						
Délai de trois jours						
IndiceTicker	-0,0006 ***					
IndiceName		-0,0003 ***				
Délai de quatre jours						
IndiceTicker			-0,0006 ***			
IndiceName				-0,0003 ***		
Constante	0,009 ***	-4,2369 ***	0,0095 ***	0,0088 ***		
Nombre d'observations	16208	16803	16459	16207		
R^2	0,0524	0,0519	0,0545	0,0522		

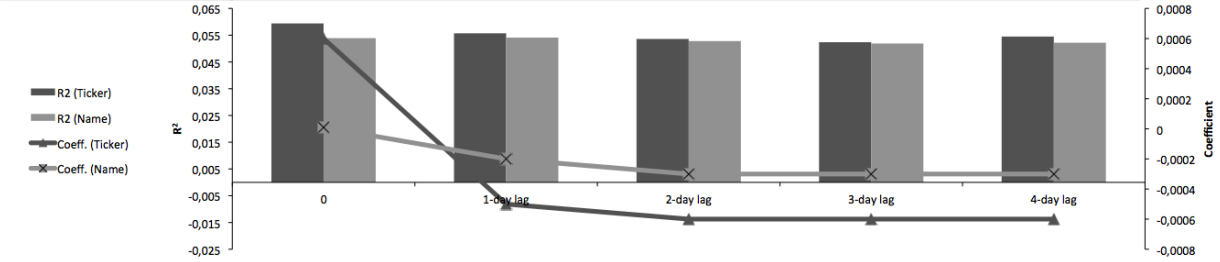

Tableau B.3 : résultats des modèles MCO pour des valeurs positives de rendements journaliers.
*** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Valeurs positives de rendements journaliers**											
Variable dépendante (t-1)	0,197	***	0,1952	***	0,1972	***	0,1986	***	0,1758	***	0,1862	***
IndiceTicker	0,0014	***			0,0014	***			0,0017	***		
IndiceName			0,0005	**			0,0005	**			0,0006	***
Jours de la semaine												
Lundi					0,0004		0,0001					
Mardi (omis)												
Mercredi					0,0005		0,0004					
Jeudi					0,0002		0,0002					
Vendredi					0,0005		0,0002					
Secteurs industriels												
Santé									-0,0021	**	-0,0022	**
Énergie									-0,0009		-0,0005	
Consommation de base									-0,003	***	-0,0023	**
Vente au détail									0,0005		0,0008	
Industries									0,0026	**	0,0025	**
Télécommunications									-0,0048	***	-0,0038	**
Services publics									-0,0064	***	-0,0034	*
Matériaux									0,0021	*	0,0036	***
Finance									-0,0001		0,0013	
Technologies de l'information (omis)												
Constante	0,0056	***	0,0075	***	0,0052	***	0,0073	***	0,0056	***	0,0073	***
Nombre d'observations	4239		4151		4239		4151		4239		4151	
R^2	0,0487		0,0484		0,0488		0,0487		0,0663		0,0626	

Tableau B.4 : résultats des modèles probit avec variables de contrôle pour les rendements journaliers. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements journaliers**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,0497 **		-0,0732 ***	-0,1627 **	-0,0085					
IndiceTicker	0,0341 ***		0,0555 ***	0,1254 ***	0,1566 *					
IndiceName										
Constante	-0,0182		-0,9516 ***	-2,7771 ***	-3,5179 ***					
Nombre d'observations	16636		16636	16636	16636					
R^2	0,0007		0,0019	0,0136	0,0158					

Variable dépendante	**Rendements journaliers**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,0448 **		-0,0739 ***	-0,1618 **	-0,006					
IndiceTicker	0,0372 ***		0,0571 ***	0,1252 ***	0,1568 *					
IndiceName										
Jours de la semaine										
Lundi	-0,1007 ***		-0,0239	-0,034	-0,1656					
Mardi (omis)										
Mercredi	-0,135 ***		-0,0262	0,017	-0,082					
Jeudi	-0,147 ***		-0,116 ***	0,0702	-0,2154					
Vendredi	0,043		-0,0027	0,0428	-0,0756					
Constante	0,0399		-0,9213 ***	-2,78 ***	-3,4216 ***					
Nombre d'observations	16636		16636	16636	16636					
R^2	0,0034		0,0028	0,0144	0,0197					

Variable dépendante	**Rendements journaliers**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,0546 ***	-0,0037	-0,0678 ***	-0,1613 **	-0,0092			-0,0634 ***		
IndiceTicker	0,0367 ***	-0,0259 **	0,0742 ***	0,1723 ***	0,2142 **					
IndiceName								-0,0212 **		
Secteurs industriels										
Santé	0,1054 ***	0,2692 ***	-0,2122 ***	*(omis)*	*(omis)*			-0,2205 ***		
Énergie	0,1314 **	0,2761 ***	-0,1792 ***	-0,0212	*(omis)*			-0,1851 ***		
Consommation de base	0,1141 ***	0,2889 ***	-0,2404 ***	-0,0401	0,188			-0,2528 ***		
Vente au détail	0,0187	-0,0232	0,0487	0,1369	-0,4807			0,0362		
Industries	0,0198	-0,0155	-0,0029	0,4721 ***	0,3526			-0,0274		
Télécommunications	0,0646	0,3088 ***	-0,3176 ***	*(omis)*	*(omis)*			-0,2771 ***		
Services publics	-0,0331	0,351 ***	-0,5726 ***	*(omis)*	*(omis)*			-0,5113 ***		
Matériaux	-0,1102 ***	-0,2365 ***	0,0874 *	0,2705 **	-0,1786			0,0702		
Finance	0,0316	0,0011	0,0578	-0,4145 **	*(omis)*			0,0685 *		
Technologies de l'information (omis)										
Constante	-0,0467	-0,4835 ***	-0,9621 ***	-2,9274 ***	-3,5251 ***			-0,7484 ***		
Nombre d'observations	16636	16636	16636	14295	11715			17147		
R^2	0,0022	0,0108	0,0108	0,0452	0,0701			0,0087		

Tableau B.5 : résultats des modèles probit avec décalage temporel pour les rendements journaliers. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements journaliers (avec décalage temporel)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)			-0,0793 ***					-0,0695 ***		-0,0136
IndiceTicker (t-1)			-0,0331 ***							
IndiceName (t-1)								-0,0166 *		0,1351 *
Constante			-0,7491 ***					-0,7939 ***		-3,5101 ***
Nombre d'observations			16636					17147		17147
R^2			0,0013					0,0008		0,0152

Variable dépendante	**Rendements journaliers (avec décalage temporel)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)			-0,0753 ***					-0,0685 ***		-0,0081
IndiceTicker (t-1)			-0,0276 ***							
IndiceName (t-1)								-0,0167 *		0,137 *
Jours de la semaine										
Lundi			-0,0404					-0,0418		-0,2014
Mardi (omis)										
Mercredi			-0,0213					-0,0254		-0,1212
Jeudi			-0,0759 **					-0,0747 **		-0,2232
Vendredi			0,0257					0,0079		-0,1171
Constante			-0,7476 ***					-0,7677 ***		-3,3954 ***
Nombre d'observations			16462					17147		17147
R^2			0,0018					0,0012		0,0197

Variable dépendante	**Rendements journaliers (avec décalage temporel)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)			-0,0737 ***					-0,0635 ***		-0,0149
IndiceTicker (t-1)			-0,0217 *							
IndiceName (t-1)								-0,0242 **		0,16 *
Secteurs industriels										
Santé			-0,2242 ***					-0,2163 ***		*(omis)*
Énergie			-0,1997 ***					-0,1802 ***		*(omis)*
Consommation de base			-0,2421 ***					-0,2559 ***		0,3499
Vente au détail			0,0387					0,0346		-0,3741
Industries			-0,0178					-0,0232		0,417 *
Télécommunications			-0,2619 ***					-0,2786 ***		*(omis)*
Services publics			-0,4752 ***					-0,5301 ***		*(omis)*
Matériaux			0,0893 *					0,0622		0,1434
Finance			0,0822 **					0,0657 *		*(omis)*
Technologies de l'information (omis)										
Constante			-0,7444 ***					-0,7392 ***		-3,546 ***
Nombre d'observations			16636					17147		12129
R^2			0,0092					0,0088		0,0586

Tableau B.6 : résultats des modèles probit avec interaction de variables pour les rendements journaliers. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements journaliers**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)			-0,0774 ***		0,0269				-0,1727 **	-0,0064
IndiceTicker (t-1)			-0,0423		-0,5438 **					
IndiceName (t-1)									-0,0134	-0,0286
Jours de la semaine										
Lundi			0,0126		-0,6774				-0,0047	-0,7860
Mardi (omis)										
Mercredi			-0,1911 **		-1,1810 *				-0,2014	-1,3619
Jeudi			-0,0341		-0,9878				0,0755	-0,2264
Vendredi			0,0080		-0,9324				0,3978	-1,8413
Lundi x *Tweet (t-1)*			-0,0269		0,2990				0,0027	0,2174
Mardi x *Tweet (t-1) (omis)*										
Mercredi x *Tweet (t-1)*			0,0740 *		0,6150 *				0,0842	0,4135
Jeudi x *Tweet (t-1)*			-0,0136		0,4781				-0,0118	0,0048
Vendredi x *Tweet (t-1)*			0,0082		0,5083				-0,1737 *	0,5509 *
Constante			-0,7096 ***		-2,1307 ***				-2,4678 ***	-2,9667 ***
Nombre d'observations			16636		16636				17147	17147
R^2			0,0023		0,0384				0,0129	0,0466

Variable dépendante	**Rendements journaliers**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,0682 ***	-0,0144	-0,0724 ***	-0,1614 **		-0,0525 ***	-0,0052	-0,0614 ***	-0,1651 **	-0,0149
IndiceTicker (t-1)	0,0085	0,0562 **	-0,0372	-0,2448 ***						
IndiceName (t-1)						0,0166	0,0065	-0,0002	0,3972 ***	0,4342 *
Secteurs industriels										
Santé	0,2316 ***	0,7066 ***	-0,6285 ***	0,1626		0,1985 *	0,5267 ***	-0,5732 ***	0,8976	*(omis)*
Énergie	0,0879	0,0527	0,1289	0,4238		0,2830	1,1203 ***	-1,2853 ***	-0,0250	*(omis)*
Consommation de base	-0,0263	-0,0630	0,1477	0,0204		0,1757 *	0,1051	0,0678	1,6778 ***	1,5217
Vente au détail	0,0021	0,0317	0,0092	-0,5445 **		0,0548	-0,1091	0,1312	1,7535 ***	1,2927
Industries	0,1496	0,3289 ***	-0,2760 **	0,6007 *		0,1654	-0,0549	0,1825	2,0307 ***	1,4247
Télécommunications	0,3416	0,7082 ***	-0,4468	*(omis)*		0,1747	0,2501	-0,1067	*(omis)*	*(omis)*
Services publics	0,2037	0,7476 **	-0,9395	*(omis)*		0,1707	0,4625 ***	-0,5168 ***	*(omis)*	*(omis)*
Matériaux	0,0975	0,0398	0,1733	-0,9393 **		0,0722	-0,0293	0,0990	1,5903 ***	0,1661
Finance	0,2199 **	0,1847 *	0,1389	-1,3064 *		0,1928 **	-0,0087	0,2422 **	0,6327	*(omis)*
Technologies de l'information (omis)										
Santé x *Tweet (t-1)*	-0,0613	-0,2118 ***	0,1959 ***	-1,2779		-0,0257	-0,0978 **	0,1508 ***	-0,4646	*(omis)*
Énergie x *Tweet (t-1)*	0,0202	0,1033	-0,1504 *	-0,2299		-0,0514	-0,3243 ***	0,4138 ***	0,0665	*(omis)*
Consommation de base x *Tweet (t-1)*	0,0695 *	0,1585 ***	-0,1838 ***	-0,0236		-0,0119	0,0922 **	-0,1451 ***	-0,6100 ***	-0,3557
Vente au détail x *Tweet (t-1)*	0,0060	-0,0229	0,0138	0,3378 ***		-0,0085	0,0394	-0,0338	-0,5392 ***	-0,5682 *
Industries x *Tweet (t-1)*	-0,0608	-0,1580 ***	0,1257 **	-0,0961		-0,0535	0,0236	-0,0788	-0,5096 ***	-0,2914
Télécommunications x *Tweet (t-1)*	-0,0923	-0,1597 **	0,0697	*(omis)*		-0,0249	0,0211	-0,0609	*(omis)*	*(omis)*
Services publics x *Tweet (t-1)*	-0,0586	-0,1427	0,1436	*(omis)*		-0,0874	-0,0970	0,0287	*(omis)*	*(omis)*
Matériaux x *Tweet (t-1)*	-0,0897 **	-0,1230 ***	-0,0372	0,5533 ***		-0,0940 **	-0,1349 **	0,0022	-0,3365 **	0,1103
Finance x *Tweet (t-1)*	-0,0719 *	-0,0826 *	-0,0202	0,4301 *		-0,0560 *	0,0093	-0,0736 *	-0,2485	*(omis)*
Technologies de l'information (omis)										
Constante	0,0257	-0,6555 ***	-0,7119 ***	-2,0705 ***		-0,0358	-0,5697 ***	-0,8128 ***	-3,8661 ***	-4,5074 ***
Nombre d'observations	16636	16636	16636	15700		17147	17147	17147	16196	12129
R^2	0,0029	0,0152	0,0128	0,0719		0,0022	0,0135	0,0124	0,0617	0,0817

ANNEXE C : RÉSULTATS DES MODÉLISATIONS POUR LES RENDEMENTS NOCTURNES

Tableau C.1 : résultats des modèles MCO pour les valeurs absolues de rendements nocturnes. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Valeurs absolues de rendements nocturnes**											
Variable dépendante (t-1)	0,048	***	0,0505	***	0,0485	***	0,051	***	0,0379	***	0,0423	***
IndiceTicker	0,0013	***			0,0013	***			0,0015	***		
IndiceName			0,0005	***			0,0005	***			0,0006	***
Jours de la semaine												
Lundi					0,0003		0,0003					
Mardi (omis)												
Mercredi					0,001	***	0,001	***				
Jeudi					0,0006	*	0,0006	*				
Vendredi					0,0018	***	0,0017	***				
Secteurs industriels												
Santé									-0,0013	***	-0,0012	***
Énergie									-0,001	*	-0,0005	
Consommation de base									-0,002	***	-0,0012	***
Vente au détail									0,0003		0,0007	**
Industries									0,0011	***	0,0012	***
Télécommunications									-0,0037	***	-0,0026	***
Services publics									-0,0056	***	-0,0026	***
Matériaux									0,0021	***	0,0034	***
Finance									0,00001	***	0,0013	***
Technologies de l'information (omis)												
Constante	0,0033	***	0,0048	***	0,0025	***	0,004	***	0,0031	***	0,0042	***
Nombre d'observations	16321		16075		16321		16075		16321		16075	
R^2	0,0133		0,0047		0,0156		0,0066		0,025		0,0133	

Tableau C.2 : résultats des modèles MCO avec décalage temporel pour les valeurs absolues de rendements nocturnes. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Valeurs absolues de rendements nocturnes**																					
Variable dépendante (t-1)	0,048	***	0,0505	***	0,0476	***	0,0514	***	0,0565	***	0,0535	***	0,049	***	0,049	***	0,0517	***	0,0504	***
IndiceTicker	0,0013	***																		
IndiceName			0,0005	***																
Délai d'un jour																				
IndiceTicker					0,0004	***														
IndiceName							0,00004													
Délai de deux jours																				
IndiceTicker									-0,0004	***										
IndiceName											-0,0002	**								
Délai de trois jours																				
IndiceTicker													-0,0004	***						
IndiceName															-0,0002	*				
Délai de quatre jours																				
IndiceTicker																	-0,0005	***		
IndiceName																			-0,0001	*
Constante	0,0033	***	0,0048	***	0,0051	***	0,0057	***	0,0067	***	-4,8403	***	0,0066	***	0,0062	***	0,0067	***	0,0061	***
Nombre d'observations	16321		16075		16321		16076		16321		16076		16320		16075		16319		16076	
R^2	0,0133		0,0047		0,0036		0,0028		0,0041		0,0031		0,0036		0,0026		0,0044		0,0028	

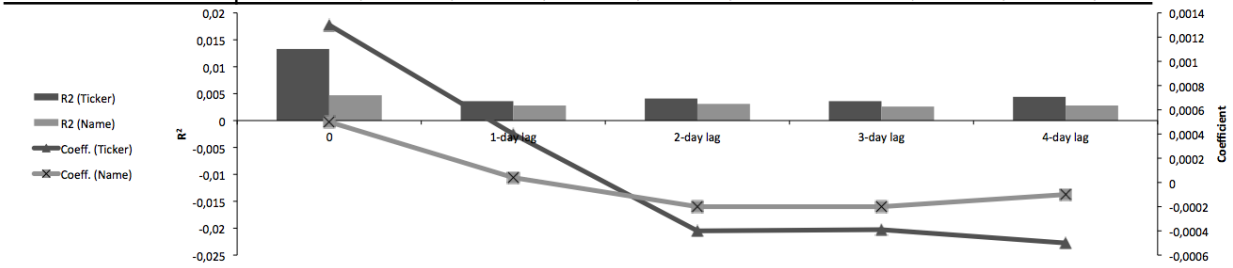

Tableau C.3 : résultats des modèles MCO pour des valeurs positives de rendements nocturnes.

*** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Valeurs positives de rendements nocturnes**											
Variable dépendante (t-1)	0,0758	***	0,0814	***	0,0745	***	0,0812	***	0,0577	***	0,0667	***
IndiceTicker	0,0015	***			0,0016	***			0,0018	***		
IndiceName			0,0004	***			0,0004	***			0,0005	***
Jours de la semaine												
Lundi					-0,0001		-0,0002					
Mardi (omis)												
Mercredi					-0,0003		-0,0002					
Jeudi					0,0008		0,0008					
Vendredi					0,0006		0,0003					
Secteurs industriels												
Santé									-0,0012	*	-0,0012	*
Énergie									-0,007		-0,0002	
Consommation de base									-0,0026	***	-0,0021	***
Vente au détail									0,0006		0,0007	
Industries									0,0018	**	0,0016	**
Télécommunications									-0,0043	***	-0,0028	**
Services publics									-0,0058	***	-0,0026	**
Matériaux									0,002	***	0,0031	***
Finance									-0,0005		0,0008	
Technologies de l'information (omis)												
Constante	0,0024	***	0,0047	***	0,0021	***	0,0045	***	0,0022	***	0,0044	***
Nombre d'observations	4105		4095		4205		4095		4205		4095	
R^2	0,0253		0,0076		0,0269		0,0089		0,048		0,0219	

Tableau C.4 : résultats des modèles probit avec variables de contrôle pour les rendements nocturnes. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements nocturnes**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,14 ***		-0,1229 ***	-0,1269	0,2358				-0,0858	0,2482
IndiceTicker	0,0379 ***		0,1112 ***	0,5184 ***	0,5189 ***					
IndiceName									0,1876 ***	0,2434 **
Constante	0,048 *		-1,6309 ***	-4,2235 ***	-4,8619 ***				-3,2893 ***	-4,077 ***
Nombre d'observations	16495		16495	16495	16495				17006	17006
R^2	0,0028		0,0077	0,1379	0,1428				0,0282	0,051

Variable dépendante	**Rendements nocturnes**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,1325 ***	-0,0866 ***	-0,1275 ***	-0,14	0,2771				-0,096	0,2846
IndiceTicker	0,0336 ***	-0,0203 **	0,114 ***	0,523 ***	0,5326 ***					
IndiceName									0,1926 ***	0,2443 **
Jours de la semaine										
Lundi	-0,4244 ***	-0,4119 ***	-0,044	0,0172	-0,3203				0,0356	-0,2615
Mardi (omis)										
Mercredi	-0,2125 ***	-0,2259 ***	0,0453	0,0593	-0,3841				0,0868	-0,3599
Jeudi	-0,0939 ***	-0,1416 ***	0,1445 ***	0,0128	0,1266				0,039	0,1089
Vendredi	-0,1508 ***	-0,221 ***	0,1787 ***	0,3376 **	-0,1264				0,3108 **	-0,1144
Constante	0,2262 ***	0,1485 ***	-1,7089 ***	-4,3389 ***	-4,8534 ***				-3,4142 ***	-4,0223 ***
Nombre d'observations	16495	16495	16495	16495	16495				17006	17006
R^2	0,0114	0,0088	0,0115	0,1493	0,1625				0,0383	0,0688

Variable dépendante	**Rendements nocturnes**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,144 ***	-0,0968 ***	-0,1407 ***	-0,1524	0,2334			-0,1338 ***	-0,0968	0,251
IndiceTicker	0,035 ***	-0,026 **	0,1317 ***	0,5975 ***	0,6165 ***					
IndiceName								0,0304 **	0,1837 ***	0,2329 **
Secteurs industriels										
Santé	-0,0753 *	-0,0096	-0,2479 ***	-0,4609 *				-0,2389 ***	-0,1789	*(omis)*
Énergie	0,0158	0,0457	-0,0932	0,0641	0,2998			-0,071	0,0457	0,4119
Consommation de base	-0,0808 *	0,0433	-0,4874 ***	-0,5335 *	0,0682			-0,4399 ***	-0,2935	0,3498
Vente au détail	-0,0017	-0,0021	-0,0197	0,1086	0,2716			-0,0127	0,1507	0,4044 *
Industries	0,0357	-0,0883 **	0,304 ***	-0,3857	0,1702			0,2808 ***	-0,3146	0,2669
Télécommunications	-0,0501	0,1079 *	-0,5377 ***	*(omis)*				-0,4513 ***	*(omis)*	*(omis)*
Services publics	-0,0469	0,1774 ***	-0,8676 ***	*(omis)*				-0,6467 ***	*(omis)*	*(omis)*
Matériaux	-0,0044	-0,2022 ***	0,4198 ***	-0,3163				0,4857 ***	0,1955	*(omis)*
Finance	0,1048 ***	0,0256	0,2103 ***	-0,439 **				0,2887 ***	-0,0663	*(omis)*
Technologies de l'information (omis)										
Constante	0,0573 *	-0,0242	-1,687 ***	-4,296 ***	-5,1311 ***			-1,4963 ***	-3,2748 ***	-4,2319 ***
Nombre d'observations	16495	16495	16495	15567	11146			17006	16063	11578
R^2	0,0039	0,0033	0,0406	0,1779	0,196			0,033	0,0397	0,0578

Tableau C.5 : résultats des modèles probit avec décalage temporel pour les rendements nocturnes. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements nocturnes (avec décalage temporel)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,1522 ***		-0,1485 ***	-0,0694			-0,1016 ***	-0,1337 ***		
IndiceTicker (t-1)	0,0288 ***		0,0381 **	0,2265 ***						
IndiceName (t-1)							0,014 *	-0,0325 ***		
Constante	0,0495 *		-1,4624 ***	-3,3643 ***			-0,1269 ***	-1,306 ***		
Nombre d'observations	16495		16495	16495			17006	17006		
R^2	0,0029		0,0036	0,0327			0,0013	0,0032		

Variable dépendante	**Rendements nocturnes (avec décalage temporel)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,141 ***		-0,1521 ***	-0,0768				-0,1386 ***		
IndiceTicker (t-1)	0,0234 **		0,0349 **	0,2267 ***						
IndiceName (t-1)								-0,0339 ***		
Jours de la semaine										
Lundi	-0,445 ***		-0,0633	0,0299				-0,0571		
Mardi (omis)										
Mercredi	-0,2033 ***		0,0526	0,0076				0,0605		
Jeudi	-0,0846 ***		0,1063 **	-0,0081				0,1277 ***		
Vendredi	-0,173 ***		0,149 ***	0,1888				0,1633 ***		
Constante	0,2363 ***		-1,5088 ***	-3,4151 ***				-1,3658 ***		
Nombre d'observations	16495		16495	16495				17006		
R^2	0,0128		0,0065	0,0371				0,0067		

Variable dépendante	**Rendements nocturnes (avec décalage temporel)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,1559 ***		-0,1658 ***	-0,0802						
IndiceTicker (t-1)	0,026 **		0,0498 ***	0,2706 ***						
IndiceName (t-1)										
Secteurs industriels										
Santé	-0,0689 *		-0,2233 ***	-0,2051						
Énergie	0,0181		-0,0786	0,0072						
Consommation de base	-0,0859 **		-0,4948 ***	-0,4043						
Vente au détail	-0,0004		-0,0243	0,1111						
Industries	0,0416		0,2908 ***	-0,3425						
Télécommunications	-0,0444		-0,4667 ***	*(omis)*						
Services publics	-0,0418		-0,7604 ***	*(omis)*						
Matériaux	-0,0005		0,4211 ***	-0,1058						
Finance	0,1011 ***		0,2208 ***	-0,2642						
Technologies de l'information (omis)										
Constante	0,0571 *		-1,5024 ***	-3,3927 ***						
Nombre d'observations	16495		16495	15567						
R^2	0,004		0,0341	0,0548						

Tableau C.6 : résultats des modèles probit avec interaction de variables pour les rendements nocturnes. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements nocturnes**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)		-0,0865 ***	-0,1516 ***	-0,0843						
IndiceTicker (t-1)		-0,0470 *	0,1450 ***	0,1989						
IndiceName (t-1)										
Jours de la semaine										
Lundi		-0,6116 ***	0,0856	0,7678						
Mardi (omis)										
Mercredi		-0,4181 ***	0,4605 ***	-0,5124						
Jeudi		-0,2028 **	0,4435 ***	-1,3375 *						
Vendredi		-0,3585 ***	0,4546 ***	0,2197						
Lundi x *Tweet (t-1)*		0,0829 **	-0,0626	-0,3471 *						
Mardi x *Tweet (t-1) (omis)*										
Mercredi x *Tweet (t-1)*		0,0885 **	-0,1787 ***	0,1802						
Jeudi x *Tweet (t-1)*		0,0375	-0,1478 ***	0,4218 *						
Vendredi x *Tweet (t-1)*		0,0583 *	-0,1330 ***	-0,0108						
Constante		0,1898 ***	-1,7620 ***	-3,3384 ***						
Nombre d'observations		16495	16495	16495						
R^2		0,0104	0,0083	0,0665						

Variable dépendante	**Rendements nocturnes**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)			-0,1679 ***	-0,0790		-0,1551 ***	-0,1023 ***	-0,1553 ***	-0,1199	0,1038
IndiceTicker (t-1)			0,0454	0,4046 ***						
IndiceName (t-1)						0,0387 *	0,0047	0,0983 ***	0,0700	1,3224 *
Secteurs industriels										
Santé			-0,3814 ***	0,9292 *		0,0277	0,0265	-0,0852	0,1978	*(omis)*
Énergie			-0,0341	-0,5702		-0,0582	0,2209	-0,7470 **	-2,4619	2,0530
Consommation de base			0,1050	0,8117		0,0965	0,0358	0,2012	0,3331	5,6843 *
Vente au détail			-0,1837	0,4640		0,0941	-0,0221	0,3332 **	0,2961	5,2217 *
Industries			0,3560 **	-0,3830		0,1096	-0,1444	0,6571 ***	0,4181	3,9129
Télécommunications			0,1870	*(omis)*		-0,0094	0,0083	-0,1640	*(omis)*	*(omis)*
Services publics			-0,8554	*(omis)*		0,2580 **	0,3009 ***	-0,3948	*(omis)*	*(omis)*
Matériaux			0,3907 ***	-0,8569		0,0642	-0,2301 **	0,7286 ***	-0,2045	*(omis)*
Finance			0,2694 *	0,7475		0,2904 ***	-0,0055	0,7303 ***	0,5357	*(omis)*
Technologies de l'information (omis)										
Santé x *Tweet (t-1)*			0,0743	-0,4698 **		-0,0376	-0,0146	-0,0412	-0,1666 *	*(omis)*
Énergie x *Tweet (t-1)*			-0,0197	0,2064		0,0332	-0,0708	0,2587 **	0,7932	-0,3155
Consommation de base x *Tweet (t-1)*			-0,2943 ***	-0,4965		-0,0637 *	0,0050	-0,2964 ***	-0,4240	-1,6852 **
Vente au détail x *Tweet (t-1)*			0,0736	-0,1291		-0,0308	0,0119	-0,1254 ***	-0,0491	-1,3685 *
Industries x *Tweet (t-1)*			-0,0322	0,0109		-0,0202	0,0334	-0,1367 **	-0,3513	-0,8927
Télécommunications x *Tweet (t-1)*			-0,2422 *	*(omis)*		-0,0095	0,0243	-0,0952	*(omis)*	*(omis)*
Services publics x *Tweet (t-1)*			0,0296	*(omis)*		-0,1925 ***	-0,1567 ***	-0,1044	*(omis)*	*(omis)*
Matériaux x *Tweet (t-1)*			0,0135	0,1924		-0,0109	0,0209	-0,1040 *	0,1624	*(omis)*
Finance x *Tweet (t-1)*			-0,0183	-0,3632		-0,0678 **	0,0111	-0,1844 ***	-0,4058 *	*(omis)*
Technologies de l'information (omis)										
Constante			-1,4917 ***	-3,7587 ***		0,0111	-0,0996	-1,6968 ***	-2,9205 ***	-8,3673 ***
Nombre d'observations			16495	15567		17006	17006	17006	16063	11578
R^2			0,0368	0,0732		0,0044	0,0035	0,0363	0,0407	0,1032

ANNEXE D : VOLUME ET RENDEMENTS ANORMAUX : MODÈLES PROBIT

Tableau D.1 : résultats des modèles probit avec interaction de variables pour les rendements concernant les volumes d'échange. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Volume**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)		-0,0997 *								-0,4708 ***
IndiceTicker (t-1)		0,0831								
IndiceName (t-1)										-0,0084
Jours de la semaine										
Lundi		0,1338								-0,3503 ***
Mardi (omis)										
Mercredi		0,2648								0,1037
Jeudi		0,5373 **								-0,0282
Vendredi		0,3199								-0,1424 *
Lundi x *Tweet (t-1)*		-0,1579								-0,0444
Mardi x *Tweet (t-1) (omis)*										
Mercredi x *Tweet (t-1)*		-0,0686								-0,0560 **
Jeudi x *Tweet (t-1)*		-0,1957 **								0,0077
Vendredi x *Tweet (t-1)*		-0,1156								0,0146
Constante		-2,4741 ***								0,0268
Nombre d'observations		16495								17006
R^2		0,0110								0,0376

Variable dépendante	**Volume**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,4964 ***			-0,1544 ***		-0,5106 ***				
IndiceTicker (t-1)	-0,0703 ***			0,0074						
IndiceName (t-1)						-0,0207				
Secteurs industriels										
Santé	-0,0911			0,1711		0,0917				
Énergie	-0,1835			0,1318		0,0897				
Consommation de base	-0,1700			-0,0821		-0,1911 *				
Vente au détail	0,0829			0,0612		-0,0695				
Industries	0,1085			0,2462		-0,0325				
Télécommunications	0,3699			0,7574 **		-0,1860				
Services publics	-0,3651			0,6675		-0,0951				
Matériaux	0,0427			-0,0090		0,0184				
Finance	-0,0464			0,3888 **		-0,0472				
Technologies de l'information (omis)										
Santé x *Tweet (t-1)*	0,0316			-0,0871		-0,0513				
Énergie x *Tweet (t-1)*	0,0915			-0,0660		-0,0347				
Consommation de base x *Tweet (t-1)*	0,0625			0,0187		0,0655 *				
Vente au détail x *Tweet (t-1)*	-0,0609 *			-0,0535		0,0085				
Industries x *Tweet (t-1)*	-0,0853			-0,1159		-0,0091				
Télécommunications x *Tweet (t-1)*	-0,0894			-0,2206 *		0,0898 *				
Services publics x *Tweet (t-1)*	0,1256			-0,1795		0,0226				
Matériaux x *Tweet (t-1)*	-0,0218			-0,0468		-0,0508				
Finance x *Tweet (t-1)*	0,0233			-0,1372 **		0,0054				
Technologies de l'information (omis)										
Constante	0,3584 ***			-1,5609 ***		0,2821 ***				
Nombre d'observations	16495			16495		17006				
R^2	0,0324			0,0079		0,0303				

Tableau D.2 : résultats des modèles probit avec interaction de variables pour les rendements anormaux. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements anormaux**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)		0,1064 ***							-0,0030	0,0169
IndiceTicker (t-1)		0,0619 **								
IndiceName (t-1)									0,0331	-0,0031
Jours de la semaine										
Lundi		0,1484 *							-0,0500	-0,4783
Mardi (omis)										
Mercredi		-0,0115							-0,0848	-1,2924 *
Jeudi		0,2122 **							0,1437	-0,0637
Vendredi		0,1532 *							0,2181	-0,3596
Lundi x *Tweet (t-1)*		-0,0727 *							-0,0243	0,1172
Mardi x *Tweet (t-1) (omis)*										
Mercredi x *Tweet (t-1)*		-0,0439							0,0063	0,3800 *
Jeudi x *Tweet (t-1)*		-0,0594							-0,1260 *	0,0340
Vendredi x *Tweet (t-1)*		-0,0374							-0,1135	0,1265
Constante		-0,8298 ***							-2,3132 ***	-2,8593 ***
Nombre d'observations		16566							17077	17077
R^2		0,0039							0,0048	0,0203

Variable dépendante	**Rendements anormaux**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	0,1155 ***	0,0949 ***	0,0353 *	0,0133	0,0131		0,0950 ***	0,0408 **	0,0115	0,0165
IndiceTicker (t-1)	0,0216	0,0267	0,0194	-0,0957	-0,2823 **					
IndiceName (t-1)							-0,0244	0,0125	0,1462 **	0,6144 ***
Secteurs industriels										
Santé	0,1771 **	0,4803 ***	-0,2274 **	-0,9974 **	*(omis)*		0,3587 ***	-0,2293 *	-0,2583	
Énergie	-0,1256	-0,0154	-0,0773	-0,2433	*(omis)*		0,7079 ***	-0,4633 **	-1,8375 **	
Consommation de base	0,0612	0,0290	0,1068	0,0344	-0,0920		0,0520	0,0319	0,3647	2,4190 ***
Vente au détail	-0,0441	-0,0447	0,0476	-0,3465 *	-1,0381 **		-0,2475 ***	0,0956	0,6866 **	2,1844 ***
Industries	0,1431	0,0996	0,0425	0,2699	-0,1787		-0,1226	0,1329	1,1339 ***	1,9648 **
Télécommunications	0,1135	0,1942	0,0153	*(omis)*	*(omis)*		0,1325	-0,0916	*(omis)*	*(omis)*
Services publics	-0,2983	0,1752	-0,3744	*(omis)*	*(omis)*		0,2628 **	-0,1767	*(omis)*	*(omis)*
Matériaux	0,1274	-0,0522	0,2878 ***	-0,4018	-1,8809 **		-0,3083 ***	0,1197	0,8586 ***	1,1692
Finance	-0,0025	-0,1612	0,2339 **	-0,5765 *			-0,3410 ***	0,3065 ***	0,4167	
Technologies de l'information (omis)										
Santé x *Tweet (t-1)*	-0,0632 *	-0,1627 ***	0,0744 *	0,2133	*(omis)*		-0,0873 **	0,0652	-0,0726	*(omis)*
Énergie x *Tweet (t-1)*	0,0684	0,0469	0,0029	0,1874	*(omis)*		-0,2494 ***	0,1598 *	0,6817 ***	*(omis)*
Consommation de base x *Tweet (t-1)*	0,0044	-0,0962 **	-0,1226 **	-0,4434	0,1169		0,0778 **	-0,0864 **	-0,3676 **	-0,7188 ***
Vente au détail x *Tweet (t-1)*	0,0043	-0,0082	-0,0157	0,1937 **	0,4503 **		0,0685 **	-0,0302	-0,2095 **	-0,7039 ***
Industries x *Tweet (t-1)*	-0,0704	-0,0844	-0,0097	0,0248	0,2284		0,0121	-0,0387	-0,2994 ***	-0,4680 *
Télécommunications x *Tweet (t-1)*	-0,0231	-0,0030	-0,0464	*(omis)*	*(omis)*		0,0193	-0,0039	*(omis)*	*(omis)*
Services publics x *Tweet (t-1)*	0,0967	0,0429	0,0110	*(omis)*	*(omis)*		0,0367	-0,0973	*(omis)*	*(omis)*
Matériaux x *Tweet (t-1)*	-0,1158 ***	-0,0829 *	-0,1200 ***	0,2814 ***	0,7419 ***		0,0331	-0,0528	-0,2494 **	-0,1355
Finance x *Tweet (t-1)*	-0,0084	0,0361	-0,0769 *	0,2290 *	*(omis)*		0,1196 ***	-0,1118 ***	-0,1373	*(omis)*
Technologies de l'information (omis)										
Constante	-0,0069	-0,7471 ***	-0,6793 ***	-2,1249 ***	-2,2837 ***		-0,6180 ***	-0,6864 ***	-2,7918 ***	-4,9343 ***
Nombre d'observations	16636	16566	16566	15634	11665		17077	17077	16130	12079
R^2	0,0034	0,0118	0,0040	0,0349	0,0411		0,0116	0,0043	0,0388	0,0575

ANNEXE E : IMPACT DES RAPPORTS TRIMESTRIELS : MODÈLES PROBIT

Tableau E.1 : impact des rapports trimestriels sur les rendements journaliers. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements journaliers (une semaine avant les rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)										
IndiceTicker (t-1)										
IndiceName (t-1)										
Rapports trimestriels										
Semaine avant										
Semaine pendant										
Semaine avant x Tweet (t-1)										
Semaine pendant x Tweet (t-1)										
Constante										
Nombre d'observations										
R^2										

Variable dépendante	**Rendements journaliers (pendant la semaine des rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)				-0,1514 **						
IndiceTicker (t-1)				-0,1360 ***						
IndiceName (t-1)										
Rapports trimestriels										
Semaine avant										
Semaine pendant				0,2946						
Semaine avant x Tweet (t-1)										
Semaine pendant x Tweet (t-1)				0,2624 **						
Constante				-2,2419 ***						
Nombre d'observations				16636						
R^2				0,0224						

Tableau E.2 : impact des rapports trimestriels sur les rendements nocturnes. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements nocturnes (une semaine avant les rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)			-0,1489 ***			-0,1496 ***	-0,1019 ***			
IndiceTicker (t-1)			0,0304 *							
IndiceName (t-1)						0,0090	0,0185 **			
Rapports trimestriels										
Semaine avant			-0,3021 *			0,1817 **	0,1576 *			
Semaine pendant										
Semaine avant x Tweet (t-1)			0,1129 *			-0,0794 **	-0,0600 *			
Semaine pendant x Tweet (t-1)										
Constante			-1,4419 ***			0,1034 ***	-0,1386 ***			
Nombre d'observations			16495			17006	17006			
R^2			0,0040			0,0028	0,0015			

Variable dépendante	**Rendements nocturnes (pendant la semaine des rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)		0,1016 ***	-0,1484 ***	-0,0304						
IndiceTicker (t-1)		0,0308 ***	0,0038	-0,0748						
IndiceName (t-1)										
Rapports trimestriels										
Semaine avant										
Semaine pendant		0,0708	0,0789	-0,1292						
Semaine avant x Tweet (t-1)										
Semaine pendant x Tweet (t-1)		-0,1189 ***	0,1133 **	0,4060 ***						
Constante		-0,1560 ***	-1,4215 ***	-2,9240 ***						
Nombre d'observations		16495	16495	16495						
R^2		0,0033	0,0109	0,1624						

Tableau E.3 : impact des rapports trimestriels sur les volumes d'échanges. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Volume (une semaine avant les rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)										
IndiceTicker (t-1)										
IndiceName (t-1)										
Rapports trimestriels										
Semaine avant										
Semaine pendant										
Semaine avant x Tweet (t-1)										
Semaine pendant x Tweet (t-1)										
Constante										
Nombre d'observations										
R^2										

Variable dépendante	**Volume (pendant la semaine des rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)	-0,5045 ***					-0,5183 ***		-0,0760 **		
IndiceTicker (t-1)	-0,0724 ***									
IndiceName (t-1)						-0,0079		0,0392 ***		
Rapports trimestriels										
Semaine avant										
Semaine pendant	0,4882 ***					0,4374 ***		0,0828		
Semaine avant x Tweet (t-1)										
Semaine pendant x Tweet (t-1)	-0,0663 *					-0,0574 *		-0,1170 *		
Constante	0,3311 ***					0,2036 ***		-1,7326 ***		
Nombre d'observations	16495					17006		17006		
R^2	0,0340					0,0322		0,0037		

Tableau E.4 : impact des rapports trimestriels sur les rendements anormaux. Les valeurs présentées correspondent aux effets marginaux des variables étudiées. *** p-value < 0,01 ; ** p-value < 0,05 ; * p-value < 0,1

Variable dépendante	**Rendements anormaux (une semaine avant les rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)										
IndiceTicker (t-1)										
IndiceName (t-1)										
Rapports trimestriels										
Semaine avant										
Semaine pendant										
Semaine avant x Tweet (t-1)										
Semaine pendant x Tweet (t-1)										
Constante										
Nombre d'observations										
R^2										

Variable dépendante	**Rendements anormaux (pendant la semaine des rapports trimestriels)**									
Seuils	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**	**0%+**	**0-1%**	**1-5%**	**5-10%**	**10%+**
Variable dépendante (t-1)		0,1006 ***	0,0314	0,0083			0,1005 ***			
IndiceTicker (t-1)		0,0321 ***	-0,0191	-0,0399						
IndiceName (t-1)							0,0175 *			
Rapports trimestriels										
Semaine avant										
Semaine pendant		0,0985	-0,2259 **	-0,0629			0,0484			
Semaine avant x Tweet (t-1)										
Semaine pendant x Tweet (t-1)		-0,0970 **	0,0659 *	0,1696 **			-0,0769 **			
Constante		-0,7414 ***	-0,6090 ***	-2,2643 ***			-0,7158 ***			
Nombre d'observations		16566	16566	16566			17077			
R^2		0,0024	0,0004	0,0122			0,0023			

ANNEXE F – REQUÊTES UTILISÉES POUR L'ANALYSE DES TWEETS FINANCIERS

Tableau F.1 : liste des requêtes utilisées correspondant aux 400 compagnies sélectionnées du S&P500

Requête	Compagnie	Requête	Compagnie	Requête	Compagnie	Requête	Compagnie
$FB	Facebook	$AAPL	Apple Inc.	$CPB	Campbell Soup	$STZ	Constellation Brands
$TWTR	Twitter	$AMAT	Applied Materials Inc	$COF	Capital One Financial	$GLW	Corning Inc.
$MMM	3M	$ADM	Archer-Daniels-Midland Co	$CAH	Cardinal Health Inc.	$COST	Costco Co.
$HSY	The Hershey Company	$AIZ	Assurant Inc	$CFN	Carefusion	$COV	Covidien plc
$VZ	Verizon	$T	AT&T Inc	$KMX	Carmax Inc	$CCI	Crown Castle International Corp.
$ANF	Abercrombie & Fitch	$ADSK	Autodesk Inc	$CCL	Carnival Corp.	$CSX	CSX Corp.
$ACE	ACE Limited	$ADP	Automatic Data Processing	$CAT	Caterpillar Inc.	$CMI	Cummins Inc.
$ACN	Accenture	$AN	AutoNation Inc	$CBG	CBRE Group	$CVS	CVS Caremark Corp.
$ADBE	Adobe	$AZO	AutoZone Inc	$CELG	Celgene Corp.	$DHI	D. R. Horton
$WFC	Wells Fargo	$AVB	AvalonBay Communities, Inc.	$CNP	CenterPoint Energy	$DHR	Danaher Corp.
$YHOO	Yahoo!	$AVY	Avery Dennison Corp	$CTL	CenturyLink Inc	$DRI	Darden Restaurants
$XRX	Xerox	$AVP	Avon Products	$CERN	Cerner	$DVA	DaVita Inc.
$A	Agilent Technologies Inc.	$BHI	Baker Hughes Inc	$CF	CF Industries Holdings Inc	$DE	Deere & Co.
$GAS	AGL Resources	$BLL	Ball Corp	$SCHW	Charles Schwab	$DELL	Dell Inc.
$APD	Air Products & Chemicals Inc	$BAC	Bank of America Corp	$CHK	Chesapeake Energy	$DLPH	Delphi Automotive
$ARG	Airgas Inc	$BK	The Bank of New York Mellon Corp.	$CVX	Chevron Corp.	$DAL	Delta Airlines
$AKAM	Akamai Technologies Inc	$BCR	Bard (C.R.) Inc.	$CMG	Chipotle Mexican Grill	$DNR	Denbury Resources Inc.
$AA	Alcoa Inc	$BAX	Baxter International Inc.	$CB	Chubb Corp.	$XRAY	Dentsply International
$ALXN	Alexion Pharmaceuticals	$BBT	BB&T Corporation	$CI	CIGNA Corp.	$DVN	Devon Energy Corp.
$ATI	Allegheny Technologies Inc	$BEAM	Beam Inc.	$CINF	Cincinnati Financial	$DO	Diamond Offshore Drilling
$AGN	Allergan Inc	$BDX	Becton Dickinson	$CTAS	Cintas Corporation	$DTV	DirecTV
$YNDX	Allstate Corp	$BBBY	Bed Bath & Beyond	$CSCO	Cisco Systems	$DFS	Discover Financial Services
$ALTR	Altera Corp	$BMS	Bemis Company	$C	Citigroup Inc.	$DISCA	Discovery Communications
$MO	Altria Group Inc	$BRK	Berkshire Hathaway	$CTXS	Citrix Systems	$DG	Dollar General
$MZN	Amazon.com Inc	$BBY	Best Buy Co. Inc.	$CLF	Cliffs Natural Resources	$DLTR	Dollar Tree
$AEE	Ameren Corp	$BIIB	BIOGEN IDEC Inc.	$CLX	The Clorox Company	$D	Dominion Resources
$AEP	American Electric Power	$BLK	BlackRock	$CME	CME Group Inc.	$DOV	Dover Corp.
$AXP	American Express Co	$HRB	Block H&R	$CMS	CMS Energy	$DOW	Dow Chemical
$AIG	American Intl Group Inc	$BA	Boeing Company	$COH	Coach Inc.	$DPS	Dr Pepper Snapple Group
$AMT	American Tower Corp A	$BWA	BorgWarner	$KO	The Coca Cola Company	$DTE	DTE Energy Co.
$AMP	Ameriprise Financial	$BXP	Boston Properties	$CCE	Coca-Cola Enterprises	$DD	Du Pont (E.I.)
$ABC	AmerisourceBergen Corp	$BSX	Boston Scientific	$CTSH	Cognizant Technology Solutions	$DUK	Duke Energy
$AME	Amgen Inc	$BMY	Bristol-Myers Squibb	$CL	Colgate-Palmolive	$DNB	Dun & Bradstreet
$AMGN	Amphenol Corp A	$BRCM	Broadcom Corporation	$CMCSA	Comcast Corp.	$ETFC	E-Trade
$APH	Anadarko Petroleum Corp	$BF	Brown-Forman Corporation	$CMA	Comerica Inc.	$EMN	Eastman Chemical
$APC	Analog Devices, Inc.	$CHRW	C. H. Robinson Worldwide	$CSC	Computer Sciences Corp.	$ETN	Eaton Corp.
$ADI	Aon plc	$CA	CA, Inc.	$CAG	ConAgra Foods Inc.	$EBAY	eBay Inc.
$AON	Apache Corporation	$CVC	Cablevision Systems Corp.	$COP	ConocoPhillips	$ECL	Ecolab Inc.
$APA	Apartment Investment & Mgmt	$COG	Cabot Oil & Gas	$CNX	CONSOL Energy Inc.	$EIX	Edison Int'l
$AIV	Apollo Group Inc	$CAM	Cameron International Corp.	$ED	Consolidated Edison	$EW	Edwards Lifesciences

Requête	Compagnie	Requête	Compagnie	Requête	Compagnie	Requête	Compagnie
$EA	Electronic Arts	$GE	General Electric	$INTU	Intuit Inc.	$MTB	M&T Bank Corp.
$EMC	EMC Corp.	$GIS	General Mills	$ISRG	Intuitive Surgical Inc.	$MAC	Macerich
$EMR	Emerson Electric	$GM	General Motors	$IVZ	Invesco Ltd.	$M	Macy's Inc.
$ESV	Ensco plc	$GPC	Genuine Parts	$IRM	Iron Mountain Incorporated	$MRO	Marathon Oil Corp.
$ETR	Entergy Corp.	$GNW	Genworth Financial Inc.	$JBL	Jabil Circuit	$MPC	Marathon Petroleum
$EOG	EOG Resources	$GILD	Gilead Sciences	$JEC	Jacobs Engineering Group	$MAR	Marriott Int'l.
$EQT	EQT Corporation	$GS	Goldman Sachs Group	$JDSU	JDS Uniphase Corp.	$MMC	Marsh & McLennan
$EFX	Equifax Inc.	$GT	Goodyear Tire & Rubber	$JNJ	Johnson & Johnson	$MAS	Masco Corp.
$EQR	Equity Residential	$GOOG	Google Inc.	$JCI	Johnson Controls	$MA	Mastercard Inc.
$EL	Estee Lauder Cos.	$GWW	Grainger (W.W.) Inc.	$JOY	Joy Global Inc.	$MAT	Mattel Inc.
$EXC	Exelon Corp.	$HAL	Halliburton Co.	$JPM	JPMorgan Chase & Co.	$MKC	McCormick & Co.
$EXPE	Expedia Inc.	$HOG	Harley-Davidson	$JNPR	Juniper Networks	$MCD	McDonald's Corp.
$EXPD	Expeditors Int'l	$HAR	Harman Int'l Industries	$KSU	Kansas City Southern	$MHFI	McGraw-Hill
$ESRX	Express Scripts	$HRS	Harris Corporation	$K	Kellogg Co.	$MCK	McKesson Corp.
$XOM	Exxon Mobil Corp.	$HIG	Hartford Financial Svc.Gp.	$KEY	KeyCorp	$MJN	Mead Johnson
$FFIV	F5 Networks	$HAS	Hasbro Inc.	$KMB	Kimberly-Clark	$MWV	MeadWestvaco Corporation
$FDO	Family Dollar Stores	$HCP	HCP Inc.	$KIM	Kimco Realty	$MDT	Medtronic Inc.
$FAST	Fastenal Co	$HCN	Health Care REIT	$KLAC	KLA-Tencor Corp.	$MRK	Merck & Co.
$FDX	FedEx Corporation	$HP	Helmerich & Payne	$KSS	Kohl's Corp.	$MET	MetLife Inc.
$FIS	Fidelity National Information Services	$HES	Hess Corporation	$KRFT	Kraft Foods Group	$MCHP	Microchip Technology
$FITB	Fifth Third Bancorp	$HPQ	Hewlett-Packard	$KR	Kroger Co.	$MU	Micron Technology
$FSLR	First Solar Inc	$HD	Home Depot	$LTD	L Brands Inc.	$MSFT	Microsoft Corp.
$FE	FirstEnergy Corp	$HON	Honeywell Int'l Inc.	$LLL	L-3 Communications Holdings	$MOLX	Molex Inc.
$FISV	Fiserv Inc	$HRL	Hormel Foods Corp.	$LH	Laboratory Corp. of America Holding	$TAP	Molson Coors Brewing Company
$FLS	Flowserve Corporation	$HSP	Hospira Inc.	$LRCX	Lam Research	$MDLZ	Mondelez International
$FLR	Fluor Corp.	$HST	Host Hotels & Resorts	$LM	Legg Mason	$MON	Monsanto Co.
$FMC	FMC Corporation	$HCBK	Hudson City Bancorp	$LEG	Leggett & Platt	$MNST	Monster Beverage
$FTI	FMC Technologies Inc.	$HUM	Humana Inc.	$LEN	Lennar Corp.	$MCO	Moody's Corp
$F	Ford Motor	$HBAN	Huntington Bancshares	$LUK	Leucadia National Corp.	$MS	Morgan Stanley
$FRX	Forest Laboratories	$ITW	Illinois Tool Works	$LIFE	Life Technologies	$MOS	The Mosaic Company
$FOSL	Fossil, Inc.	$IT	Ingersoll-Rand PLC	$LLY	Lilly (Eli) & Co.	$MSI	Motorola Solutions Inc.
$BEN	Franklin Resources	$TEG	Integrys Energy Group Inc.	$LNC	Lincoln National	$MUR	Murphy Oil
$FCX	Freeport-McMoran Cp & Gld	$INTC	Intel Corp.	$LLTC	Linear Technology Corp.	$MYL	Mylan Inc.
$FTR	Frontier Communications	$ICE	IntercontinentalExchange Inc.	$LMT	Lockheed Martin Corp.	$NBR	Nabors Industries Ltd.
$GME	GameStop Corp.	$IBM	International Bus. Machines	$L	Loews Corp.	$NDAQ	NASDAQ OMX Group
$GCI	Gannett Co.	$IGT	International Flav/Frag	$LO	Lorillard Inc.	$NOV	National Oilwell Varco Inc.
$GPS	Gap (The)	$IP	International Game Technology	$LOW	Lowe's Cos.	$NTAP	NetApp
$GRMN	Garmin Ltd.	$IPG	International Paper	$LSI	LSI Corporation	$NFLX	NetFlix Inc.
$GD	General Dynamics	$IFF	Interpublic Group	$LYB	LyondellBasell	$NWL	Newell Rubbermaid Co.

Requête	Compagnie	Requête	Compagnie	Requête	Compagnie	Requête	Compagnie
$NFX	Newfield Exploration Co	$OKE	ONEOK	$PNW	Pinnacle West Capital	$PWR	Quanta Services Inc.
$NEM	Newmont Mining Corp. (Hldg. Co.)	$ORCL	Oracle Corp.	$PXD	Pioneer Natural Resources	$QCOM	QUALCOMM Inc.
$NWSA	News Corporation	$OI	Owens-Illinois Inc	$PBI	Pitney-Bowes	$DGX	Quest Diagnostics
$NEE	NextEra Energy Resources	$PCG	PG&E Corp.	$PCL	Plum Creek Timber Co.	$RRC	Range Resources Corp.
$NLSN	Nielsen Holdings N.V.	$PCAR	PACCAR Inc.	$PNC	PNC Financial Services	$RTN	Raytheon Co.
$NKE	NIKE Inc.	$PLL	Pall Corp.	$RL	Polo Ralph Lauren Corp.	$RHT	Red Hat Inc.
$NI	NiSource Inc.	$PH	Parker-Hannifin	$PPG	PPG Industries	$REGN	Regeneron
$NE	Noble Corp	$PDCO	Patterson Companies	$PPL	PPL Corp.	$RF	Regions Financial Corp.
$NBL	Noble Energy Inc	$PAYX	Paychex Inc.	$PX	Praxair Inc.	$RSG	Republic Services Inc
$JWN	Nordstrom	$BTU	Peabody Energy	$PCP	Precision Castparts	$RAI	Reynolds American Inc.
$NSC	Norfolk Southern Corp.	$JCP	Penney (J.C.)	$PCLN	Priceline.com Inc	$RHI	Robert Half International
$NTRS	Northern Trust Corp.	$PNR	Pentair Ltd.	$PFG	Principal Financial Group	$ROK	Rockwell Automation Inc.
$NOC	Northrop Grumman Corp.	$PBCT	People's United Bank	$PG	Procter & Gamble	$COL	Rockwell Collins
$NU	Northeast Utilities	$POM	Pepco Holdings Inc.	$PGR	Progressive Corp.	$ROP	Roper Industries
$NRG	NRG Energy	$PEP	PepsiCo Inc.	$PLD	Prologis	$ROST	Ross Stores
$NUE	Nucor Corp.	$PKI	PerkinElmer	$PRU	Prudential Financial	$TGT	Target
$NVDA	Nvidia Corporation	$PRGO	Perrigo	$PEG	Public Serv. Enterprise Inc.	$R	Ryder System
$NYX	NYSE Euronext	$PETM	PetSmart, Inc.	$PSA	Public Storage	$SWY	Safeway Inc.
$ORLY	O'Reilly Automotive	$PFE	Pfizer Inc.	$PHM	Pulte Homes Inc.	$SBUX	Starbucks
$OXY	Occidental Petroleum	$PM	Philip Morris International	$PVH	PVH Corp.	$SNDK	SanDisk Corporation
$OMC	Omnicom Group	$PSX	Phillips 66	$QEP	QEP Resources	$SLB	Schlumberger