Three essays on the informational efficiency of financial markets through the use of Big Data Analytics

Thomas Renault

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

ECOLE DOCTORALE - SCIENCES DE GESTION

LABORATOIRE PRISM

Three Essays on the Informational Efficiency of Financial Markets through the use of Big Data Analytics

---

Présentée et soutenue publiquement le 6 septembre 2017 en vue de l'obtention du Doctorat en Sciences de Gestion par THOMAS RENAULT

Directeur de thèse:
ROLAND GILLET, Professeur à l'Université Paris 1 Panthéon-Sorbonne

Rapporteurs:
ALAIN DURRÉ, Professeur Associé à l'IÉSEG School of Management
JEAN-FRANÇOIS GAJEWSKI, Professeur à l'Université Savoie Mont Blanc

Suffragants:
PETER POPE, Professeur à la London School of Economics
JEAN-PAUL LAURENT, Professeur à l'Université Paris 1 Panthéon-Sorbonne

"*The history of thought in financial markets has shown a surprising lack of consensus about a very fundamental question: what ultimately causes all those fluctuations in the price of speculative assets like corporate stocks, commodities, or real estate? One might think that so basic a question would have long ago been confidently answered. But the answer to this question is not so easily found.*"

Robert J. Shiller, 2014

Résumé

L'augmentation massive du volume de données générées chaque jour par les individus sur Internet offre aux chercheurs la possibilité d'aborder la question de la prédictibilité des marchés financiers sous un nouvel angle. Sans prétendre apporter une réponse définitive au débat entre les partisans de l'efficience des marchés et les chercheurs en finance comportementale, cette thèse vise à améliorer notre compréhension du processus de formation des prix sur les marchés financiers grâce à une approche Big Data.

Plus précisément, cette thèse porte sur (1) la mesure du sentiment des investisseurs à fréquence intra-journalière, et le lien entre le sentiment des investisseurs et les rendements agrégés du marché, (2) la mesure de l'attention des investisseurs aux informations économiques et financières en temps réel, et la relation entre l'attention des investisseurs et la dynamique des prix des actions des sociétés à forte capitalisation, et enfin, (3) la détection des comportements suspicieux pouvant amoindrir le rôle informationnel des marchés financiers, et le lien entre le volume d'activité sur les réseaux sociaux et le prix des actions des entreprises de petite capitalisation.

Le premier essai propose une méthodologie permettant de construire un nouvel indicateur du sentiment des investisseurs en analysant le contenu des messages publiés sur le réseau social StockTwits. En examinant les caractéristiques propres à chaque utilisateur (niveau d'expérience, approche d'investissement, période de détention), cet essai fournit des preuves empiriques montrant que le comportement des investisseurs naïfs, sujets à des périodes d'excès d'optimisme ou de pessimisme, a un impact sur la valorisation du marché action, et ce en accord avec les théories de la finance comportementale.

Le deuxième essai propose une méthodologie permettant de mesurer l'attention des investisseurs aux informations en temps réel, en combinant les données des médias traditionnels avec le contenu des messages envoyés par une liste d'experts sur la plateforme Twitter. Cet essai démontre que lorsqu'une

information attire l'attention des investisseurs, les mouvements de marchés sont caractérisés par une forte hausse des volumes échangés, une hausse de la volatilité et des sauts de prix. Cet essai démontre également qu'il n'y a pas de fuite d'information significative lorsque les sources d'informations sont combinées pour corriger un potentiel problème d'horodatage.

Le troisième essai étudie le risque de manipulation informationnelle en examinant un nouveau jeu de données de messages publiés sur Twitter à propos des entreprises de petite capitalisation. Cet essai propose une nouvelle méthodologie permettant d'identifier les comportements anormaux de manière automatisée en analysant les interactions entre les utilisateurs. Etant donné le grand nombre de recommandations suspicieuses d'achat envoyées par certains groupes d'utilisateurs, l'analyse empirique et les conclusions de cet essai soulignent la nécessité d'un plus grand contrôle par les régulateurs de l'information publiée sur les réseaux sociaux ainsi que l'utilité d'une meilleure éducation des investisseurs individuels.

Mots clés: Evaluation des actifs financiers, Sentiment des investisseurs, Attention des investisseurs, Manipulation de marché, Efficience des marchés, Finance comportementale, Réseaux sociaux, Analyse textuelle, Machine learning

Abstract

The massive increase in the availability of data generated everyday by individuals on the Internet has made it possible to address the predictability of financial markets from a different perspective. Without making the claim of offering a definitive answer to a debate that has persisted for forty years between partisans of the efficient market hypothesis and behavioral finance academics, this dissertation aims to improve our understanding of the price formation process in financial markets through the use of Big Data analytics.

More precisely, it analyzes: (1) how to measure intraday investor sentiment and determine the relation between investor sentiment and aggregate market returns, (2) how to measure investor attention to news in real time, and identify the relation between investor attention and the price dynamics of large capitalization stocks, and (3) how to detect suspicious behaviors that could undermine the informational role of financial markets, and determine the relation between the level of posting activity on social media and small-capitalization stock returns.

The first essay proposes a methodology to construct a novel indicator of investor sentiment by analyzing an extensive dataset of user-generated content published on the social media platform StockTwits. Examining users' self-reported trading characteristics, the essay provides empirical evidence of sentiment-driven noise trading at the intraday level, consistent with behavioral finance theories

The second essay proposes a methodology to measure investor attention to news in real-time by combining data from traditional newswires with the content published by experts on the social media platform Twitter. The essay demonstrates that news that garners high attention leads to large and persistent change in trading activity, volatility, and price jumps. It also demonstrates that the pre-announcement effect is reduced when corrected newswire timestamps are considered.

The third essay provides new insights into the empirical literature on small capitalization stocks

market manipulation by examining a novel dataset of messages published on the social media platform Twitter. The essay proposes a novel methodology to identify suspicious behaviors by analyzing interactions between users and provide empirical evidence of suspicious stock recommendations on social media that could be related to market manipulation. The conclusion of the essay should reinforce regulators' efforts to better control social media and highlights the need for a better education of individual investors.

Keywords: Asset pricing, Investor sentiment, Investor attention, Market manipulation Efficient market hypothesis, Behavioral finance, Social media, Textual analysis, Machine learning

Remerciements

Tout d'abord, je tiens à exprimer ma sincère gratitude à mon directeur de thèse, le Professeur Roland Gillet, pour son soutien continu et les nombreux échanges passionnants que nous avons eus au cours de ces dernières années. Je le remercie tout particulièrement pour la confiance et la liberté de recherche qu'il m'a accordées. Grâce à ses conseils et sa bienveillance, j'ai pu m'épanouir personnellement et intellectuellement dans mon travail de thèse, et, pour cela, je lui en suis extrêmement reconnaissant.

J'aimerais aussi exprimer ma reconnaissance toute particulière au Professeur Alain Durré, grâce à qui cette thèse a pu voir le jour. Je le remercie de m'avoir ouvert l'esprit sur le monde de la recherche et d'avoir toujours été présent pour m'aider à avancer. Nos nombreuses discussions m'ont poussées à viser l'excellence et à avancer avec rigueur et ambition dans mon travail, et pour cela, je lui exprime toute ma gratitude.

J'adresse aussi mes remerciements au Professeur Jean-François Gajewski pour ses précieux commentaires lors de ma pré-soutenance de thèse, ainsi qu'au Professeur Jean-Paul Laurent au Professeur Peter Pope pour avoir accepté d'être membre de mon jury de thèse. C'est un réel honneur de soutenir ma thèse devant un tel jury et c'est pourquoi je tiens ici à tous les en remercier.

Mes sincères remerciements vont aussi à l'institution IÉSEG School of Management pour le soutien financier et matériel accordé, et tout particulièrement à Jean-Philippe Ammeux et à Anna Canato pour la confiance qu'ils m'ont temoignée. Merci aussi à Sandra et Abida pour le support administratif apporté lors de mes innombrables demandes de financement pour des soumissions de papiers ou des conférences. Je tiens également à remercier Angela Armakolla pour les relectures au milieu de la

nuit et Matthieu Picault pour nos échanges stimulants sous perfusion de caféine et pour ses nombreux commentaires sur mes recherches.

Un grand merci à tous mes amis de m'avoir accompagné (et supporté) dans cette longue aventure. Plus particulièrement, et en acronyme comme à mon habitude, à AL, AM, DM, JdC, MO, MT, VT, TV, ALS, CL, CD, HC, SG, J2M et OF, qui se reconnaitront. Enfin une pensée toute particulière pour mes parents ainsi que pour mes frères et sœur. Merci de m'avoir choyé, de m'avoir fait comprendre qu'on pouvait s'amuser en travaillant (ou l'inverse), et de m'avoir toujours soutenu et encouragé dans mes choix.

Contents

List of Figures

List of Tables

Introduction générale

Malgré plusieurs décennies de recherches empiriques et théoriques, il n'existe toujours pas de consensus au sein de la communauté académique en ce qui concerne la prédictibilité des marchés financiers. Selon l'hypothèse d'efficience des marchés (Fama, 1965), l'information est instantanément intégrée dans les prix. Le comportement des investisseurs irrationnels est neutralisé par celui des arbitragistes rationnels qui corrigent immédiatement toutes anomalies. Il est donc impossible de réaliser un profit ajusté du niveau de risque en investissant sur les marchés (Jensen, 1978). Pour reprendre une ancienne plaisanterie largement répandue parmi les économistes à propos de l'hypothèse d'efficience des marchés, si vous voyez un billet de 100$ sur le trottoir, il n'est même pas nécessaire de vous pencher pour essayer de le ramasser, car, s'il s'agissait d'un véritable billet de 100$, quelqu'un l'aurait déjà pris à votre place (Lo, 2008).

La recherche académique a cependant fortement évolué depuis l'époque où l'hypothèse d'efficience des marchés était établie au-delà de tout doute raisonnable (Shiller, 2003). Appliquant des principes issus de la psychologie et de la sociologie aux marchés financiers, les chercheurs en finance comportementale ont depuis rapporté un certain nombre d'irrégularités semblant contredire l'hypothèse d'efficience des marchés. Selon ces derniers, les investisseurs ne sont pas des robots aux comportements parfaitement rationnels mais des agents aux capacités cognitives limitées, pouvant être sujets à des périodes d'excès d'optimisme ou de pessimisme (Baker & Wurgler, 2007; Li & Yu, 2012). D'une part, en raison de limites à l'arbitrage, les prix peuvent donc diverger significativement de leurs valeurs fondamentales en présence d'investisseurs irrationnels dont les stratégies d'investissement sont motivées par leurs sentiments (De Long et al., 1990). D'autre part, l'attention étant une ressource rare (Kahneman, 1973), le niveau d'attention des investisseurs lors de l'arrivée d'une nouvelle information peut entraîner une sous-réaction ou une sur-réaction des prix boursiers (Huberman & Regev, 2001; Barber & Odean, 2008).

Sans prétendre apporter une réponse définitive à un débat vieux de plus de 40 ans entre les partisans de l'efficience des marchés et les chercheurs en finance comportementale, l'objectif de cette thèse est d'améliorer notre compréhension du processus de formation des prix grâce à une approche *Big Data*. L'augmentation massive du volume et de la variété des données disponibles permettent en effet d'aborder le problème de la prédictibilité des marchés financiers sous un nouvel angle. A la manière de l'apparition du microscope en microbiologie (King, 2011), les chercheurs et les praticiens en science sociales ont dorénavant accès à plusieurs milliards de données granulaires générées chaque jour par les individus sur les réseaux sociaux, les blogs, et les forums de discussion. Les comportements humains, tel que le sentiment ou l'attention, peuvent désormais être analysés à l'échelle individuelle. Selon Einav & Levin (2014), l'expansion du volume de données collectées à propos de l'activité économique et sociale devrait avoir des effets profonds sur la manière dont est conduite la recherche empirique en économie. Au travers des trois essais constituant cette thèse, nous allons démontrer pourquoi nous pensons que la "révolution des données" affectera aussi profondément le monde de la recherche en finance.

Cette thèse apporte des contributions méthodologiques et empiriques sur trois thématiques relatives à l'efficience informationnelle des marchés financiers. Plus précisément, nous avons focalisé notre attention sur (1) la mesure du sentiment des investisseurs à fréquence intra-journalière, et le lien entre le sentiment des investisseurs et les rendements agrégés du marché, (2) la mesure de l'attention des investisseurs aux informations économiques et financières en temps réel, et la relation entre l'attention des investisseurs et la dynamique des prix actions des sociétés à forte capitalisation, et enfin, (3) la détection des comportements suspicieux pouvant amoindrir le rôle informationnel des marchés financiers, et le lien entre le volume d'activité sur les réseaux sociaux et le prix des actions des entreprises de petite capitalisation.

A cet égard, et pour chacun des trois essais, nous avons construit des jeux de données uniques à partir de messages publiés sur les réseaux sociaux afin de créer de nouveaux indicateurs permettant (1) de mesurer le sentiment des investisseurs, (2) d'analyser l'attention des investisseurs aux informations publiées dans les médias, et (3) de détecter les recommandations d'achats suspicieuses pouvant se rapprocher de tentatives de manipulations. A l'aide de différents outils, tel que l'analyse textuelle, la

théorie des réseaux, les régressions prédictives ou les études d'évènements, cette thèse apporte des preuves empiriques établissant que le contenu publié sur les réseaux sociaux permet d'améliorer notre compréhension du processus de formation des prix sur le marché action.

En plus de son intérêt pour la recherche académique traitant de l'efficience informationnelle des marchés, des dynamiques de trading à haute-fréquence, et des manipulations de marché, les conclusions de ces trois essais nous semblent aussi être pertinentes pour les praticiens et les régulateurs. D'un côté, les professionnels de la finance pourraient trouver intéressant les résultats empiriques de stratégies d'investissement utilisant des nouvelles mesures du sentiment et de l'attention des investisseurs en tant que signaux de trading. De nombreux articles de presse rapportent un usage croissant des réseaux sociaux par les participants de marchés, par exemple pour créer un indicateur du sentiment de marché à partir des messages publiés sur les réseaux sociaux[1] ou bien pour mettre en place des stratégies de trading automatisées analysant les tweets de Donald Trump et en plaçant automatiquement des ordres en fonction des entreprises nommées par le président américain.[2] Cependant, en dehors de quelques exemples anecdotiques, la profitabilité réelle de telles stratégies de trading est inconnue.

Enfin, les régulateurs trouveront dans le troisième essai de cette thèse différents arguments plaidant pour un plus fort contrôle de l'information publiée sur les réseaux sociaux afin d'éviter les manipulations de marché. Bien que les réseaux sociaux puissent bénéficier aux investisseurs et améliorer la transparence des marchés, ces nouveaux moyens de communication offrent aussi aux fraudeurs de nouvelles opportunités pour manipuler les marchés en diffusant de fausses informations. En cela, la Security and Exchange Commission (SEC) et l'Autorité des Marchés Financiers (AMF) pourraient voir en la méthodologie d'analyse de réseaux que nous proposons un outil novateur permettant d'identifier les comportements suspicieux de manière automatisée.

Nous restons bien évidemment à la disposition de l'ensemble des professionnels (académiques, investisseurs, régulateurs) souhaitant obtenir davantage d'informations sur les méthodes utilisées et les résultats empiriques présentés tout au long de cette thèse.[3]

[1] "Firms Analyze Tweets to Gauge Stock Sentiment" (The Wall Street Journal, 6 juillet 2015).
[2] "A Little Birdie Told Me: Playing the Market on Trump Tweets" (The New York Times, 8 février 2017).
[3] Mail: thomas.renault@univ-paris1.fr

1 Premier essai

Selon Baker & Wurgler (2007), la question n'est plus de savoir, comme c'était le cas il y a quelques décennies, si le sentiment des investisseurs affecte le prix des actifs financiers, mais plutôt de définir comment mesurer le sentiment des investisseurs et en quantifier les effets. Depuis les travaux fondateurs de Tumarkin & Whitelaw (2001) et Antweiler & Frank (2004), les chercheurs en finance comportementale ont porté une attention particulière à la construction de nouveaux indicateurs de sentiment en utilisant des données en provenance d'Internet. Extraire et analyser plusieurs millions de messages publiés sur Internet, pour ensuite créer un indicateur du sentiment des investisseurs en temps réel, peut, à première vue, sembler attrayant afin de réduire certains biais de mesure dus à l'échantillonnage des répondants (indicateur basé sur les sondages), à la présence d'un composant idiosyncratique non lié au sentiment (indicateur basé sur les données de marchés) ou bien à un lien de causalité difficile à établir (indicateur basé sur les données textuelles des médias). Cependant, bien que des résultats encourageants aient été identifiés sur les entreprises de faible capitalisation boursière (Sabherwal et al., 2011; Leung & Ton, 2015), les résultats empiriques sur les sociétés à forte capitalisation et sur les indices boursiers sont quant à eux mitigés (Nardo et al., 2016).

Dans le premier essai de cette thèse, intitulé "*Intraday online investor sentiment and return patterns in the U.S. stock market*", nous réexaminons la relation entre, d'un côté, le sentiment individuel des messages publiés sur les réseaux sociaux et, de l'autre côté, le rendement des indices boursiers américains. Plus particulièrement, notre étude se focalise sur les dynamiques intra-journalières à partir de l'analyse d'une base de données de plus de 60 millions de messages publiés par des investisseurs sur le réseau social StockTwits. L'utilisation de ce nouveau jeu de données offre de nombreux avantages. Premièrement, StockTwits propose aux investisseurs de révéler explicitement le sentiment associé à chaque message publié sur la plateforme (positif/bullish ou négatif/bearish). Cette fonctionnalité permet d'éviter d'introduire un biais de subjectivité inhérent aux études se basant sur un jeu de données classifié manuellement par les auteurs. Deuxièmement, lors de l'inscription sur la plateforme, StockTwits demande aux investisseurs de renseigner leur niveau d'expérience (novice, intermédiaire, expert), leur approche d'investissement (technique, fondamentale, momentum...) et leur période moyenne de détention (intra-journalière, journalière, long terme...). La disponibilité de ces

données et la possibilité d'analyser l'hétérogénéité de la base d'investisseurs permet ensuite de relier plus simplement les résultats empiriques aux différentes théories financières (efficience des marchés, finance comportementale, présence d'investisseurs naïfs ou irrationnels...).

La première contribution de cet essai concerne la méthodologie utilisée afin d'obtenir un indicateur du sentiment des investisseurs à partir du contenu publié sur les réseaux sociaux. Au travers différents tests, nous montrons empiriquement que la mesure du sentiment des investisseurs est fortement dépendante de l'approche choisie pour transformer chaque document en une variable quantitative de sentiment. Les approches traditionnelles basées sur l'utilisation d'un dictionnaire générique sont à cet effet inappropriées pour classifier les messages courts et informels publiés sur les réseaux sociaux. Etant donné ce résultat, nous proposons une nouvelle approche conçue spécifiquement pour analyser le contenu des messages publiés sur les réseaux sociaux, et ce afin d'améliorer la classification et la qualité des indicateurs de sentiment. Cette méthode se base sur la construction automatique d'un nouveau lexique pondéré de mots positifs et négatifs utilisés par les investisseurs lorsqu'ils partagent leurs opinions et idées à propos des marchés financiers sur les réseaux sociaux. Pour faciliter la réplicabilité des résultats et encourager la recherche dans ce domaine, l'ensemble des scripts, outils d'analyses, et lexiques utilisés dans cet essai sont disponibles en ligne.[4]

La seconde contribution de cet essai porte sur la relation entre le sentiment des investisseurs et les rendements du marché action. En divisant chaque journée de trading (9h30-16h) en 13 intervalles de temps de 30 minutes, nous démontrons ainsi que la variation du sentiment des investisseurs durant les 30 premières minutes de la journée permet de prévoir le rendement du marché durant les 30 dernières minutes de la journée. Après contrôle des rendements passés du marché et des annonces de chiffres macro-économiques, la variation du sentiment des investisseurs demeure le seul prédicteur de l'évolution du prix des indices boursiers américains durant la dernière demi-heure de trading de la journée. Ensuite, en examinant les caractéristiques propres à chaque utilisateur (niveau d'expérience, approche d'investissement, période de détention), nous établissons que cet effet est principalement causé par les variations du sentiment des investisseurs inexpérimentés. A notre connaissance, cet essai est ainsi le premier à fournir des résultats empiriques montrant que le comportement des investisseurs

[4]http://www.thomas-renault.com

naïfs, sujets à des périodes d'excès d'optimisme ou de pessimisme, a un impact sur la valorisation du marché action, en accord avec les théories de la finance comportementale.

Cet essai s'inscrit dans le cadre de plusieurs champs de recherche. Premièrement, la méthodologie proposée afin de construire automatiquement un lexique pondéré à partir d'un vaste jeu de messages pré-classifiés publiés sur StockTwits s'insère dans la littérature grandissante relative à l'analyse textuelle appliquée à la finance (Loughran & McDonald, 2011; Kearney & Liu, 2014; Das, 2014). A ce propos, nous mettons en avant la nécessité de développer des méthodes spécifiques selon le type de texte étudié, et nous soulignons qu'une méthodologie transparente et réplicable apporte des résultats qualitativement similaires à des méthodologies "boîtes noires" plus complexes (machine learning). Deuxièmement, les conclusions présentées dans cet essai à propos de la prédictibilité de la dernière demi-heure d'échange nous semblent pertinentes pour la littérature empirique sur la dynamique des prix à haute-fréquence (Heston et al., 2010; Sun et al., 2016). Nous émettons à ce sujet l'hypothèse que l'effet "momentum" à l'échelle intra-journalière, documenté récemment par Gao et al. (2017), est en réalité un effet "sentiment" dû à la présence d'investisseurs peu sophistiqués. En cela, une analyse plus détaillée de cette dynamique constituerait, selon nous, une piste de recherche future intéressante. Enfin, cet essai s'inscrit dans la continuité de littérature utilisant des données nouvelles en provenance d'Internet et de réseaux sociaux afin de prévoir l'évolution des marchés financiers (voir Nardo et al., 2016, pour une revue de la littérature). Contrairement aux études peu concluantes mesurant le sentiment des investisseurs à partir des messages publiés sur les forums de discussions Yahoo! Finance et Raging Bull (Antweiler & Frank, 2004; Leung & Ton, 2015), les résultats présentés dans cet essai montrent que les réseaux sociaux permettent de créer des indicateurs fiables du sentiment des investisseurs à échelle intra-journalière. Etant donné le très fort volume de messages publiés chaque jour sur les réseaux sociaux (plusieurs centaines de millions de messages) et la disponibilité d'informations détaillées sur les utilisateurs, nous pensons donc que la recherche académique devrait porter une attention toute particulière à ces nouvelles données. Nous travaillons d'ailleurs actuellement à une extension de ces résultats en considérant l'effet du sentiment sur d'autres classes d'actifs et autour d'annonce macro-économiques spécifiques.

2 Deuxième essai

Selon l'hypothèse d'efficience des marchés, toute l'information est instantanément et parfaitement intégrée dans le prix des actifs financiers. D'un point de vue pratique, cela implique donc que les investisseurs (ou, tout au moins, que certains investisseurs) soient capables d'identifier en temps réel les informations importantes parmi les milliers de nouvelles économiques et financières publiées chaque jour par les médias traditionnels. L'immédiateté à laquelle l'information doit être intégrée dans les prix semble cependant difficile à réconcilier avec la rareté des ressources cognitives des investisseurs. A cet égard, le niveau d'attention apporté par les investisseurs à une nouvelle pourrait donc avoir un impact sur la vitesse à laquelle une information est intégrée dans les prix, et, éventuellement, causer des sur-réactions ou des sous-réactions des prix boursiers.

Dans ce deuxième essai, intitulé "*Market reaction to news and investor attention in real time*", nous, avec mes co-auteurs Deniz Erdemlioglu et Roland Gillet, examinons la réaction des marchés à haute-fréquence en proposant pour cela un nouveau cadre permettant de mesurer en temps réel l'attention des investisseurs aux nouvelles économiques non-anticipées. Dans la littérature, différentes approches indirectes ont été utilisées pour mesurer l'attention des investisseurs, comme par exemple le volume d'échange sur les marchés, la présence d'un plus haut annuel ou d'un rendement extrême, ou bien encore le niveau de couverture médiatique. Cependant, l'utilisation de ces indicateurs comporte différents biais pouvant rendre complexe l'analyse de causalité entre le niveau d'attention et l'évolution des marchés financiers. D'un côté, les données de marchés (comme les volumes échangés ou les rendements extrêmes) contiennent une composante idiosyncratique qui n'est pas liée à l'attention des investisseurs. De l'autre, considérer simplement le niveau de couverture médiatique ne permet pas de prendre en compte l'importance de certaines nouvelles et peut être affecté, entre autres, par la présence d'informations non pertinentes et par les communiqués de presse émis directement par les entreprises.

La première contribution de cet essai est méthodologique. Nous proposons d'utiliser l'activité sur les réseaux sociaux autour de la publication de nouvelles économiques comme nouvel indicateur de l'attention des investisseurs. Notre méthodologie se base sur une analyse de similarité entre le contenu publié par les médias traditionnels et le contenu généré par une liste d'experts sur le réseau social

Twitter. Cette approche, transparente et automatique, permet d'identifier en temps réel les nouvelles pertinentes attirant l'attention des experts et de déterminer le moment précis où une information est effectivement disponible pour les acteurs de marchés.

La seconde contribution est empirique. En analysant la manière dont les marchés réagissent aux publications des nouvelles économiques, nous démontrons que lorsqu'une information attire l'attention des investisseurs, les mouvements de marchés sont caractérisés par une forte hausse des volumes échangés, une hausse de la volatilité et des sauts de prix. Au contraire, lorsque les investisseurs ne portent pas attention à une nouvelle, la publication de celle-ci a un impact très limité sur les marchés. Nous rapportons de plus que les mouvements identifiés dans la littérature avant la publication ne sont pas nécessairement le fait de la présence d'information privée, mais peuvent aussi être expliqués par une mauvaise identification de la minute exacte à laquelle une information est publique. En utilisant la première mention d'une nouvelle sur Twitter pour corriger un potentiel problème d'horodatage dans les données médias, nous montrons qu'il n'y a pas de fuite significative d'information et que les mouvements de marchés sont liés à l'information publique.

Les résultats présentés dans cet essai nous semblent être pertinents pour plusieurs pans de la littérature. Premièrement, la littérature s'intéressant à l'impact de l'arrivée de nouvelles informations non planifiées sur les dynamiques de marchés (Ranaldo, 2008; Groß-Klußmann & Hautsch, 2011; Boudt & Petitjean, 2014) pourrait trouver intéressante la méthodologie présentée permettant d'identifier à quel instant une information est effectivement rendue publique.[5] Une identification plus précise permet ensuite de mieux comprendre le fonctionnement des marchés financiers en ce qui concerne le rôle de l'information privée, complétant ainsi les résultats de Bradley et al. (2014) sur l'importance d'un horodatage exact dans les études d'évènements à l'échelle intra-journalière. Deuxièmement, cet essai contribue à la littérature sur l'attention des investisseurs, en proposant un nouveau cadre permettant d'identifier en temps réel les informations importantes, sans être dépendant de "boîtes noires" et sans choisir subjectivement les nouvelles qui semblent pertinentes (Antweiler

[5]Bien que nous ayons focalisé notre étude sur les nouvelles spécifiques aux entreprises cotées, la méthodologie de mesure de l'attention proposée peut être appliquée de la même manière à l'analyse de nouvelles macroéconomiques (annonce surprise de politique monétaire...) ou aux "breaking news" (catastrophe naturelle, changement politique...).

& Frank, 2006; Boudoukh et al., 2013). Cette méthode permet d'éviter d'avoir des inférences statistiques biaisées à cause de la présence d'articles non pertinents ou de communiqués de presse directement envoyés par les entreprises. Pour finir, nous pensons que les résultats présentés dans cet essai peuvent être utiles aux participants de marché et aux traders intra-journalier. Avant de passer un ordre, les études d'évènements réalisées montrent que les investisseurs doivent prendre en compte les biais potentiels dans l'horodatage des nouvelles économiques et utiliser l'attention des investisseurs afin de séparer les nouvelles pertinentes du bruit. Sauf en passant un ordre à l'exacte seconde de publication d'une nouvelle, les résultats de notre étude montre qu'il est impossible de réaliser un profit économique ajusté du niveau de risque en suivant une stratégie de trading basée sur la publication d'information publique, en accord avec l'hypothèse d'efficience des marchés. Nous encourageons à ce propos les chercheurs en finance à combiner des données en provenance des réseaux sociaux avec des données médias afin de confirmer ce résultat sur d'autres classes d'actifs ou d'autres périodes de temps.

3 Troisième essai

L'efficience des marchés suppose que les investisseurs soient capables d'identifier en temps réel les informations pertinentes, mais aussi d'en vérifier instantanément la véracité. Cependant, en réalité, s'assurer de l'exactitude d'une information prend du temps (Foucault et al., 2016). Les investisseurs se retrouvent alors face à un arbitrage entre investir rapidement, au risque de se baser sur un faux signal, ou bien investir après avoir analysé plus en détail l'information, au risque que celle-ci soit déjà intégrée dans les prix. A cet égard, la possibilité que les marchés soient manipulés est donc une question importante pour l'efficience informationnelle des marchés (Aggarwal & Wu, 2006).

Dans le troisième essai, intitulé "*Market manipulation and suspicious stock recommendations on social media*", nous examinons le lien entre l'information publiée sur les réseaux sociaux et l'évolution du prix des actions des entreprises de petite capitalisation, en prêtant une attention particulière à la véracité et à la pertinence de l'information. En effet, bien que les réseaux sociaux puissent aider les investisseurs souhaitant recueillir et partager des informations à propos des marchés financiers, ces nouveaux modes de communication représentent aussi une opportunité pour les fraudeurs

souhaitant envoyer des rumeurs et des fausses informations aux investisseurs.

La première contribution de cet essai est empirique. En examinant l'ensemble des actions judiciaires conduites par la SEC entre 1996 et 2015, nous démontrons tout d'abord que les manipulations de marché ciblent principalement les actions des entreprises à faible capitalisation boursière. Les fraudeurs utilisent divers canaux de communication pour envoyer de fausses informations sur les marchés, dont les communiqués de presse, les sites web, les forums de discussions et les réseaux sociaux. Dans un second temps, nous identifions, grâce à une étude d'évènement, qu'un pic d'activité à propos d'une entreprise sur Twitter est associé à une hausse du prix de l'action de l'entreprise concernée le jour même, suivie par une nette baisse la semaine suivante. Ce résultat est cohérent avec une de manipulation de type "pump-and-dump", au cours de laquelle un manipulateur utilise les réseaux sociaux afin de gonfler artificiellement le prix des actions des petites entreprises (pump), avant que le prix ne chute fortement une fois la manipulation terminée (dump).

La seconde contribution de cet essai est méthodologique. Nous proposons ainsi une nouvelle approche permettant d'identifier les comportements suspicieux afin de faire la distinction entre un retournement de prix dû à un excès d'optimisme "naturel" de la part des investisseurs et un retournement de prix suite à une hausse artificielle due à des recommandations d'achat suspicieuses. Cette approche est issue de la théorie des réseaux et inspirée des travaux de Diesner et al. (2005) sur l'identification de comportements anormaux lors du scandale Enron. En analysant les interactions entre les utilisateurs sur Twitter, nous avons identifié plusieurs groupes d'utilisateurs ayant des comportements suspicieux (utilisation de faux comptes, envoi de messages synchronisés par des robots, campagne de promotion non-déclarée...), favorisant ainsi l'hypothèse de manipulation/promotion. A notre connaissance, cet essai est le premier à présenter des preuves empiriques de manipulation où des utilisateurs se servent des réseaux sociaux afin d'envoyer des informations fausses ou trompeuses aux acteurs de marché.

Les résultats présentés dans cet essai nous semblent être d'un intérêt tout particulier pour les régulateurs de marchés, venant faire écho à deux actions civiles menées en 2011 et en 2015 par la SEC contre des individus (ou groupes d'individus) ayant utilisé le réseau social Twitter pour manipuler les marchés. Les preuves empiriques de comportements suspicieux présentées dans cet essai devraient, selon nous, inciter les régulateurs à renforcer leurs efforts pour mieux contrôler l'information diffusée

sur les réseaux sociaux. La méthodologie basée sur la théorie des réseaux que nous proposons pourrait aider les régulateurs à identifier de manière automatique les comportements anormaux et les groupes d'utilisateurs ayant des interactions suspicieuses. D'un point de vue académique, les chercheurs travaillant sur les manipulations de marché devraient quant à eux être intéressés par l'analyse fournie de l'ensemble des actions civiles menées par la SEC, permettant d'étendre les résultats d'Aggarwal & Wu (2006) et apportant de nouvelles informations concernant les outils utilisés par les fraudeurs dans le cadre de manipulation de marché basée sur la divulgation d'informations fausses ou trompeuses. Bien que des preuves de manipulation aient déjà été reportées par Böhme & Holz (2006), Nelson et al. (2013) et Sabherwal et al. (2011) à partir de base de données de spam (email) et de messages publiés sur les forums de discussion, les conclusions de cet essai montrent l'importance de considérer l'activité sur les réseaux sociaux pour mieux comprendre les manipulations de marché. L'analyse des interactions entre les utilisateurs, la disponibilité d'un horodatage précis, et le volume d'activité très important sur les réseaux sociaux font de Twitter une source de données pertinentes dans le cadre d'études empiriques à propos des manipulations de marché. Enfin, les investisseurs individuels peuvent voir en les conclusions de cet essai un rappel des risques inhérents aux investissements dans des entreprises à faible capitalisation échangées sur les marchés de gré-à-gré. A cet égard, nous plaidons donc à la fois pour un plus fort contrôle des informations publiées sur les réseaux sociaux, mais aussi pour une meilleure éducation des investisseurs individuels afin de limiter la portée des manipulations informationnelles.

General introduction

Despite several decades of empirical and theoretical research, there is still no consensus in academia on the predictability of financial markets. According to the efficient market hypothesis (Fama, 1965), the returns from speculative assets are unforecastable. Information is instantaneously and perfectly integrated into prices, and irrational investors are met in the markets by rational arbitrageurs who trade against them, immediately eliminating any anomalies. Therefore, it should be impossible to make risk-adjusted economic profits by trading (Jensen, 1978). To borrow an old joke widely told among economists about the efficient market hypothesis, if a $100 bill is lying on the floor, it is not even necessary to bend down to pick it up, because, if it was genuine, someone would have already taken it (Lo, 2008).

However, academic finance has evolved a long way from the days when the efficient market hypothesis was widely considered to be proven beyond any reasonable doubt (Shiller, 2003). By applying the principles of psychology and sociology to financial markets, behavioral finance academics have reported a number of irregularities and predictable patterns that contradict the efficient market hypothesis. They have noted that investors are not perfectly rational investing robots; rather, they are agents with limited cognitive abilities who may be subject to periods of excessive optimism or pessimism (Baker & Wurgler, 2007; Li & Yu, 2012). On the one hand, due to the limits to arbitrage, prices can diverge significantly from their fundamental values in the presence of sentiment-driven irrational noise traders (De Long et al., 1990). On the other hand, as attention is a scarce resource (Kahneman, 1973), the level of investors' attention to news may results in stock price underreaction or overreaction (Huberman & Regev, 2001; Barber & Odean, 2008).

Intellectually, a strong departure from the efficient market hypothesis would imply that an "unlimited wealth-machine" would exist for traders exploiting anomalies, which is impossible in a stable economy (Timmermann & Granger, 2004). Although it is unlikely that irrationalities will last long,

pricing anomalies can appear over time and persist for short periods (Malkiel, 2003). Moreover, a small departure from the efficient market theory is necessary to solve the fundamental conflict between the speed at which market professionals spread information and the incentive to acquire costly information (Grossman & Stiglitz, 1980). Although a $100 bill lying on the floor of a stock exchange might not be there for long, a few sophisticated investors processing information quickly and using innovative trading strategies might still be able to pick it up before it disappears.

Without making the claim of offering a definitive answer to a debate that has persisted for forty years between partisans of the efficient market hypothesis and behavioral finance academics, this dissertation aims to improve our understanding of the price formation process in financial markets through the use of Big Data analytics. Indeed, the massive increase in the availability of digital footprints generated by individuals on the Internet has made it possible to address the predictability of financial markets from a different perspective. Analogous to the development of the microscope for microbiologists (King, 2011), academics and practitioners in the social sciences now have access to several billions of granular data generated every day by individuals on social media sites, blogs, and message boards. Human behaviors, such as sentiment and attention, can now be analyzed at the micro-level. According to Einav & Levin (2014), the expansion of data on social and economic activity that is being collected is likely to have profound effects on empirical economic research. In three essays, we demonstrate that the data revolution will also profoundly affect financial research.

This dissertation makes methodological and empirical contributions to three issues related to the informational efficiency of financial markets through the use of Big Data analytics. More precisely, it analyzes: (1) how to measure intraday investor sentiment and determine the relation between investor sentiment and aggregate market returns, (2) how to measure investor attention to news in real time, and identify the relation between investor attention and the price dynamics of large capitalization stocks, and (3) how to detect suspicious behaviors that could undermine the informational role of financial markets and determine the relation between the level of posting activity on social media and small-capitalization stock returns.

In that regard, the research design of each essay involves the construction of new datasets of messages published on social media sites to create novel indicators in order to: (1) measure investor

sentiment, (2) proxy investor attention to news, and (3) detect suspicious stock recommendations that could be related to market manipulation. Using textual analysis, network theories, event studies, or predictive regressions, this dissertation provides empirical evidence that textual content published on social media contains value-relevant information about asset price formation.

In addition to its academic interest for researchers working on the informational efficiency of financial markets, on high-frequency trading dynamics, and on market manipulation, the results are also of interest for practitioners and regulators. Indeed, practitioners might be interested in empirical results from strategies using online investor attention and sentiment as trading signals. The media report an increasing usage of social media by market participants, for example to gauge overall market sentiment from messages posted on social media[6] or to implement automated trading strategies by analyzing Donald Trump's tweets.[7] However, apart from anecdotal evidence, little is known about the real profitability of such trading strategies.[8]

Last but not least, regulators might be interested in analyzing the information disseminated on social media more thoroughly to detect suspicious online behaviors and to avoid (or pursue) potential fraudsters. While social media could provide benefits to investors and improve market transparency, it also presents opportunities for fraudsters who may attempt to manipulate share prices by using social media to spread false or misleading information about stocks. Recently, the Security and Exchange Commission has conducted two civil actions against individuals who use Twitter to manipulate stock markets. Nevertheless, a better assessment of the prevalence of frauds on social media and a better understanding of the techniques used by fraudsters to manipulate stock markets is necessary to help regulators fight fraud.

3 First essay

According to Baker & Wurgler (2007), the question is no longer, as it was a few decades ago, whether investor sentiment affects stock prices, but rather how to measure investor sentiment and

[6] "Firms Analyze Tweets to Gauge Stock Sentiment" (The Wall Street Journal, July 6, 2015).

[7] "A Little Birdie Told Me: Playing the Market on Trump Tweets" (The New York Times, February 8, 2017).

[8] According to Nardo et al. (2016), "in spite of the disappointing results, mood extraction seems to also be good business according to the number of new firms selling it."

quantify its effects. Since the seminal papers from Tumarkin & Whitelaw (2001) and Antweiler & Frank (2004), researchers in behavioral finance have focused on the construction of investor sentiment proxies using data from the Internet. Extracting and analyzing millions of messages published on the Web to measure investor sentiment might sound appealing, as it could overcome issues related to answering bias (survey-based indicators), idiosyncratic non-sentiment-related components (market-based indicators), or confounding causality (media-based indicators). However, while encouraging results have been identified for small capitalization stocks (Sabherwal et al., 2011; Leung & Ton, 2015), until now, the empirical results have been disappointing for large capitalization stocks and aggregate market returns (Nardo et al., 2016).

The first essay, entitled "*Intraday online investor sentiment and return patterns in the U.S. stock market*", re-examines the relation between the sentiment of individual messages published on social media and stock returns. To provide new insights on the topic, the scope of this essay focuses on intraday market dynamics by using a novel dataset of several million messages published on the social media platform StockTwits. The use of this new data set offers many advantages. First, StockTwits allows online investors to explicitly reveal their sentiment when publishing a message (bullish/positive or bearish/negative), avoiding the bias and small sample issues that can arise when the sentiment polarity of messages is defined manually. Second, when investors register on the StockTwits platform, they can declare their experience level, trading approach, and holding period. Thus, analyzing the self-reported trading characteristics of users and investor base heterogeneity can be beneficial in exploring the different hypotheses (market efficiency, sentiment-driven noise trading, and informed trading) and linking empirical results with theoretical models.

The first contribution of this essay is related to the methodology used to derive investor sentiment indicators from textual content published on social media sites. The research provides empirical evidence showing that the measure of investor sentiment is highly dependent on the approach selected to convert each document into a quantitative sentiment variable. Standard dictionary-based approaches used in the literature are, in fact, an inappropriate way to analyze short informal messages published on social media. Thus, this essay proposes a novel approach designed specifically to analyze messages published by individuals on social media to improve classification accuracy and the quality of

the investor sentiment indicator derived from textual content. This method is based on the development of a field-specific weighted lexicon of positive and negative words used by investors when they share opinions and ideas about financial markets on social media. To favor the replicability of the results and encourage future research in this area, all the codes developed in the Python programming language and all the lexicons used in this essay are shared with the public.[9]

The second contribution of this essay concerns the relation between investor sentiment and aggregate stock market returns. Dividing each trading day into 13 half-hour time intervals, this essay demonstrates that the first half-hour shift in investor sentiment predicts that last-half hour returns. Even after controlling for previous market return and macro-economic announcements, investor sentiment remains the only predictor of market return at the intraday level. Then, examining users' self-reported investment approach, holding period and experience level, this essay demonstrates that the intraday sentiment effect is driven by the shift in sentiment of novice traders. To the best of our knowledge, this essay is the first to provide direct empirical evidence of sentiment-driven noise trading at the intraday level, consistent with behavioral finance theories.

This essay is related to several strands of literature. First, the methodology proposed to automatically construct a weighted lexicon from a large set of pre-classified messages published on StockTwits is part of the growing literature on textual analysis applied to finance (Loughran & McDonald, 2011; Kearney & Liu, 2014; Das, 2014). In this regard, this essay emphasizes the need to develop specific methods based on the type of data studied, while also stressing that a transparent and replicable methodology can provide results similar to those of more complex "black box" methodologies (machine learning). Secondly, the conclusions about the predictability of the last half-hour of exchange presented in this essay seem to be relevant to the empirical literature on intraday price dynamics (Heston et al., 2010; Sun et al., 2016). As such, it conjectures that the "intraday momentum" effect, recently documented by Gao et al. (2017), is, in fact, a "sentiment" effect due to the presence of unsophisticated investors. In this respect, a more detailed analysis of this dynamic is an interesting topic for future research. Finally, this essay is a continuation of the literature that uses new data from the Internet and social media to predict the evolution of financial markets (see, Nardo et al., 2016, for a review of the literature). Contrary to the inconclusive studies measuring investor sentiment from the

[9] http://www.thomas-renault.com

content published on the Yahoo! Finance and Raging Bull message boards (Antweiler & Frank, 2004; Leung & Ton, 2015), the results presented in this essay show that social media sites make it possible to create reliable intraday indicators of investor sentiment. Given the very high volume of messages published daily on social media (several hundred million messages) and the availability of detailed information about users, academic researchers should pay attention to these new types of data. We are currently working on an extension of these results by considering the effect that sentiment has on other asset classes and on specific macroeconomic announcements.

3 Second essay

According to the efficient market hypothesis, all relevant information is fully and instantaneously reflected in asset prices. From a practical point of view, this implies that investors (or at least a few investors) are able to identify relevant news in the real-time flow of the several tens of thousands of news stories that are published every day by traditional newswires. However, the scarcity of investor cognitive resource seems difficult to reconcile with the immediacy at which information should be integrated. In this regard, recent research on behavioral finance argues that the level of investor attention to news may impact the speed at which information is integrated into asset prices, and it might create short-term price overreaction or underreaction.

In the second essay, entitled "*Market reaction to news and investor attention in real time*", I and my co-authors Deniz Erdemlioglu and Roland Gillet, examine the high-frequency impact of unscheduled news on financial markets by proposing a new framework to measure investor attention. In the literature, indirect proxies have been used to measure investor attention, such as the traded volume, a 52-week high or an extreme return, or the level of media coverage. However, these indicators have different biases that can make the analysis of causality between the level of attention and the evolution of financial markets complex. On the one hand, market data (such as traded volume or extreme returns) contain an idiosyncratic component that is not tied to the attention of investors. On the other hand, simply considering the level of media coverage does not make it possible to consider the importance of specific news items, and it may be affected, inter alia, by press releases issued directly by companies or the presence of irrelevant news information.

The first contribution of this essay is methodological. It proposes to use social media activity around the release of unscheduled news announcements as a novel indicator of investor attention. This methodology relies on the textual similarities between news stories released on traditional newswires and the content that financial experts post on the social media platform Twitter. This transparent and automated approach enables the real-time identification of relevant news that attracts the attention of experts and allows for the identification of the precise moment at which information was available for market participants.

The second contribution of this essay is empirical. Analyzing market reaction to unscheduled news, it demonstrates, at the firm level, that news that garners high attention leads to large and persistent changes in trading activity, volatility, and price jumps. However, when investors pay little attention to the news, the effect on those trading patterns tends to be lower, and it dissipates rather quickly. This essay further reports that the pre-announcement effect might not be explained only by private information; rather, the effect could be related to the poor identification of the exact minute news is made public. Using the first mention of the news on Twitter to control for biases, it was found that, if corrected newswire timestamps are considered, the pre-announcement effect is partially eliminated. Thus, combining Twitter data with traditional newswires sources improves the understanding of how and when information affects financial markets.

The results presented in this essay contribute to several strands of literature. First, the literature on the impact that unscheduled news flows may have on trading patterns (Ranaldo, 2008; Groß-Klußmann & Hautsch, 2011; Boudt & Petitjean, 2014) may find the methodology in this thesis to be useful, as it automatically locates the exact timestamp at which a news was made public.[10] In turn, this allows researchers to better understand how financial markets work with respect to the role of private information, market efficiency, and trade timing, thereby extending Bradley et al. (2014) findings on the utmost importance of having precise timestamps when conducting an intraday event study. Second, this essay contributes to the literature on investor attention by proposing a framework that allows one to trace news events that matter in a continuous flow of information without relying on "black box" pre-processed data or subjectively selecting seemingly relevant news (Antweiler &

[10] Although this essay focuses on firm-specific unscheduled news, the proposed framework can also be relevant to research on the price impact of unscheduled non-firm-specific events (natural disasters, political events, unscheduled monetary policy announcements).

Frank, 2006; Boudoukh et al., 2013). This type of filtering reduces or eliminated the bias of statistical inferences that arise due to routine news stories or irrelevant articles. Finally, the results presented in this essay will be useful for market participants and intraday traders. When timing their trades, investors should consider biases in both news releases and investor attention. With the exception of trading almost instantaneously on news releases, this thesis found empirical evidence that market participants cannot make any risk-adjusted short-term economic profits (after transaction costs) when trading on public information. This finding is consistent with the efficient market hypothesis, and it tends to discourage news-based trading strategies. Nonetheless, further research in this area is recommended in order to confirm this novel result.

3 Third essay

Market efficiency assumes that investors can identify and process relevant news in real time, inter alia, by filtering out fake information instantaneously. However, discovering whether a signal is true or false takes time (Foucault et al., 2016), and investors face a trade-off between trading fast on a signal, at the risk of trading on false news, or trading after processing the signal, at the risk that the prices already reflect that information. In this regard, the possibility that the stock market can be manipulated is an important issue for market efficiency (Aggarwal & Wu, 2006).

The third essay, entitled "*Market manipulation and suspicious stock recommendations on social media*", provides new insights into the empirical literature on market manipulation by examining a novel dataset of messages published on the social media platform Twitter. In fact, while social media can help investors gather and share information about stock markets, it also presents opportunities for fraudsters to spread false or misleading statements in the marketplace. In order to answer this question, this essay examines all the civil actions carried out by the Security and Exchange Commission (SEC), and then it presents an analysis of a unique dataset of several million posts published on Twitter about small-capitalization stocks.

The first contribution of this essay is empirical. After analyzing all SEC litigation releases during a 14-year period, it reports that stock market manipulation mainly targets small capitalization stocks that trade over-the-counter (OTC). Fraudsters use various channels of communication to send false

or misleading statements to the marketplace, including press releases, websites, message boards, and social media sites. Then, analyzing a dataset of several hundred thousand messages posted on Twitter, this essay demonstrates that a spike in posting activity about small capitalization stocks is associated with a contemporaneous price increase, followed by a sharp price reversal over the next trading week. The findings are consistent with the patterns of a pump-and-dump scheme, where fraudsters use social media to temporarily inflate the price of small capitalization stocks.

The second contribution of this essay is methodological. It proposes a novel approach to disentangle a price reversal due to overly "natural" optimistic sentiment from the effects of suspicious stock recommendation. The approach is derived from a methodology used by Diesner et al. (2005) in the Enron fraud scandal to identify suspicious behaviors from email communications. Analyzing interactions between users, this essay identify several clusters of users with suspicious online activity (stock promoters, fake accounts, automatic postings), favoring the manipulation/promotion hypothesis. To the best of our knowledge, this essay is the first to present empirical evidence of market manipulation where (potential) fraudsters use social media to send false or misleading statement to the marketplace.

The results presented in this essay are of a particular interest to market regulators, echoing two civil actions initiated by the SEC, in 2011 and 2015, against individuals who use Twitter to manipulate stock markets. The empirical evidence presented in this essay should reinforce regulators' efforts to better control social media, as it identifies numerous suspicious stock recommendations and suspicious behaviors. Furthermore, the methodology proposed in this essay can help regulators automatically identify suspicious clusters before taking judicial action, or deciding not to do so. Academic researchers working on market manipulation might be interested in the analysis of all the SEC litigation releases, in which the findings reported by Aggarwal & Wu (2006) are extended by providing new information on the tools used by fraudsters for information-based manipulation. Although evidence of market manipulation has been reported by others using data from stock spam e-mails and message boards (Böhme & Holz, 2006; Nelson et al., 2013; Sabherwal et al., 2011), the essay shows that empirical researchers should more thoroughly consider messages posted on social media as focusing on this source of data has many advantages (precise timestamp, measurable interactions,

higher volume of messages). Finally, the conclusions presented in this essay remind individual investors of the risks inherent in investing in small-cap companies traded on OTC markets. Given the risk of manipulation, low liquidity, and the negative average abnormal return, investors should be very cautious before choosing to invest in this type of asset.

Chapter 1

Intraday online investor sentiment and return patterns in the U.S. stock market

Forthcoming, Journal of Banking and Finance

Abstract

We implement a novel approach to derive investor sentiment from messages posted on social media before we explore the relation between online investor sentiment and intraday stock returns. Using an extensive dataset of messages posted on the microblogging platform StockTwits, we construct a lexicon of words used by online investors when they share opinions and ideas about the bullishness or the bearishness of the stock market. We demonstrate that a transparent and replicable approach significantly outperforms standard dictionary-based methods used in the literature while remaining competitive with more complex machine learning algorithms. Aggregating individual message sentiment at half-hour intervals, we provide empirical evidence that online investor sentiment helps forecast intraday stock index returns. After controlling for past market returns, we find that the first half-hour change in investor sentiment predicts the last half-hour S&P 500 index ETF return. Examining users' self-reported investment approach, holding period and experience level, we find that the intraday sentiment effect is driven by the shift in the sentiment of novice traders. Overall, our results provide direct empirical evidence of sentiment-driven noise trading at the intraday level.

Keywords: Asset pricing, Investor sentiment, Intraday return predictability, Textual analysis, Machine learning, Social media

JEL classification: G02, G12, G14.

"*There's one more thing, Ben, and this is important. We're counting cards, we're not gambling. We're following a specific set of rules and playing a system. [...] Now, I've seen how crazy it can get at those tables, and sometimes, people lose control. They give in to their emotions. You will not.*"

21, Dir. Robert Luketic. Columbia Pictures, 2008. Movie.

1.1 Introduction

Since the pioneering work of Antweiler & Frank (2004) and Das & Chen (2007) on the predictability of stock markets using data from Internet message boards, a growing number of researchers have tried to "explore" the Web to provide forecasts for the financial markets (see, Nardo et al., 2016, for a survey of the literature). From both theoretical and empirical perspectives, two main elements can explain why messages posted by investors on the Internet could give rise to periods of departure from the efficient market hypothesis.[1]

First, given the tremendous increase in the flow of textual content published every day on the Internet, we may wonder whether value-relevant information about fundamental stock prices could be identified and exploited by traders able to process information and trade quickly. This situation would be consistent with the Grossman & Stiglitz (1980) framework of market efficiency, in which small excess returns simply represent the compensation for investors who spend time and money to continuously monitor a wide variety of information sources. Developing and maintaining infrastructures and algorithms to analyze billions of messages posted on the Internet every day has a cost, and an albeit low level of predictability can be viewed as a financial reward that helps to solve the fundamental conflict between the efficiency with which markets spread information and the incentives for acquiring information. Nonetheless, this value-relevant information should be short-lived, as fast-moving traders will compete to take advantage of any existing anomalies. Testing this hypothesis empirically would thus require combining intraday stock market data with high-granularity time-stamped textual data.

Second, studies in behavioral finance argue that stock prices may deviate temporarily from their fundamental values in the presence of sentiment-driven noise traders with erroneous stochastic beliefs (De Long et al., 1990) and limits to arbitrage (Pontiff, 1996; Shleifer & Vishny, 1997). According to Baker & Wurgler (2007), the question is no longer whether investor sentiment affects stock prices, but how to measure investor sentiment and quantify its effects. Various proxies have been used in the literature, and a significant degree of stock return predictability has been identified using investor

[1] In the sense of Jensen (1978), "a market is efficient with respect to information set θ_t if it is impossible to make economic profits by trading on the basis of information set θ_t".

sentiment proxies from surveys (Brown & Cliff, 2005), market data (Baker & Wurgler, 2006) or traditional media content (Tetlock, 2007). Recently, researchers in behavioral finance have also paid special attention to the construction of investor sentiment proxies using data from the Internet. Extracting and analyzing millions of messages published on the Web to measure investor sentiment may, at first sight, sound appealing, as it could overcome issues related to answering bias (survey-based indices), idiosyncratic non-sentiment-related components (market-based measures) or confounding causality (media-based variables). However, while encouraging results have been identified for small capitalization stocks (Sabherwal et al., 2011; Leung & Ton, 2015), until now, the empirical results for large stocks and market indices have been disappointing (Nardo et al., 2016). Computing investor sentiment using machine learning algorithms on data from Yahoo! Finance message boards, Antweiler & Frank (2004) and Das & Chen (2007) find no economically significant relation between user-generated content and stock returns. These results were confirmed recently by Kim & Kim (2014) on an extensive dataset of 32 million of messages and for a longer sample period: investor sentiment proxied by user-generated content is positively affected by previous stock performances but does not help predict future stock returns, volume or volatility.

However, today communication on social media is very different from chatter on message boards several years ago. Numerous articles report increasing use of social media by market participants, from large quantitative hedge funds to family offices and high-frequency-trading firms.[2] Little anecdotal evidence, like the integration of Twitter and StockTwits feeds into financial platforms (Bloomberg Terminal and Thomson Reuters Eikon), seems to confirm this phenomenon. Given the evolution of the regulatory framework[3] and the constantly changing nature of communication on the Internet, we believe that the "news or noise" question raised by Antweiler & Frank (2004) must be reassessed frequently. Thus, we add to the recent and expanding literature that examines new data from the Internet to forecast stock markets (see, among others, Da et al., 2015; Moat et al., 2013; Avery et al., 2016; Chen et al., 2014; Blankespoor et al., 2013; Sprenger et al., 2014) by focusing on user-generated content published on the social media platform StockTwits.

[2]See, for example, "The Wall Street Journal - Firms Analyze Tweets to Gauge Stock Sentiment"

[3]See "Commission Guidance on the Use of Company We Sites" and "SEC Says Social Media OK for Company Announcements if Investors Are Alerted"

This paper contributes both to the literature on intraday return predictability and to the literature on textual analysis in finance. Analyzing ETF price dynamics, Gao et al. (2017) (GHLZ hereafter) provide empirical evidence showing that the first half-hour return predicts positively the last half-hour return. Theoretically, the market intraday momentum is consistent with an infrequent rebalancing mechanism (Bogousslavsky, 2016) and with the presence of late-informed traders in the market. Extending GHLZ model by exploring the relationship between intraday stock market returns and intraday sentiment, Sun et al. (2016) (SNS hereafter) find that the change in investor sentiment has predictive value for the intraday market returns. The signs of the estimated coefficients for the change in investor sentiment are positive on all regressions: sentiment-driven optimistic (pessimistic) traders create short-term upward (downward) price pressure, especially during the end of the trading day. One potential explanation proposed by both GHLZ and SNS is related to limits to arbitrage, as risk averse market makers might hesitate to trade against over-optimistic (over-pessimistic) noise traders during the last half-hour of the trading days to avoid exposures to overnight risks, resulting in a short-term pricing anomaly.

Regarding textual analysis in finance, one of the many challenges faced by academics and practitioners in this field concerns the methodology used to automatically convert a qualitative variable—a message, a blog post, or a tweet—into a quantitative sentiment variable. Two main methods are used for textual sentiment analysis in finance: dictionary-based approaches and machine learning techniques (see, Kearney & Liu, 2014; Das, 2014, for surveys of methods and models). Whereas dictionary-based methods that use the Harvard-IV dictionary or the Loughran & McDonald (2011) dictionary (LM hereafter) are widely used in the literature to measure sentiment in articles published in traditional media (Tetlock, 2007; Tetlock et al., 2008; Engelberg et al., 2012; Dougal et al., 2012; Garcia, 2013), textual sentiment analysis of user-generated content published on the Internet mainly relies on machine learning algorithms (Antweiler & Frank, 2004; Das & Chen, 2007; Sprenger et al., 2014; Leung & Ton, 2015; Ranco et al., 2015). Although each method has its own advantages and limits, as we will discuss later, one simple reason that explains the predominance of machine learning techniques to quantify individual messages posted on message boards and social media is the absence of a field-specific dictionary. Messages published by online investors on the Internet are usually

shorter and less formal than content published on traditional media, making the correct classification of tone difficult (Loughran & McDonald, 2016). Nonetheless, as stated by Nardo et al. (2016), "a good text classifier for a financial corpus is a good avenue for future research," as it could facilitate the comparability and enhance the replicability of previous findings.

In this paper, we first implement a novel approach to construct a lexicon of words used by investors when they share ideas and opinions about the bullishness or bearishness of the stock market on social media. Following Oliveira et al. (2016), we use a subset of 750,000 messages already tagged by online investors as bullish (positive) or bearish (negative) to automatically construct a field-specific weighted lexicon (L_1 hereafter). We also develop a field-specific non-weighted lexicon (L_2 hereafter) by examining and classifying manually all words that appear at least 75 times in the sample, adopting a methodology close to Loughran & McDonald (2011). Then, we use L_1 and L_2 to derive sentiment in a subset of 250,000 tagged messages, and we compare the out-of-sample classification accuracy with three baseline methods: a dictionary-based approach using the LM dictionary (B_1 hereafter), a dictionary-based approach using the Harvard-IV dictionary (B_2 hereafter) and a supervised machine learning algorithm using a maximum entropy classifier (M_1 hereafter). We find that L_1, L_2 and M_1 significantly outperform the standard dictionary-based approaches B_1 and B_2. Thus, the results confirm Kearney & Liu (2014) conclusion about the need to construct more authoritative and extensive field-specific dictionaries in order to enhance replicability and facilitate future work in the area.

Then, we examine the relation between online investor sentiment and intraday stock returns using an extensive dataset of nearly 60 million messages published by online investors over a five-year period, from January 2012 to December 2016. We compute five distinct intraday investor sentiment measures by aggregating the sentiment of individual messages posted on the microblogging platform StockTwits at half-hour intervals. We follow Heston et al. (2010) by dividing each trading day into 13 half-hour trading intervals, and we reassess the intraday momentum and the intraday sentiment effect documented by Gao et al. (2017) and Sun et al. (2016). We find that when investor sentiment is computed using L_1, L_2 and M_1, the first half-hour change in investor sentiment predicts positively the last half-hour S&P 500 index ETF returns. After controlling for the lagged market return and the first half-hour return, we find that first half-hour change in investor sentiment remains the only significant

predictor of the last half-hour market return. In contrast, the predictability disappears when sentiment is computed using B_1 or B_2.

Analyzing users' self-reported information on their investment approach (technical, fundamental, momentum, value, growth or global macro), holding period (day trader, swing trader, position trader or long-term investor) and experience level (novice, intermediate or professional), we construct intraday investor sentiment indicators for each group of users. We find that the intraday sentiment effect is mainly driven by the shift in the sentiment of novice traders. Implementing a trading strategy using the change in novice traders' sentiment as a trading signal to buy (sell) the S&P 500 ETF during the last half-hour of the trading day before selling (buying) it at market close, we demonstrate that a sentiment-driven strategy delivers a significantly higher risk-adjusted performance compared to baseline strategies (momentum, long-only, first half-hour and random strategies).

This paper supports the role of investor sentiment in predicting intraday stock returns and adds to the existing literature for various reasons. First, our results contrast with previous findings from GHLZ by demonstrating that the intraday price momentum has disappeared during the most recent sample period. While SNS find that both the sentiment effect and lagged return variables help predict the last half-hour return on a sample period from 1998 to 2016, we find that the sentiment effect is the only predictor of the last half-hour return on a sample period from 2012 to 2016. Second, we demonstrate that the intraday sentiment-driven anomaly is very short-lived: a positive sentiment-driven price pressure on day t is followed by a price reversal on the next trading day, consistent with the noise trading hypothesis. Third, and contrary to the measure of investor sentiment based on the proprietary Thomson Reuters MarketPsych Indices (TRMI) used by SNS, our investor sentiment measure is transparent, replicable, and allows us to provide a more direct test of the noise trading hypothesis. Exploring investor base heterogeneity and focusing on users' experience level (novice, intermediate, professional), we provide to the best of our knowledge the first direct empirical evidence of intraday sentiment-driven noise trading.

The paper is structured as follows. Section 1.2 describes the StockTwits platform and gives details about the data. Section 1.3 reviews the differences between dictionary-based methods and machine-learning techniques and compares the classification accuracy of L_1 and L_2 with other baseline methods

used in the literature. Section 1.4 explores the relation between online investor sentiment and intraday stock returns. Section 1.5 concludes and discusses further research.

1.2 Data

StockTwits is a social microblogging platform dedicated to financial markets on which individuals, investors, market professionals and public companies can publish 140-character messages to “Tap into the Pulse of the Markets". According to StockTwits.com, more than 300,000 users now use the platform to share information and ideas, producing streams that are viewed by an audience of more than 40 million across the financial web and social media platforms. In September 2012, StockTwits implemented a new feature that allows users to express their sentiment directly when they publish a message on the platform. More precisely, every time a user chooses to post a message on StockTwits, he or she can classify his or her message as “bearish" (negative) or “bullish" (positive) by simply clicking on a toggle button below his or her message. Figure 1.1 shows a screenshot from the StockTwits platform, with a bearish message, an unclassified message and a bullish message.

[Insert Figure 1.1 about here]

Using the Python library *BeautifulSoup*, we extract all messages published on StockTwits between January 1, 2012, and December 31, 2016, and we store them in a MongoDB NoSQL database. For each message, we collect the following information: (1) a unique identifier, (2) the username of the user who sent the message, (3) the message content, (4) the time stamp with a one-second granularity and (5) the sentiment (“bullish", “bearish" and “unclassified") associated with the message. Table 1.1 shows a sample of messages from the database, with the sentiment variable associated. Our final dataset contains 59,598,856 messages from 239,996 distinct users. Overall, 9,434,321 messages are classified as bullish (15.85%) and 2,286,292 as bearish (3.84%), and the remaining are unclassified. The 4 to 1 ratio between positive and negative messages shows that online investors are, on average, optimistic about the stock markets, as already documented in the literature (see, e.g., Kim & Kim, 2014; Avery et al., 2016).

Table 1.2 presents descriptive statistics of StockTwits messages during the sample period. Figure 1.2 represents the volume of messages per 30-minute intervals during a representative week, illustrating the intraday and weekly seasonality of message posted on the social media platform. Intraday activity on StockTwits usually peaks at market opening (between 9:30 a.m. and 10:00 a.m.), decreases at lunchtime and increases again before market close (between 3:30 p.m. and 4:00 p.m.). During non-trading hours and weekends, the average number of messages per 30-minutes interval is approximately 10 times lower than during trading hours (over the whole sample period).

[Insert Tables 1.1 and 1.2 about here]

[Insert Figure 1.2 about here]

1.3 Textual sentiment analysis

Before assessing whether user-generated content can help predict stock returns, academics and practitioners have to implement specific procedures to convert unstructured qualitative information into structured quantitative sentiment variables. In this section, we briefly review the two distinct approaches used for textual sentiment analysis, before we detail the methodology we implement to construct field-specific lexicons and compare our results with the benchmark classifiers used in the literature.

1.3.1 Dictionary-based classification

In the simplest form, a dictionary-based “bag-of-words" approach consists of computing a sentiment variable by counting the number of positive words and the number of negative words in a document, using a predefined list of signed words. For example, in a simple 4-word lexicon where “good" and “love" are defined as positive and “bad" and “hate" are defined as negative, the sentence “I love Facebook $FB company" is classified as positive with a score of +1.

Three main procedures can be implemented to create lexicons for sentiment analysis. The first technique relies on pure experts’ views, in which researchers create from scratch a list of positive and negative words, based on their knowledge and expertise. The second technique, used, for example,

to construct the LM dictionary, is a two-step process in which a vector of words is automatically generated by analyzing a list of non-classified documents. Then, each word is manually classified as positive, negative or neutral by an expert.[4] The last technique consists of creating or extracting a list of pre-classified documents and, for each word, computing statistical measures based on the term's frequency (and/or document frequency) in each class of documents. Term frequency thresholds are then used to classify each word as positive, neutral or negative.

Although a dictionary-based approach is easy to implement, and if the list of signed words is public, enables replicability, this approach has some limitations. First, it is necessary to develop field-specific dictionaries for each domain of research, as a word may not have the same meaning in two different contexts. For example, words like "liability", "capital" and "cost" are classified as negative in the Harvard-IV psychosocial dictionary but should be considered otherwise in finance (Loughran & McDonald, 2011). Furthermore, even in a given area like financial markets, formal articles written by financial journalists on traditional media are very different from user-generated content published by individual investors on the Internet. According to Loughran & McDonald (2016), the use of slang, sarcasm, emoticons and the constantly changing vocabulary on social media makes accurate classification of tone difficult. Second, except for rare exceptions (Jegadeesh & Wu, 2013), the vast majority of dictionary-based approaches uses an equal-weighting scheme, where each word in the dictionary is supposed to have the same explanatory power. Although term-weighting has the potential to increase the accuracy of textual analysis, the large number of available weighting procedures may give too many degrees of freedom to researchers in selecting the best possible empirical specification (Loughran & McDonald, 2016), creating a risk of overfitting.

1.3.2 Machine learning classification

The objective of a machine learning classification is to provide a prediction of Y given a set of features X. For a 2-class sentiment analysis problem, Y represents sentiment classes $Y_1 = positive$ and

[4]For example, Loughran & McDonald (2011) extract all words occurring in at least 5% of 121,217 10-K reports downloaded directly from the Security and Exchange Commission website, before manually classifying the "eligible words" as positive, negative or neutral.

Y_2 = *negative* and X is a vector of words. A supervised learning classification problem can be decomposed in three steps: (1) learn in-sample, (2) measure accuracy out-of-sample and (3) predict. First, a training dataset of n documents d pre-classified as positive or negative is used to fit the algorithm (see, Pang et al., 2002 for a description and a mathematical explanation of three of the most widely used classifiers in the literature: naive Bayes, support vector machine and maximum entropy). Then, features identified during the learning phase are used to predict the Y class on a testing dataset of n' pre-classified documents d'. Classification accuracy is computed by comparing the classifier prediction to the known value of Y for all documents in d'. When the accuracy of the prediction cannot be improved by modifying or fine-tuning the parameters and/or is in line with previous findings in the literature, then the algorithm is used to predict the outcome Y for all documents where class Y is unknown.

A machine learning technique has many advantages compared to a dictionary-based approach. Instead of relying on a (somehow subjective and limited) list of signed words, it allows the automatic construction of a very large set of features specific to the domain of interest and to the type of data. Furthermore, machine learning algorithms can provide answers to problems related to the weighting procedure or the non-independence of words in a sentence. However, this does not come without limitations. The first difficulty is to create or extract a sufficiently large list of labeled documents to construct a training dataset and a testing dataset. In most cases, documents are labeled manually by the author(s) or by financial expert(s) so there is subjectivity.[5] Second, machine learning accuracy can be very sensitive to the size and the construction of the training dataset. For example, Antweiler & Frank (2004) manually labeled only 1,000 messages from Yahoo! Finance message boards (55 negative, 693 neutral and 252 positive) to train their classifier, raising concerns about the accuracy of the classification when the algorithm is fitted on such a low number of messages. Third, supervised classification accuracy can change significantly depending on the algorithm used (naive Bayes, support vector machine, maximum entropy, random forests, neural network...) and few fine-tuning arbitrary parameters. As most papers use a (private) manually labeled training dataset and a specific set

[5] A system in which each message is classified by two different reviewers can be implemented to partly overcome this issue. However, as shown by Das & Chen (2007) on a sample of 438 messages posted on Yahoo! Finance message boards, the level of agreement between two human experts can be very low, with a mismatch percentage of 27.5% in their sample.

of (often) unpublished rules, filters or parameters to fit the data, replicability and comparison across studies are often impossible.

1.3.3 Creating an investor lexicon

To create our lexicon, we follow Oliveira et al. (2016) automated procedure by focusing on messages in which sentiment is explicitly revealed by online investors. We first randomly select a list of 375,000 “bullish" messages and 375,000 “bearish" messages published on StockTwits between June 2013 and August 2014. As in Pang et al. (2002), we impose a maximum of 375 messages per user and per class (or 0.1% of the whole corpus) to avoid domination of the corpus by a small number of prolific reviewers. We implement a data cleaning process similar to Sprenger et al. (2014), except that we choose to keep the punctuation (question marks and exclamation marks) and we do not remove the morphological endings from words. To take negation into account, we add the prefix “negtag_" to all words following “not", “no", “none", “neither", “never" or “nobody".

Although various natural language processing approaches could have been applied (lemmatization, stemming, part-of-speech tagging), we choose to use a conservative approach by removing only three stopwords from all messages (“a", “an" and “the").[6] We also convert positive emoticons into a common word “emojipos" and negative emoticons into a common word “emojineg"[7], as in Go et al. (2009). We replace all tickers ($SPY, $AAPL, $BOA, $XOM...) with a common word “cashtag", all links by a common word “linktag", all numbers by a common word “numbertag" and all mentions of users by a common word “usertag". Table 1.3 shows several examples of messages before and after data pre-processing.

[Insert Table 1.3 about here]

We use a bag-of-words approach to extract all unigrams (one word) and bigrams (two words) appearing at least 75 times in the sample of 750,000 messages. While the Harvard-IV and the LM dictionary consider only unigrams, we find that adding bigrams provides additional information and

[6] We choose a conservative approach as we find that the words “short", “shorts", “shorted", “shorter", “shorters" and “shorties" are used by online investors to express very distinct feelings. The same is true for the words “call", “calls", “called", “calling", “caller", “callers" and for a subsequent number of words.

[7] ;) :) :-) =) :D as “emojipos". :(:-(=(as “emojineg

improves the accuracy of the classification.[8] For each of the 19,665 terms t identified (5,786 unigrams and 13,879 bigrams), we count the number of occurrences of t in the 375,000 bullish documents ($n_{d_{pos},t}$) and the number of occurrences of t in the 375,000 bearish documents ($n_{d_{neg},t}$). We define the sentiment weight (SW) for each word as:

$$SW(t) = \frac{n_{d_{pos},t} - n_{d_{neg},t}}{n_{d_{pos},t} + n_{d_{neg},t}} \tag{1.1}$$

Table 1.4 shows a list of selected n-grams with their associated sentiment weight. For example, the word “buy" was used 20,837 times in bullish messages and 12,654 times in bearish messages, leading to a SW of 0.2443. Interestingly, we find that the bigrams “buy !" and “strong buy" convey a much more positive sentiment than the unigram “buy", with an SW equal to 0.6052 and 0.8250, respectively. The bigram “buy ?" is approximately neutral (SW equals 0.0331) whereas “negtag_buy" (“not buy", “never buy"...) conveys a negative sentiment (SW equals -0.4534).

[Insert Table 1.4 about here]

Then, we sort all 19,665 n-grams by their SW, and we define a weighted field-specific lexicon L_1 by considering all terms in the first quintile (negative terms) and all terms in the last quintile (positive terms). Manually examining all words included in lexicon L_1 (approximately 8,000 n-grams), we identify a few anomalies and misclassifications. For example, the word “further" is classified as negative, as it appears 1,260 times in the 375,000 negative documents and 506 times in the 375,000 positive documents, leading to an SW of -0.4270 (in the first quintile). Analyzing the n-gram frequencies, we find that the word “further" is often used in combination with verbs like “drop," “down" and “fall" (“drop further", “down further," “fall further"), in such a way that the negativity does not come from the word “further" by itself but from the verb associated with it in the bigrams. Another anomaly is related to non-equity assets. For example, the unigram “commodity" is considered negative in L_1, because, during the sample period, commodity prices dropped, and investors were mainly

[8] For example, the sentence “What a bear trap!" should be not be classified as negative (i.e., “bear trap" is an expression used in technical analysis to indicate that a security should go up) even if “bear" and “trap" are individually considered negative.

commenting on past movements using bearish vocabulary. The same is true for the unigrams “Euro" and “EURUSD" as the euro currency depreciates sharply against the dollar during the sample period.

Thus, we adopt a methodology close to Loughran & McDonald (2011) to create a manually cleaned equal-weighted field-specific lexicon. More precisely, we examine all n-grams in L_1, and we manually classify each n-gram as positive (+1), negative (-1) or neutral (0). We also add typical inflections of root words defined as positive or negative to extend our lexicon. For example, we manually classify the words “bankrupt" and “bankruptcy" as negative, and we add the inflections “bankrupts", “bankrupted", “bankrupting" and “bankruptcies". We end up with a total of 543 positive terms and 768 negative terms, and we denote this lexicon L_2. L_1 and L_2 are available online.[9]

1.3.4 Message sentiment and classification accuracy

To assess the accuracy of L_1 and L_2, we use a time-order evaluation holdout. We randomly select a list of 125,000 bullish messages and 125,000 bearish messages published on StockTwits between September 2014 and April 2015. We use the same pre-processing techniques and the same limit of messages for a given user as for the training dataset (maximum 0.1% of the whole corpus). For each message, we compute a sentiment score by considering five classifiers:

- L_1 - Weighted field-specific lexicon: approximately 4,000 negative outlook terms and 4,000 positive outlook terms. $SW(t)$ as defined previously.
- L_2 - Manual field-specific lexicon: 768 negative outlook terms and 543 positive outlook terms. $SW(t)$ equals 1 for positive terms and -1 for negative terms.
- B_1 - Loughran-McDonald dictionary: 2,355 negative outlook terms and 354 positive outlook terms. $SW(t)$ equals 1 for positive terms and -1 for negative terms.
- B_2 - Harvard-IV psychosocial dictionary: 2,007 negative outlook terms and 1,626 positive outlook terms. $SW(t)$ equals 1 for positive terms and -1 for negative terms.
- M_1 - Supervised machine learning algorithm (maximum entropy): Implemented using scikit-learn, a machine learning package in Python. Default parameters and equal prior probabilities.

[9] http://www.thomas-renault.com

For L_1, L_2, B_1 and B_2, the individual message sentiment score is defined as the average $SW(t)$ of the terms present in the message. Given the standardized number of words in each document (maximum 140 characters), we find that using a simple relative word count weighting scheme gives slightly better results than a Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme (see Appendix A for details). This result is consistent with those of Smailović et al. (2014), who find, using data from Twitter, that the term-frequency (TF) approach is statistically significantly better than the TD-IDF based approach. For M_1, individual message sentiment score is given by the probability estimates that a message m belongs to the bullish or the bearish class. See Appendix B for a detailed description. For all messages in the testing dataset, we compare the sentiment expressed by the investor who sent the message (the real sentiment) with the sentiment score computed using the five classifiers (the estimated sentiment). We compute the percentage of correct classification excluding unclassified messages CC (i.e, bearish-declared messages with a sentiment score lower than 0 and bullish-declared messages with a sentiment score greater than 0), the percentage of correct classification per class (CC_{bull} and CC_{bear}, respectively), the percentage of classified messages CM (message with a sentiment score different from zero) and the percentage of classified messages per class (CM_{bull} and CM_{bear}). Table 1.5 presents the results.

[Insert Table 1.5 about here]

We find a percentage of correct classification of 74.62% for L_1 and 76.36% for L_2. As the number of features is much greater in L_1 (approximately 8,000 n-grams) than in L_2 (approximately 1,300 n-grams), the percentage of classified messages CM is greater for L_1 (90.03%) than for L_2 (61.78%), leading to an expected arbitrage between accuracy and exhaustiveness. Interestingly, and contrary to Oliveira et al. (2016), we find that the accuracy and the percentage of the classified messages are nearly equivalent for the bullish and bearish messages for L_1.[10] However, the percentage of correct classification of benchmark dictionary-based approaches B_1 (LM) and B_2 (Harvard-IV) is significantly lower, with an accuracy of 63.06% and 58.29%, respectively. Furthermore, the percentage of

[10] As we focus our analysis on financial messages published on social media with self-reported sentiment, we cannot compare directly the accuracy of our field-specific approach with previous results from the literature on textual analysis. However, out-of-sample classification accuracy between 75% and 80% is standard on user-generated content sentiment analysis (see Pang et al. (2002), Go et al. (2009) or Smailović et al. (2014), among others).

classified messages in B_1 is very low (27.70%) as numerous messages published on social media do not contain any words included in the LM word lists. The LM dictionary was created by examining formal corporate 10-K reports in such a way that it is not well suited to analyze informal messages published on social media. This first result confirms Kearney & Liu (2014) discussion on the need to construct more authoritative and extensive field-specific dictionaries in order to improve textual analysis classification.

We also find that the classification accuracy of the supervised machine learning method M_1 is slightly better (75.16%) than that of L_1 (74.62%). However, as we will show later, results for the relation between investor sentiment and stock returns are qualitatively similar when intraday investor sentiment indicators are computed using L_1, L_2 or M_1. As field-specific dictionary-based approaches are more transparent than machine learning techniques, we believe that researchers should consider thoroughly implementing both methods when quantifying textual content published on the Internet. This dual approach would enhance the replicability and comparability of the findings while ensuring that the results are robust to the methodology used to convert a text into a quantitative sentiment variable. Thus, we re-affirm Loughran & McDonald (2016) conclusion by recommending that alternative complex methods (machine learning) should be considered only when they add substantive value beyond simpler and more transparent approaches (bag-of words).

1.4 Intraday online investor sentiment and stock returns

In this section, we explore the relation between online investor sentiment and intraday stock returns. We first detail the methodology we use to derive the investor sentiment indicators by aggregating the sentiment of individual messages. Then, we reassess the intraday momentum patterns documented by GHLZ by considering an augmented sentiment-based model. Last, we analyze whether users' self-reported investment approach, holding period and experience level contain value-relevant information to understand the reason behind the intraday sentiment effect.

1.4.1 Intraday investor sentiment indicators

We use our five classifiers to derive a sentiment score between -1 and +1 for all 59,598,856 messages published on StockTwits between January 1, 2012, and December 31, 2016. Then, we compute five intraday investor sentiment indicators by averaging, at half-hour intervals, the sentiment score of individual messages published per 30-minute period. We denote those indicators s_x where x={L_1, L_2, B_1, B_2, M_1}. To control for the increase in message volume and the seasonality of posting patterns on social media, we standardize s_x by dividing each indicator by its rolling one-week standard deviation. Table 1.6 shows the correlation between the five s_x indicators.

[Insert Table 1.6 about here]

The very high correlation coefficient between s_{L1} and s_{M1} (0.9341) seems to confirm that quantifying the sentiment of individual messages using a weighted field-specific lexicon is competitive with more complex machine learning methods. However, the correlation coefficients of s_{B1} and s_{B2} with our field-specific approach are low (from 0.2292 to 0.3365) demonstrating that the methodology used to derive quantitative indicators from textual content can widely affect investor sentiment measures.

1.4.2 Predictive regressions

Following Heston et al. (2010), we divide each trading day into 13 half-hour intervals. We denote $r_{i,t}$ the *i-th* half-hour return of the S&P 500 ETF on day *t*. As in GHLZ, $r_{1,t}$ is the first half-hour return using the closing price on day *t-1* and the price at 10:00 a.m. on day *t*. $r_{13,t}$ denotes the last half-hour return using the ETF price at 3:30 p.m. and 4:00 p.m. on day *t*. In a similar fashion, we denote $\Delta s_{i,t}$ the change in intraday investor sentiment in the *i-th* half-hour trading interval on day *t*. For example, $\Delta s_{1,t}$ denotes the difference between the first half-hour investor sentiment (the average sentiment of all messages sent between 9:30 a.m. and 10:00 p.m.) on day *t* and the last half-hour sentiment on day *t-1* (the average sentiment of all messages sent between 3:30 p.m. and 4:00 p.m. on the previous trading day). $\Delta s_{13,t}$ denotes the difference between the last half-hour investor sentiment and the 12th half-hour investor sentiment on day t.

As in SNS, we run predictive regressions to explore the relation between changes in intraday investor sentiment and the half-hour S&P 500 index ETF return. Given GHLZ empirical evidence showing that the first half-hour return predicts the last half-hour return, we also include the first half-hour change in investor sentiment. Thus, we consider the following model:

$$r_{i,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 \Delta s_{i,t-1} + \epsilon_t \tag{1.2}$$

where i represents the i-th half-hour time interval. Table 1.7 shows the regression results for i={11,12,13}.[11] We present the results when investor sentiment is computed using the five classifiers (L_1, L_2, B_1, B_2 and M_1). The regressions are based on 1,258 observations (251 or 252 trading days per year from 2012 to 2016).

[Insert Table 1.7 about here]

We find evidence that when investor sentiment is computed using L_1, L_2 or M_1, the first half-hour change in investor sentiment predicts the last half-hour stock market return. Coefficients are significant and positive at the 0.1% level when investor sentiment is computed with L_1 or M_1 and at the 1% level when investor sentiment is computed with L_2. The R^2 values of 1.35% (L_1) and 1.33% (M_1) are comparable to those reported by SNS on the predictability of the last half-hour return using the change in investor sentiment based on the Thomson Reuters MarketPsych Indices (1.43%). However, when investor sentiment is computed using B_1 or B_2, we do not find any predictability. This finding reinforces our conclusion that the Loughran-McDonald and the Harvard-IV psychosocial dictionaries are inappropriate for deriving the sentiment of short informal messages published on social media.

We then control for lagged market returns to assess if the predictability of stock index return using past change in investor sentiment is not caused by a contemporaneous correlation between sentiment and return (as documented, among others, by Kim & Kim, 2014). Based on the results in Table 1.7, we focus on $i = 13$ and on the first half-hour change in investor sentiment. More precisely, we

[11] As we do not find significant results for i={2,...,10}, we do not present those results for readability.

consider the following model:

$$r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \beta_4 r_{13,t-1} + \epsilon_t. \tag{1.3}$$

The inclusion of $r_{1,t}$ is motivated by GHLZ who find that the first half-hour return predicts the last half-hour return for a wide range of ETFs. The inclusion of $r_{13,t-1}$ is motivated by Heston et al. (2010) who identify return continuation at half-hour intervals that are exact multiples of a trading day. Table 1.8 presents the results.

[Insert Table 1.8 about here]

Even after controlling for lagged market returns, the first half-hour change in investor sentiment remains the only significant predictor of the last half-hour market return. Sentiment-driven optimistic (pessimistic) traders create short-term upward (downward) price pressure at the end of the trading day. This finding provides evidence that the intraday sentiment effect is distinct from the intraday momentum effect. Interestingly, we also demonstrate that the intraday momentum effect documented by GHLZ do not hold during the most recent period. Although we find evidence of intraday momentum effect when we consider a longer time period from 1998 to 2017, with R^2 values and coefficients very similar to those reported by GHLZ on a time period from 1993 to 2013, we do not find significant intraday momentum effect when we focus on recent years (2012 to 2017). Academic research may have destroyed stock return predictability (McLean & Pontiff, 2016), or previous results may have been caused by data-snooping, market frictions or omitted variables. We leave this question for further research.

We also examine whether the intraday sentiment effect is driven by the release of macroeconomics news before the market opens or during the trading day. For this purpose, we re-run Equation 1.3 by dividing all trading days into two groups: days with news releases and days without. We focus on three major macroeconomics announcements: Non-Farm Payroll (NFP, monthly at 8.30 a.m.), the Michigan Consumer Sentiment Index (MSCI, preliminary and final releases, monthly at 10:00 a.m.) and the Federal Open Market Committee meeting (FOMC, every six weeks at 2:00 p.m.). To account for FOMC pre-meeting or post-meeting announcement drift, we include one day before and one day

after the meetings. Table 1.9 reports the results. For readability, we present the results only when field-specific lexicon L_1 is used to derive investor sentiment, but we find similar results for L_2 and M_1, and no significant results for B_1 and B_2, as previously.

[Insert Table 1.9 about here]

We find that the intraday sentiment effect is concentrated on days without macroeconomic news announcements. The first half-hour shift in investor sentiment is not significant on NFP days, MSCI days, and [-1:+1] days around FOMC meetings. Investor sentiment, thus, is not a mere reflection of macroeconomics news announcements. This result is consistent with the fact that on days with macroeconomic news announcements, the last half-hour return is mainly driven by the news announcements in such a way that sentiment-driven traders do not affect prices. However, on days with no news, investor sentiment affects stock prices.

As in GHLZ and SNS, we then analyze whether the sentiment effect is significant for other domestic ETFs, sector indices, international ETFs and bond ETFs. Table 1.10 reports the results. As above, we report only the results when we use L_1 to measure investor sentiment, but the results are similar for L_2 and M_1. We confirm that the first half-hour change in investor sentiment predicts the last half-hour return for a diverse set of ETFs. We also find that the associated R^2 decreases for international equity indices and small capitalization ETFs (Russell 2000) and is not significant for bond market ETFs. This result is consistent with the fact that users on StockTwits mainly discuss the development of the U.S. stock market indices and the cross-section of large and medium capitalization stock returns. These complementary results provide evidence that analyzing data from StockTwits allows researchers to construct a value-relevant intraday measure of U.S. investor sentiment.

[Insert Table 1.10 about here]

Last but not least, we investigate whether the predictability identified previously is driven by fundamental end of day demand or by noise trading. To do so, we consider the following model:

$$r_{i,t+1} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{13,t} + \epsilon_t. \tag{1.4}$$

If the predictability is driven by noise trading, we should identify a price reversal over the next trading day (i.e, a negative coefficient on β_1). This situation would be consistent with the presence of uninformed sentiment-driven traders causing a price run up on day t (as shown previously) followed by a price reversal afterwards on day $t + 1$ when arbitrageurs step in to correct the anomaly. Table 1.11 shows the regression results for i={1,2,...,13}.[12] We identify a significant price reversal on day $t + 1$ during the 8th, the 11th and the 12th half-hour of the trading days, favoring the noise trading hypothesis over the fundamental end of day demand hypothesis. This result is consistent with the evidence of sentiment-driven short-term price pressures followed by price reversals documented in the literature (Tetlock, 2007; Garcia, 2013). However, to the best of our knowledge, we provide the first evidence of sentiment-driven price pressure followed by a price reversal at the intraday level, consistent with limits to arbitrage during the last half-hour of the trading day.[13]

[Insert Table 1.11 about here]

1.4.3 Exploring investor base heterogeneity

Contrary to the Thomson Reuters MarketPsych Index (TRMI) used by SNS as a proxy for intraday investor sentiment (a “black box" aggregate indicator), focusing on data from StockTwits allows researchers to test directly whether the predictability is driven (or not) by noise trader sentiment. StockTwits provides unique information about users’ self-reported investment approach (technical, fundamental, global macro, momentum, growth, or value), holding period (day trader, swing trader, position trader, or long-term investor), and experience level (novice, intermediate, or professional). For example, using data from StockTwits and exploiting investor base heterogeneity, Cookson & Niessner (2016) find that investor disagreement robustly forecasts abnormal trading volume at a daily frequency. In a similar fashion, we assess in this subsection whether a specific type of trader or a specific trading strategy drives the sentiment effect identified previously. Although reporting the investment approach, the holding period and the experience level is not required to register to StockTwits, we still observe a self-reported trading strategy for a large number of users (84,891 users)

[12] For readability, we only present our results when L_1 is used to compute investor sentiment.

[13] Sun et al. (2016) find that "there appears to be some evidence of reversal at longer horizons", but the coefficient estimates for β_1 are not significant in most of their regressions, with the exception of the 11th half-hour.

and messages (35,436,607 messages). Table 1.12 presents the distribution of users by the investment approach, holding period and experience level.

[Insert Table 1.12 about here]

As in the previous subsection, we construct intraday investor sentiment indicators at half-hour time intervals. However, instead of considering all messages, we create intraday investor sentiment indicators for each investment approach, each holding period and each experience level by considering only the messages of users who self-reported the given information in their profile. We find qualitatively similar results when we use L_1, L_2 or M_1 but no significant results when we use B_1 and B_2, confirming previous findings. For readability, we present the results only when field-specific lexicon L_1 is used to quantify individual message sentiment. As only 1.01% of users self-declared themselves as following a "Global Macro" trading approach, we remove this strategy as in Cookson & Niessner (2016). The correlation coefficient between the 12 investor sentiment indicators at half-hour time intervals range from 0.0780 (between "fundamental traders" and technical traders") to 0.6216 (between "technical traders" and "swing traders"). See Appendix C for details. We denote with $\Delta s_{1,t,x}$ the first half-hour change in investor sentiment on day t for users' self-reported characteristic x. Then, we estimate the following predictive regression:

$$r_{13,t} = \alpha + \beta_1 \Delta s_{1,t,x} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \epsilon_t. \tag{1.5}$$

where $r_{13,t}$ is the last half-hour return, $r_{1,t}$ is the first half-hour return, $r_{12,t}$ the 12th half-hour return and $\Delta s_{1,t,x}$ represents the change in sentiment the first half-hour of day t for each investor type $x = \{x_1, x_2, x_3\}$. We consider each investor depending on his or her trading approach ($x_1 = \{$*technical, fundamental, momentum, growth, value*$\}$), his or her holding period ($x_2 = \{$*day, swing, position, long-term*$\}$) and his or her experience ($x_3 = \{$*novice, intermediate, professional*$\}$). Table 1.13 presents the results by investment approach, holding period and experience level.

[Insert Table 1.13 about here]

Analyzing each investment approach separately, and controlling for lagged market returns, we find significant results for traders with technical, growth and value investing strategies and for position traders (i.e., holding periods from a few days to a few weeks). We also find that the significance of the results decreases with traders' self-reported experience. The first half-hour change in novice investor sentiment is significant at the 1% level (Adj-R^2 equal to 1.77%) whereas the first half-hour change in intermediate investor sentiment is significant only at the 5% level (Adj-R^2 equal to 1.51%), and the first half-hour change in professional investor sentiment is not significant (Adj-R^2 equal to 1.33%). We also consider all possible approach and experience, approach and period, and period and experience doublets (60 combinations). We find that the last last half-hour return is robustly forecasted by the first half-hour change in novice investor sentiment. Looking at 10 doublets with the highest Adj-R^2, we find that all the best combinations, except one, include the change in novice investor sentiment (with R^2 ranging from 1.69 to 2.05). The only other characteristic that adds value when combined with the "novice experience" is the trading approach "technical analysis" (significant at the 10% level).

Last, we simulate a trading strategy buying (selling) the S&P 500 ETF at 3.30 p.m. on days with an increase in novice investor sentiment during the first half-hour of that day, and selling (buying) at 4:00 p.m. We present the results when the performance of the trading strategies is evaluated using the Sharpe ratio, but the results are robust to the performance evaluation metrics as all trading strategies exhibit very similar volatility. We compare the performance of a "sentiment-driven" strategy with an *Always Long Strategy* buying the ETF at the beginning of the last half-hour and selling it at market close. We also consider a *First Half-Hour Return Strategy* buying (selling) the ETF on days with a positive (negative) first half-hour return and selling (buying) it at market close, and a *12th Half-Hour Return Strategy* buying (selling) the ETF on days with a positive (negative) 12th half-hour return and selling (buying) it at market close. As in Roger (2014), we compare the Sharpe ratio of each strategy to the simulated Sharpe ratio distribution by generating 10,000 strategies randomly buying (selling) the S&P 500 ETF. Table 1.14 reports the results.

[Insert Table 1.14]

We find that the average annualized return of a strategy using half-hour change in novice investor

sentiment as a trading signal is equal to 4.55%, with a Sharpe ratio of 1.496. Although the annualized return might not seem impressive at first sight, the return is remarkable as we hold a position only during 30 minutes per day and we do not keep any position overnight. Comparing the location of the sentiment-driven strategy Sharpe ratio in the simulated Sharpe ratio distribution, we find that only 9 random strategies out of the 10,000 simulated ones have a Sharpe ratio greater than 1.496. Thus, the observed profitability is significant at the 0.1% level. We also demonstrate that a sentiment-driven strategy significantly outperforms other benchmark strategies. Overall, the results provide empirical evidence of sentiment-driven noise trading at the intraday level.

1.4.4 Discussion of empirical results

According to GHLZ, there are two explanations for why the first half-hour return predicts the last half-hour return. First, strategic informed traders might time their trade for periods of high trading volume. On days with positive overnight night news, informed traders are likely to trade very actively at the market opening before reinforcing their position during the last half-hour. Second, on days with a sharp overnight and first half-hour increase in the stock market index, some traders might expect a price reversal over the following hours and short the market. As typical day traders are flat at the end of the day, they are likely to unwind their position during the last half-hour return which, in turn, will push prices up. Closer to our paper, SNS provide two reasons to explain why investor sentiment has predictive value for intraday market returns and why the sentiment effect is concentrated on the end of the trading day. First, due to risk aversion, investors trading the S&P 500 index ETF might prefer to wait a few hours before taking a position on the market. Second, risk-averse arbitrageurs may be more likely to trade against sentiment traders at the beginning of the day than later in the day due to the uncertainty introduced by overnight news.

Our findings provide direct empirical evidence for the two hypotheses proposed by SNS. First, we find that when investors are more optimistic during the first 30 minutes on day *t* than during the last 30 minutes of day *t-1*, the S&P 500 index ETF significantly increase during the last half-hour of the trading day. However, all other variations in investor sentiment ($\Delta s_{i,t}$ for i={2,...12}) are not significant in predictive regressions. This finding illustrates the “timing effect" as investors seem to

prefer to wait until “the dust is about to settle" before buying or selling the S&P 500 index ETF based on their initial sentiment. This finding is also consistent with the explanation based on the presence of late-informed investors provided by GHLZ.

Furthermore, analyzing users' self-reported experience, we find that the last half-hour predictability is driven by the shift in the sentiment of novice traders, and, to a lesser extent, by the shift in the sentiment of traders following technical analysis strategies. This finding is consistent with Hoffmann & Shefrin (2014) who find, using private data from a sample of discount brokerage clients, that individual investors who use technical analysis are disproportionately likely to speculate in the short-term stock market. Examining the impact of aggregate investor sentiment on trading volume and long-run price reversal, SNS document that the investor sentiment effect is driven by noise trading. In this paper, using self-reported experience level instead of making indirect inferences by analyzing market reactions, we provide, to the best of our knowledge, the first direct empirical evidence of intraday sentiment-driven noise trading.

1.5 Conclusion

Improving the transparency and replicability of results are of utmost importance for the big-data and finance environment. Although developing public field-specific lexicons will obviously not solve all issues related to replicability and comparability, it still constitutes an important step to facilitate further research in this area, as stated by Nardo et al. (2016) in a recent survey of the literature of financial market prediction using the Web. In the first part of this paper, we construct a lexicon of words used by online investors when they share opinions and ideas about the bullishness or bearishness of the stock market by using an extensive dataset of messages for which sentiment is explicitly revealed by investors. We demonstrate that a transparent and replicable approach significantly outperforms the benchmark dictionaries used in the literature while remaining competitive with more complex machine learning algorithms. The findings provide empirical evidence to Kearney & Liu (2014) conclusion about the need to develop a more authoritative field-specific lexicon and of Loughran & McDonald (2016) recommendations that alternative complex methods (machine learning) should be

considered only when they add substantive value beyond simpler and more transparent approaches (bag-of words).

In the second part, we explore the relation between online investor sentiment and intraday S&P 500 index ETF returns. We find that the first half-hour change in investor sentiment predicts the last half-hour return, even after controlling for lagged market returns (first half-hour return and lagged half-hour return). This finding holds for a wide range of ETFs and is robust to macroeconomic news announcements. We also demonstrate that the short-term sentiment-driven price pressure is followed by a price reversal on the next trading day, consistent with the noise trading hypothesis. Then, analyzing users' self-reported investment approach, holding period and experience level, we find that the sentiment effect is mainly driven by the shift in the sentiment of novice traders. We confirm this result by showing that a strategy that uses changes in novice investors' sentiment as trading signals significantly outperforms other baseline strategies (risk-ajusted performance). Overall, the results provide direct empirical evidence of intraday sentiment-driven noise trading.

Although we focused on the predictability of aggregate market returns, we believe that the evolution of intraday investor sentiment over time and across users with different trading approaches, experiences and investment horizons can also be useful in many other situations, such as explaining the cross-section of average stock returns or forecasting stock market volatility. We encourage further research in this area by making public the field-specific weighted lexicon we developed for this paper.

1.6 Appendix A - Weighting scheme

The standard TF-IDF weighting scheme, often used in information retrieval and text mining, can be computed as:

$$tf\text{-}idf(t,d) = \frac{n_{d,t}}{n_{d,T}} * log\frac{N_d}{N_{d,t}} \tag{1.6}$$

where t is a term (unigram or bigram), d is a collection of documents, $n_{d,t}$ is the number of occurrences of term t in documents d, $n_{d,T}$ is the total number of terms in documents d, N_d is the total number of documents d, $N_{d,t}$ is the total number of documents d containing term t. Then, the sentiment weight for each term t can be computed as in Oliveira et al. (2016) as:

$$SW_{tf\text{-}idf}(t) = \frac{tf\text{-}idf(t,d_{pos}) - tf\text{-}idf(t,d_{neg})}{tf\text{-}idf(t,d_{pos}) + tf\text{-}idf(t,d_{neg})}, \tag{1.7}$$

where d_{pos} is a collection of positive documents, and d_{neg} is a collection of negative documents. In the paper, we choose to adopt a very simple relative word count (wc) term-weighting, defined as:

$$SW_{wc}(t) = \frac{n_{d_{pos},t} - n_{d_{neg},t}}{n_{d_{pos},t} + n_{d_{neg},t}} \tag{1.8}$$

Given the maximum length of the messages published on social media (140 characters), $N_{d,t} \approx n_{d,T}$ (as a given word very rarely appears twice in the same tweet). Furthermore, in our empirical analysis, the number of bullish (positive) documents in the training dataset is equal to the number of bearish (negative) documents (375,000) ($n_{d_{pos},T} \approx n_{d_{neg},T}$ and $N_{d_{pos}} \approx N_{d_{neg}}$). From previous equations, it thus can be easily seen that $SW_{tf\text{-}idf}(t) \approx SW_{wc}(t)$.

Analyzing all n-grams that appear at least 75 times in our training dataset, we find an absolute difference between $SW_{tf\text{-}idf}(t)$ and $SW_{wc}(t)$ equal to 0.024. Comparing out-of-sample classification accuracy, we find qualitatively similar results when a TF-IDF scheme is used to compute the terms' weight and to identify relevant features (n-grams). Table A-1 presents the out-of-sample classification accuracy of a subset of 250,000 messages. Furthermore, the results for the predictability of intra-day returns are qualitatively similar when investor sentiment is derived using a relative word-count

Table A-1 - Classification accuracy - TD-IDF and relative word count weighting scheme

Classifier	CC	CC_{bull}	CC_{bear}	CM	CM_{bull}	CM_{bear}
L_1 (TF-IDF)	74.53%	73.82%	75.23%	89.96%	89.31%	90.61%
L_1 (Word Count)	74.62%	73.98%	75.24%	90.03%	89.32%	90.73%

Notes: This tables shows the out-of-sample classification accuracy when terms' weight are computed using a relative word count weighting scheme or a TF-IDF weighting scheme. We also present results from a simple relative word count weighting scheme (as used in the paper). We report the percentage of correct classification excluding unclassified messages CC, the percentage of correct classification per class (respectively CC_{bull} and CC_{bear}), the percentage of classified messages CM (message with a sentiment score different from zero) and the percentage of classified messages per class (CM_{bull} and CM_{bear}).

weighting scheme or a TF-IDF scheme. Table A-2 presents the results. Overall, we find that the results are robust to the method used for term-weighting. As the term-weighing scheme lacks theoretical motivation (Loughran & McDonald, 2016), we favor the simplest approach due to the standardized (and short) size of the messages posted on social media. Recently, Smailović et al. (2014) confirmed that the TF approach is statistically significantly better than the TD-IDF-based approach to data from Twitter.

Table A-2 - Predictive regressions - Investor sentiment and half-hour market return

	α	β_1	β_2	$Adj R^2$ (%)
L_1 (TF-IDF)	-0.0001 (-1.3099)	0.0316*** (3.9785)	-0.0083 (-0.6618)	1.36
L_1 (Word Count)	-0.0001 (-1.4169)	0.0312*** (4.1339)	-0.0087 (-0.6879)	1.44

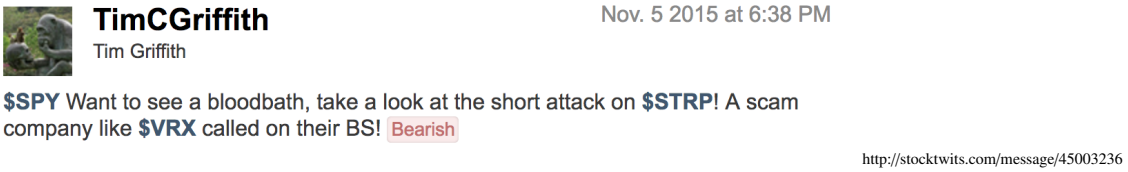
Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 \Delta s_{12,t} + \epsilon_t$ when the change in investor sentiment is computed using a relative word count weighting scheme or a TF-IDF weighting scheme. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1,258 observations).

1.7 Appendix B - Message classification

We compute a sentiment score between -1 and +1 for all messages published on StockTwits ($SS(m)$) by adopting dictionary-based approaches and a machine learning method.

Dictionary-based approaches

For dictionary-based approach L_1, we use a methodology similar to Oliveira et al. (2016). Message sentiment is equal to the average $SW(t)$ of the terms present in the message and included in lexicon L_1. When a bigram is present in the text, we do not take into account the score of the individual unigram included in the bigram to avoid double counting. For example, considering the following message:



Using the field-specific lexicon L_1, we find that the following terms are present in the message above (within the brackets the SW computed as in Equation 1.1):

- cashtag ! [SW = 0.3069]
- cashtag called [SW = -0.3033]
- bloodbath [SW = -0.6600]
- short [SW = -0.5811]
- scam [SW = -0.8493]

Taking the average $SW(t)$, we find a sentiment score equals -0.4069. In this example, the classification is correct as the message was classified as “Bearish" by the user who sent the tweet, and we obtain a sentiment score lower than 0. We use a similar methodology to compute $SS(m)$ for the other dictionary-based approaches L_2, B_1 and B_2, except that we consider an equal-weighting scheme by

giving all words in the positive lists a weight of +1 and all words in the negative lists a weight of +1. Using the previous example, we identify the following terms:

- L_2 : bloodbath [-1], short [-1], scam [-1]
- B_1 : None of the words are present in the LM dictionary
- B_2 : short [-1], attack [-1], company [+1], like [+1]

We end up with a sentiment score for the message equal to -1 for L_2, 0 for B_1 (no term identified) and 0 for B_2 (two positive terms and two negative terms).

Machine learning methods

We experiment three machine algorithms as in Pang et al. (2002) and Go et al. (2009): naive Bayes (NB), maximum entropy (MaxEnt) and support vector machines (SVM). We report results only for MaxEnt, as we find that MaxEnt provides better results than NB (we conjecture due to the overlapping in NB) and similar (but with a lower computational complexity) than SVM. For MaxEnt, the probability that document d belongs to class c given a weight vector δ is equal to:

$$P(c|d,\delta) = \frac{exp[\sum_i \delta \tilde{} i f_i(c,d)]}{\sum_c exp[\sum_i \delta \tilde{} i f_i(c,d)]} \tag{1.9}$$

where $f_i = \{f_1, f_2, .., f_m\}$ is a predefined set of m features (unigram or bigram) that can appear in a document. The weight vector is found by numerical optimization of the lambdas to maximize the conditional probability. We use the "liblinear" package for this purpose. Considering the previous message ($SPY Want to see a bloodbath, take a look at the short attack on $SRTP! A scam company like $VRX called on their BS!), we find using MaxEnt: $P(c_{pos}) = 0.12$ and $P(c_{neg}) = 0.88$. To obtain an $SS(m)$ between -1 and +1, we define:

$$SS(m)_{MaxEnt} = (P(c_{pos}|m,\delta) - 0.5) * 2. \tag{1.10}$$

In the previous example, we find $SS_{MaxEnt} = -0.76$. We then consider all messages with an $SS_{MaxEnt} < 0$ (equivalent to a $P(c_{pos}) < 0.5$) as negative, and all messages with an $SS_{MaxEnt} > 0$ as

positive. When a message does not contain any features included in $\{f_1, f_2, .., f_m\}$, then $SS_{MaxEnt} = 0$, and we consider the message as unclassified.

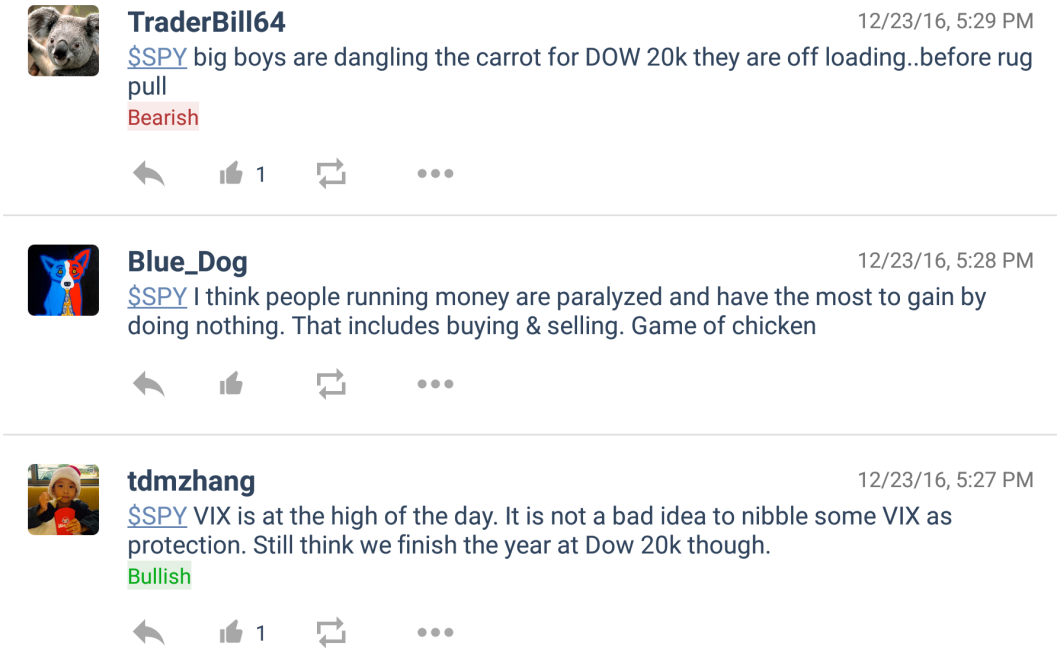
Appendix C - Trading strategy correlation

Intraday investor sentiment - Self-reported trading strategy correlation

	Technical	Fundamental	Momentum	Growth	Value	Day	Swing	Position	Long-Term	Novice	Intermediate	Professional
Technical	1.0000											
Fundamental	0.1037	1.0000										
Momentum	0.1664	0.0844	1.0000									
Growth	0.1154	0.1202	0.1170	1.0000								
Value	0.1126	0.0780	0.0792	0.0984	1.0000							
Day	0.4816	0.1103	0.2429	0.0950	0.0889	1.0000						
Swing	0.6216	0.1978	0.3520	0.2193	0.1464	0.1806	1.0000					
Position	0.3146	0.2421	0.2412	0.2295	0.2240	0.1224	0.1880	1.0000				
Long	0.1659	0.3569	0.1374	0.3829	0.4118	0.0878	0.1597	0.1585	1.0000			
Novice	0.2309	0.1867	0.2425	0.3285	0.1534	0.1753	0.3131	0.2035	0.3535	1.0000		
Intermediate	0.4778	0.2716	0.3401	0.2846	0.1905	0.3161	0.4873	0.4588	0.2837	0.1773	1.0000	
Professional	0.4778	0.2411	0.2261	0.1687	0.3019	0.3804	0.4224	0.3631	0.2986	0.1386	0.2050	1.0000

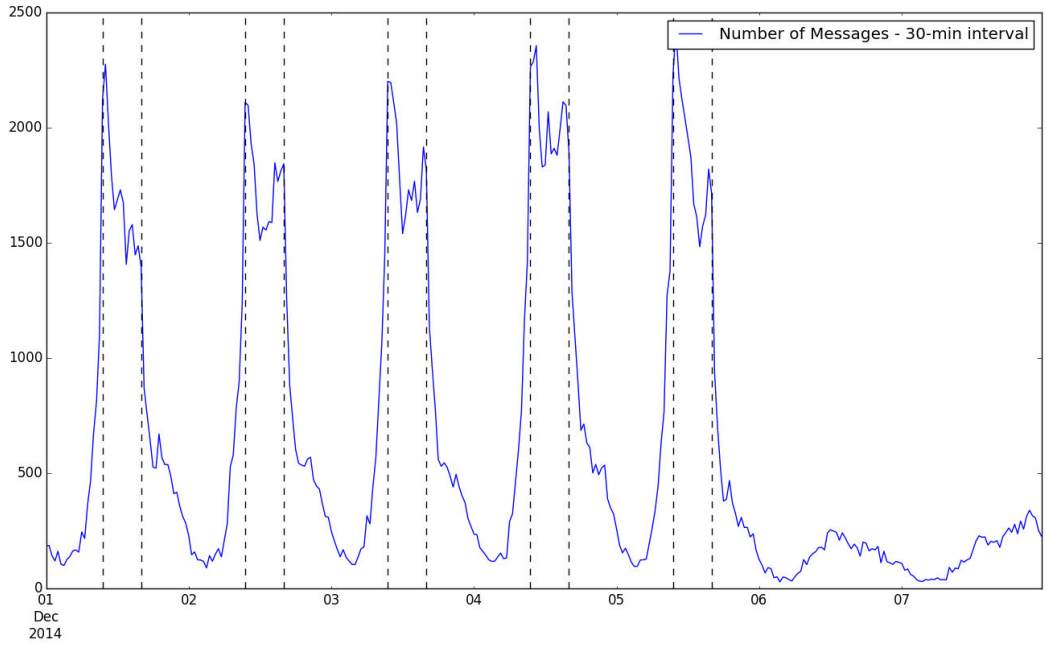
Notes: This tables shows the correlation matrix of intraday investor sentiment indicators for each investment approach, each holding period and each experience level. Results are presented when investor sentiment indicators are computed from individual message quantification using L_1.

FIGURE 1.1: StockTwits platform - Explicitly revealed sentiment



Notes: This figure shows a screenshot from StockTwits platform on December 23, 2016. The first message was self-classified as bearish (negative) by the investor who wrote the tweet (TraderBill64). The second message was not classified. The third was classified as bullish (positive) by the investor who wrote the tweet (tdmzhang). $SPY is the cashtag associated with the S&P 500 index ETF.

FIGURE 1.2: StockTwits - Number of messages per 30-minute interval



Notes: This figure shows the number of messages published on the platform StockTwits for each 30-minute interval on a representative week, from Monday, December 1, to Sunday, December 7, 2014. Dashed vertical lines represent market opening hours (9:30 a.m.) and market closing hours (4 p.m.).

TABLE 1.1: StockTwits - Sample messages

Date	Sentiment	User	Content
2014-11-11 14:42:49	Bearish	albogolfer	$BABA Dont catch the falling knife !! Fraud company !! All Lies
2014-11-11 14:42:49	N/A	Saratoga733	@jahnman @123xyz what is time frame on the 80mil or is it only on as needed basis?
2014-11-11 14:42:49	Bearish	lbailey33	$BABA IS PURE TRASH!!! JUNK PRODUCTS!!!!
2014-11-11 14:42:51	N/A	Akcoline	$BABA would it be reason if they will be able to fill the orders of $9.6 bil
2014-11-11 14:42:54	N/A	2Struong	$BABA will dive more in the last 30 min of trading today
2014-11-11 14:42:56	N/A	Poweruhvone	@neve1064 When did SIRI IPO? What did it do for monthsyear after the IPO? Your [...].
2014-11-11 14:42:57	Bullish	Dannydyn	$TSLA BURN YOU SHORTS, BUUUUUUUUUUUUUUURRRRNNNNN!!!!
2014-11-11 14:42:58	Bullish	rosswestgate	$KERX why are we stuck? other firms would be at 35 by now lol
2014-11-11 14:42:59	N/A	DZee	@Ben_Frank_Beatdown Now that MADE ME LAUGH!!!
2014-11-11 14:43:00	N/A	seer	$PLI.CA Chen will likely be sharing Lenovo partnership talk and more tomorrow
2014-11-11 14:43:00	N/A	GrindTime	@BehavioralEdge GILD already has 100% success rate in 500+ person Japan trial. Big deal [...]
2014-11-11 14:43:00	N/A	Dawnr48	@Crackerjax if holds 28.12 good r:r maybe to 29.
2014-11-11 14:43:00	N/A	adamh9999	$CODI has good earnings and the offering is at 17.50 a share. This will go up from here.
2014-11-11 14:43:01	N/A	kim2020	this will turn tomorrow!!
2014-11-11 14:43:02	Bearish	OutOfBounds1	$SPY Oh just fall already.
2014-11-11 14:43:04	N/A	ThunderRoad	@MrMulally When that happens, this could be bigger than Apple.
2014-11-11 14:43:05	N/A	SmokeyNYY	@BiopharmaPro The funny thing is Benitec is further along with their Macular Degeneration than RXII
2014-11-11 14:43:05	N/A	FinancesStocks	$AMSWA American Software, Inc. Invites You to Join Its Second Quarter Fiscal Year [...]
2014-11-11 14:43:06	N/A	iowasooner	$APP imo probably best to sell wsmall profit & possibly revisit after next ER....BOD aren't helping
2014-11-11 14:43:08	N/A	Gbones2692	@Daks Yes very confident with UGAZ to 24 min by mid jam
2014-11-11 14:43:08	Bullish	Beautifulla	@viktor89 Nice I got a big big smiling on my face hahaha
2014-11-11 14:43:15	N/A	Pisces_19	@LMikeH Thanks Mike! Hope this is just temporary as these providers sure need upgrades soon.
2014-11-11 14:43:16	N/A	dinovix2778	@jgp0006 it's up to you. Don't take any advice. Do your own math and sell when you feel good [...]
2014-11-11 14:43:19	N/A	kaluhiokalani71	$PCLN oh gawwd! this is tearing faces off shortz!
2014-11-11 14:43:19	N/A	stckwhisperer	@BounceBabyBounce Do you mean limit order?
2014-11-11 14:43:20	N/A	joelyboyblue	$FOSL Long some
2014-11-11 14:43:20	Bullish	Farne	@jp007 we go green hornet brotha
2014-11-11 14:43:21	N/A	Kroberg	@Trade4moneydaily Don't have a crystal ball, sorry
2014-11-11 14:43:21	N/A	regent17	@MoreCash2014 @IndianTrader I feel it, today bought 1500 shares, on may, i did the same thing at $8 !
2014-11-11 14:43:23	N/A	tubetoppuller	$ASTI float still around 10M?
2014-11-11 14:43:23	N/A	M10	$JNUG CPCharts-You got it wrong. Weak yen=strong dollar=lower gold. Wrong information [...]
2014-11-11 14:43:27	N/A	milkball	$TSLA Who is holding their calls overnight?

Notes: This table shows all messages sent on StockTwits between 14:42:49 and 14:43:27 on November 11, 2014. Messages can be classified as positive (bullish) or negative (bearish) by the investor/user who sent the tweet, or can be published without any associated sentiment (N/A).

TABLE 1.2: Descriptive statistics - StockTwits messages

Period	Mean	Std-Dev	Min	Max	Total
30-min (All)	886.65	679.61	0	8,248	59,598,856
30-min (Trading-hours)	1092.67	1918.98	0	8,248	31,383,060
Daily (All)	26,649.86	32,621.16	1,127	132,063	59,598,856
Daily (2012)	6,805.23	11,488.46	1,127	27,831	4,204,778
Daily (2013)	10,251.82	17,786.75	2,070	46,501	6,492,164
Daily (2014)	16,088.48	29,765.41	4,100	59,310	10,864,373
Daily (2015)	21,766.21	42,323.00	6,442	80,936	15,447,896
Daily (2016)	32,435.49	61,720.34	9,153	132,063	22,589,645

Notes: This table shows descriptive statistics about the quantity of messages posted on the platform StockTwits. We present statistics at half-hour time interval and at a daily frequency for each year in our sample (2012-2016).

TABLE 1.3: StockTwits messages - Data pre-processing

Message before pre-processing	@lololemon $BABA IS PURE TRASH !!
Message after pre-processing	usertag cashtag is pure trash ! !
Message before pre-processing	$FB dropping now! not good :(
Message after pre-processing	cashtag dropping now ! negtag_good emojineg
Message before pre-processing	$MSFT Short the POP
Message after pre-processing	cashtag short pop
Message before pre-processing	$GILD moves like Jagger! http://stks.co/r0nUR
Message after pre-processing	cashtag moves like jagger ! linktag

Notes: This table shows four examples of messages before and after data pre-processing (removing stopwords, adding prefix for negation, replacing users' mention by "usertag", tickers by "cashtag", links by "linktag"...).

TABLE 1.4: Selected sample of n-grams and associated Sentiment Weight (SW)

n-grams	n_{total}	n_{pos}	n_{neg}	SW
awesome	1,447	1,077	370	0.4886
bear	5,669	1,506	4,163	-0.4687
bear trap	393	250	143	0.2723
beast mode	182	172	10	0.8901
bottomed-out	137	127	10	0.8540
bullish	11,483	7,812	3,671	0.3606
bullish engulfing	121	112	9	0.8512
buy	33,491	20,837	12,654	0.2443
buy !	765	614	151	0.6052
buy ?	302	156	146	0.0331
cashtag junk	95	1	94	-0.9789
down	4,2391	11,388	31,003	-0.4627
down further	145	25	120	-0.6552
emojineg	1,885	401	1,484	-0.5745
emojipos	15,223	10,091	5,132	0.3258
great	11,952	8,380	3,572	0.4023
great fundamentals	126	120	6	0.9048
intraday	1,334	557	777	-0.1649
investor	1,493	869	624	0.1641
like	35,756	17,845	17,911	-0.0018
media	1,038	557	481	0.0732
negtag_buy	1,577	431	1,146	-0.4534
negtag_short	781	290	491	-0.2574
optimism	185	91	94	-0.0162
poor	1,467	333	1,134	-0.5460
poor fundamental	136	0	136	-1.0000
price	20,730	10,393	10,337	0.0027
pump	4,501	659	3,842	-0.7072
scam	1,540	116	1,424	-0.8494
sell	23,183	6,637	16,546	-0.4274
sentiment	1,982	619	1,363	-0.3754
short	47,856	10,022	37,834	-0.5812
stock	32,781	13,928	18,853	-0.1502
strong	8,223	5,966	2,257	0.4511
strong buy	557	507	50	0.8205
timber	398	17	381	-0.9146
today	38,761	21,604	17,157	0.1147
trading	8,383	3,934	4,449	-0.0614
trap	1,867	426	1,441	-0.5437
up	61,337	37,823	23,514	0.2333
up up	786	720	66	0.8321
word	817	473	344	0.1579

Notes: This table shows the Sentiment Weight (SW) of a sample of selected words. For example, over the 750,000 messages we use to construct our lexicon, the word “buy" appears 33,491 times in the positive training dataset (375,000 messages) and 20,837 times in the negative training dataset (375,000 messages), leading to a sentiment weight SW of (33,491 - 20,837) / (33,491 + 20,837) = 0.2443. Red and green colors represent n-grams with a SW respectively in the first and last quintile (when sorting all 19,665 n-grams by their SW).

TABLE 1.5: Classification accuracy - Investor social lexicons

Classifier	CC	CC_{bull}	CC_{bear}	CM	CM_{bull}	CM_{bear}
L_1	74.62%	73.98%	**75.24%**	**90.03%**	**89.32%**	**90.73%**
L_2	**76.36%**	**79.10%**	73.72%	61.78%	60.61%	62.95%
B_1	63.06%	57.99%	67.86%	27.70%	26.88%	28.50%
B_2	58.29%	63.63%	53.02%	58.09%	57.72%	58.47%
M_1	75.16%	75.98%	74.36%	**90.03%**	**89.32%**	**90.73%**

Notes: This tables shows the out-of-sample classification accuracy for classifiers L_1, L_2, B_1, B_2 and M_1, computed on 250,000 messages from the testing dataset (125,000 positive and 125,000 negative). We report the percentage of correct classification excluding unclassified messages CC, the percentage of correct classification per class (respectively CC_{bull} and CC_{bear}), the percentage of classified messages CM (message with a sentiment score different from zero) and the percentage of classified messages per class (CM_{bull} and CM_{bear}).

TABLE 1.6: Intraday investor sentiment indicators - Correlation matrix

	s_{L1}	s_{L2}	s_{B1}	s_{B2}	s_{M1}
s_{L1}	1.0000				
s_{L2}	0.6250	1.0000			
s_{B1}	0.2292	0.3365	1.0000		
s_{B2}	0.2328	0.3000	0.3112	1.0000	
s_{M1}	0.9341	0.6581	0.2629	0.2361	1.0000

Notes: This tables shows the correlation matrix of our five intraday investor sentiment indicators s_x, where x=$\{L_1, L_2, B_1, B_2, M_1\}$.

TABLE 1.7: Predictive regressions - Investor sentiment and half-hour market return

	α	β_1	β_2	Adj-R^2 (%)
11th half-hour return				
L_1	0.0000 (0.1671)	0.0031 (0.4809)	0.0005 (0.0568)	-0.14
L_2	0.0000 (0.2262)	0.0057 (0.8112)	0.0080 (0.9700)	-0.01
B_1	0.0000 (0.4161)	0.0081 (0.8771)	0.0038 (0.3940)	-0.08
B_2	0.0000 (0.3183)	-0.0082 (-0.7383)	-0.0140 (-1.5655)	0.06
M_1	0.0000 (0.1493)	0.0047 (0.7144)	-0.0001 (-0.0093)	-0.11
12th half-hour return				
L_1	0.0001 (1.1835)	-0.0093 (-1.3883)	0.0050 (0.5527)	0.06
L_2	0.0000 (1.0038)	-0.0027 (-0.3930)	0.0036 (0.4338)	-0.13
B_1	0.0000 (0.8201)	-0.0096 (-0.8781)	-0.0010 (-0.1119)	-0.08
B_2	0.0001 (1.2040)	-0.0117 (-0.9928)	0.0031 (0.2922)	-0.04
M_1	0.0001 (1.0658)	-0.0055 (-0.7922)	0.0061 (0.7040)	-0.05
Last half-hour return				
L_1	-0.0001 (-0.9945)	0.0274*** (4.1448)	-0.0181 (-1.5949)	1.35
L_2	-0.0000 (-0.2838)	0.0227** (3.1837)	-0.0086 (-0.8755)	0.71
B_1	-0.0000 (-0.2310)	0.0075 (0.6176)	-0.0097 (-0.9079)	-0.07
B_2	-0.0000 (-0.6261)	0.0071 (0.6144)	-0.0099 (-0.7517)	-0.08
M_1	-0.0001 (-0.9649)	0.0273*** (3.9754)	-0.0194 (-1.7576)	1.33

Notes: This table reports the results of the equation $r_{i,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 \Delta s_{i,t-1} + \epsilon_t$ for i={11,12,13}. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1,258 observations).

TABLE 1.8: Predictive regressions - Investor sentiment and lagged market return

	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
Last half-hour return						
L_1	-0.0001 (-1.1662)	0.0274*** (3.4025)	0.0111 (0.5610)	0.1086 (1.2903)	0.0508 (1.1349)	2.13
L_2	-0.0000 (-0.4378)	0.0216** (2.6833)	0.0142 (0.7337)	0.1047 (1.2400)	0.0523 (1.1456)	1.68
B_1	-0.0000 (-0.5873)	0.0052 (0.4468)	0.0248 (1.4088)	0.1051 (1.2392)	0.0392 (0.8589)	1.10
B_2	-0.0000 (-0.7841)	0.0074 (0.6651)	0.0251 (1.4145)	0.1054 (1.2448)	0.0391 (0.8590)	1.12
M_1	-0.0001 (-1.0671)	0.0269** (3.2612)	0.0108 (0.5456)	0.1062 (1.2626)	0.0518 (1.1533)	2.04
GHLZ [2012-2016]	-0.0000 (-0.7003)		0.0255 (1.4390)	0.1039 (1.2263)		1.10
GHLZ [1998-2016]	-0.0000 (-0.7903)		0.0673*** (4.3443)	0.1246** (2.7420)		2.91

Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \beta_4 r_{13,t-1} + \epsilon_t$. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1,258 observations). This table also reports the results of the equation $r_{13,t} = \alpha + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \epsilon_t$ as in GHLZ.

TABLE 1.9: Predictive regressions - News and no-news trading days

	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)	Obs.
NFP							
Release	0.0000 (0.0185)	-0.0386 (-1.1573)	-0.0057 (-0.1732)	0.1353 (0.6669)	0.2164 (1.3349)	0.53	58
No Release	-0.0001 (-1.4401)	0.0310*** (3.6609)	0.0115 (0.5373)	0.1074 (1.2339)	0.0481 (1.0551)	2.39	1,200
MSCI							
Release	0.0001 (0.7152)	0.0046 (0.1700)	0.0426 (1.6016)	-0.0840 (-0.5957)	0.2955** (3.1112)	8.88	116
No Release	-0.0001 (-1.3211)	0.0282*** (3.3813)	0.0087 (0.4071)	0.1173 (1.3396)	0.0229 (0.4919)	2.13	1,142
FOMC Meetings							
Release	-0.0001 (-0.6180)	0.0193 (1.0068)	0.0823* (2.3597)	0.0168 (0.1069)	-0.1118 (-1.1740)	4.50	120
No Release	-0.0001 (-1.1176)	0.0302*** (3.4959)	0.0028 (0.1286)	0.1162 (1.2819)	0.0702 (1.4009)	2.33	1,138
NFP or MSCI or FOMC							
Release	0.0001 (0.5122)	0.0127 (0.8540)	0.0234 (0.9985)	0.0019 (0.0157)	0.1092 (1.4222)	0.98	238
No Release	-0.0001 (-1.5408)	0.0334*** (3.5410)	0.0028 (0.1107)	0.1260 (1.2988)	0.0355 (0.6672)	2.53	993

Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \beta_4 r_{13,t-1} + \epsilon_t$ for days with (release) or without (no release) macroeconomic news announcements. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016.

TABLE 1.10: Predictive regression - Other ETFs.

US ETF	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
SPY [S&P 500]	-0.0001 (-1.1662)	0.0274*** (3.4025)	0.0111 (0.5610)	0.1086 (1.2903)	0.0508 (1.1349)	2.13
DIA [Dow]	-0.0001* (-1.8996)	0.0260*** (3.3277)	-0.0005 (-0.0290)	0.1303 (1.4043)	0.0441 (0.9877)	1.97
QQQ [NASDAQ]	-0.0001 (-0.8698)	0.0340*** (3.6179)	-0.0090 (-0.4489)	0.0544 (0.7179)	0.0289 (0.6330)	1.26
XLF [Finance]	-0.0000 (-0.7034)	0.0340*** (4.0151)	0.0110 (0.8614)	0.0939 (1.4558)	0.0287 (0.7112)	2.15
IYR [Real Estate]	0.0002** (2.5444)	0.0321*** (4.1693)	0.0233* (1.8391)	-0.0091 (-0.1106)	0.0534 (1.5668)	2.04
IWM [Small-Cap]	0.0001 (1.3709)	0.0236*** (2.6280)	0.0132 (1.0224)	-0.0009 (-0.0167)	0.0294 (0.9111)	0.76
Non-US ETF	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
EEM [Emerging]	-0.0000 (-0.5131)	0.0215*** (2.8544)	-0.0009 (-0.0922)	0.0808 (1.2928)	0.0342 (0.8164)	0.95
FXI [China]	-0.0001 (-1.0609)	0.0223*** (2.7922)	-0.0101 (-1.6133)	-0.0109 (-0.1602)	0.0636* (1.7049)	0.92
EFA [Non-US]	0.0000 (1.0330)	0.0127** (2.1457)	-0.0016 (-0.2057)	0.0418 (0.7786)	-0.0109 (-0.2509)	0.24
VWO [Emerging]	-0.0001 (-1.2608)	0.0169** (2.2976)	-0.0035 (-0.3749)	0.0790 (1.2339)	0.0447 (1.0145)	0.75
Non-Equity ETF	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
TLT [Bond Market]	0.0001 (1.3886)	0.0020 (0.3879)	0.0238*** (3.4643)	0.0092 (0.2548)	-0.1601*** (-4.9402)	3.56

Notes: This table reports the results of the equation $r_{13,t,x} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t,x} + \beta_3 r_{12,t,x} + \beta_4 r_{13,t-1,x} + \epsilon_t$, where x={SPY, QQQ, XLF, IWM, DIA, EEM, FXI, EFA, VWO, IYR, TLT}. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1,258 observations).

TABLE 1.11: Predictive regression - Price reversal over the next trading day

Period	α	β_1	β_2	β_3	Adj-R^2 (%)
1st (First) Half-Hour	0.0003 (1.7010)	0.0125 (0.5416)	-0.0566 (-1.0475)	-0.0798 (-0.5147)	0.13
2nd Half-Hour	0.0000 (0.2412)	-0.0039 (-0.3034)	0.0366 (1.4359)	0.0416 (0.5363)	0.15
3rd Half-Hour	0.0000 (0.1393)	0.0002 (0.0232)	-0.0143 (-1.1623)	-0.0210 (-0.5346)	-0.04
4th Half-Hour	0.0001 (1.4095)	0.0071 (1.1097)	-0.0013 (-0.1008)	0.0510 (1.4513)	0.14
5th Half-Hour	0.0000 (0.6032)	0.0001 (0.0183)	0.0116 (0.9775)	0.0162 (0.3548)	-0.06
6th Half-Hour	0.0000 (0.2331)	-0.0005 (-0.0883)	-0.0005 (-0.0572)	-0.0304 (-0.8702)	-0.10
7th Half-Hour	0.0001* (2.2981)	0.0023 (0.4475)	0.0076 (0.8243)	0.0181 (0.5778)	-0.02
8th Half-Hour	-0.0000 (-0.2736)	-0.0132* (-2.1961)	0.0310* (2.3822)	-0.0181 (-0.5229)	1.21
9th Half-Hour	-0.0000 (-0.1111)	-0.0050 (-0.9159)	-0.0017 (-0.1612)	0.0045 (0.1302)	-0.13
10th Half-Hour	0.0001 (1.5648)	-0.0019 (-0.2860)	-0.0019 (-0.1623)	-0.0317 (-0.7760)	-0.12
11th Half-Hour	0.0000 (0.5419)	-0.0163* (-2.5282)	0.0175 (1.3305)	-0.0032 (-0.0641)	0.39
12 Half-Hour	0.0001 (1.2226)	-0.0163* (-2.1070)	0.0285 (1.5629)	0.0161 (0.3186)	0.71
13th (Last) Half-Hour	-0.0000 (-0.5232)	0.0074 (0.8749)	-0.0009 (-0.0330)	-0.0718 (-1.3239)	0.24

Notes: This table reports the results of the equation $r_{i,t+1} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{13,t} + \epsilon_t$ for i={1,12,...,13}. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1,257 observations).

TABLE 1.12: Distribution of users' self-reported investment approach, holding period and experience level

	Users		**Messages**	
	Number	Percentage	Number	Percentage
Investment Approach				
Technical	29,104	12.12%	13,177,530	22.11%
Fundamental	9,541	3.97%	3,936,066	6.60%
Global Macro	2,425	1.01%	872,404	1.46%
Momentum	13,533	5.64%	6,003,008	10.07%
Growth	13,111	5.46%	4,590,279	7.70%
Value	7,295	3.04%	3,346,318	5.61%
Holding Period				
Day Trader	16,462	6.86%	6,046,038	10.14%
Swing Trader	29,956	12.48%	13,223,008	22.18%
Position Trader	15,514	6.46%	6,003,489	10.07%
Long-Term Investor	15,026	6.26%	6,344,566	10.64%
Experience Level				
Novice	25,686	10.70%	5,260,787	8.83%
Intermediate	36,082	15.03%	14,499,167	24.32%
Professional	14,619	6.09%	11,779,219	19.76%

Notes: This table reports the distribution of users' self-reported investment approach, holding period and experience level. Percentage is calculated as the number of users (or messages) who self-reported a given trading strategy in their profile divided by the total number of users (or messages) in the sample.

TABLE 1.13: Predictive regression - Investor sentiment by investment approach, holding period and experience level.

Investment Approach	[1]	[2]	[3]	[4]	[5]
$r_{1,t}$	0.0156	0.0248	0.0226	0.0210	0.0239
	(0.7946)	(1.3942)	(1.2225)	(1.1514)	(1.3368)
$r_{12,t}$	0.1065	0.1039	0.1051	0.1030	0.1032
	(1.2613)	(1.2259)	(1.2462)	(1.2275)	(1.2317)
$\Delta s_{1,t,technical}$	0.0217*				
	(2.5564)				
$\Delta s_{1,t,fundamental}$		0.0037			
		(0.4132)			
$\Delta s_{1,t,momentum}$			0.0163		
			(1.3456)		
$\Delta s_{1,t,growth}$				0.0212*	
				(2.1436)	
$\Delta s_{1,t,value}$					0.0210*
					(2.1051)
Adj-R^2(%)	1.65	1.03	1.19	1.38	1.44

Holding Period	[1]	[2]	[3]	[4]
$r_{1,t}$	0.0233	0.0195	0.0208	0.0240
	(1.2949)	(1.0120)	(1.1219)	(1.3328)
$r_{12,t}$	0.1034	0.1055	0.1012	0.1037
	(1.2256)	(1.2486)	(1.2031)	(1.2277)
$\Delta s_{1,t,day}$	0.0154			
	(1.2547)			
$\Delta s_{1,t,swing}$		0.0178		
		(1.7557)		
$\Delta s_{1,t,position}$			0.0206*	
			(2.0494)	
$\Delta s_{1,t,long}$				0.0097
				(1.1156)
Adj-R^2 (%)	1.17	1.31	1.36	1.10

Experience Level	[1]	[2]	[3]
$r_{1,t}$	0.0194	0.0186	0.0194
	(1.0796)	(0.9882)	(0.9950)
$r_{12,t}$	0.1054	0.1051	0.1050
	(1.2551)	(1.2504)	(1.2410)
$\Delta s_{1,t,novice}$	0.0306**		
	(3.2360)		
$\Delta s_{1,t,intermediate}$		0.0243*	
		(2.2976)	
$\Delta s_{1,t,professional}$			0.0154
			(1.7427)
Adj-R^2 (%)	1.77	1.51	1.33

Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t,x} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \epsilon_t$. As the constant α is not significant in any regression, we do not report results for α for readability. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1,258 observations).

TABLE 1.14: Trading strategy performance

Strategy	Mean (%)	Std Dev (%)	Sharpe Ratio
Sentiment-Driven Strategy	4.55	3.042	1.496***
Always Long Strategy	-0.632	3.055	-0.207
First Half-Hour Strategy	1.66	3.054	0.544
12th Half-Hour Strategy	0.566	3.055	0.185

Notes: This table reports the annualized mean returns, standard deviations and Sharpe ratios of trading strategies relying on different signals to buy (sell) S&P 500 ETF index at 3:30 p.m. on day t and sell (buy) it at market close on the same trading day. Superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively, using a simulation-based p-value for the Sharpe ratio significance level.

Chapter 2

Market reaction to news and investor attention in real time

Abstract

We propose a new framework for proxying investor attention in real time by analyzing the Twitter messages of financial experts around the release of unscheduled news announcements. Using high-frequency data on large-cap U.S. stocks from January 2013 to December 2015, we find evidence that news events receiving attention on social media lead to large and persistent changes in trading activity, volatility and price jumps. When investors do not pay attention to news, however, the effects of news on such trading patterns tend to be smaller and vanish quickly. With respect to timing, we find that approximately one fourth of the news stories arrive first on Twitter before being reported by Bloomberg. This result suggests that movements prior to news releases may not be explained only by private information, but could also be related to timestamp delays. We control such potential biases with attention-adjustment and newswire-corrected timestamps, which partially eliminates the pre-announcement effect.

Keywords: Investor attention, News announcements, Intraday market dynamics, Social media

JEL classification: D83, G12, G14.

Co-authored with Deniz Erdemlioglu (IÉSEG School of Management) and Roland Gillet (Université Paris 1 Panthéon-Sorbonne)

"*You see, my competitors, they were fixated on sucking it up and monetizing via shopping and social media. They thought that search engines were a map of what people were thinking. But actually they were a map of how people were thinking. Impulse. Response. Fluid. Imperfect. Patterned. Chaotic.*"

Ex Machina, Dir. Alex Garland. DNA Films, 2015. Movie.

2.1 Introduction

The flow of information plays a central role in financial markets. Macroeconomic announcements and firm-specific news often affect trading activity, market volatility and price dynamics. To understand the link between trading patterns and news flows, it is important to study how investors process the news and react accordingly by filtering value-relevant information from noise. In this paper, we propose a new framework for examining market reaction to news through investor attention to information in real time. Developing an attention network and using intraday one-minute data on large-cap U.S. stocks, we find evidence that news strongly influences trading activity only when investors pay attention to news announcements. In the absence of such attention, however, the effect of news on returns, volume, volatility and price jumps appears to remain weak and short-lived.

Measuring attention to financial news is challenging in continuous time. To cope with this difficulty, we use the flow of tweets from financial experts around the release of firm-specific unscheduled news. Although most of the 500 million messages sent every day on the micro-blogging platform Twitter are noisy, visual inspection typically suggests that certain Twitter posts—conveying value-relevant information—lead to large movements in financial markets. For example, on March 30, 2015, Tesla's stock price jumped by nearly 2% after a tweet from Tesla CEO Elon Musk announcing a new product line (Figure 2.1). On April 28, 2014, Twitter's stock price dropped by 5% after a tweet from Selerity, a FinTech company, leaking Twitter earnings results one hour early (Figure 2.2). On June 24, 2015, Netflix's stock price decreased sharply following a tweet from the investor Carl Icahn, who announced that he had sold his stake in the company (Figure 2.3). All these events, characterized by extremely active posting activity on social media, are associated with a large and long-lasting impact on trading volume, volatility and asset prices.

[Insert Figures 2.1–2.3 about here]

Even though these examples primarily reveal the reactions of individual stocks, tweets can also amplify marketwide movements. For example, on April 23, 2013, at 1:08 p.m., the Dow Jones Industrial Average index plunged by nearly 145 basis points in one minute after a fake tweet announced the bombing of the White House. Although the Dow Jones recovered to its previous level a few minutes

after the news proved to be false, this exceptional event illustrates the speed at which information can be shared and disseminated on social media and the link between investor attention and information in financial markets.

In light of these examples, the objective of this paper is to examine the intraday response of stocks to news announcements by proxying investor attention to information. Our analysis has two main steps. First, we develop a procedure that uses Tweet flow as a proxy for investor attention. In real time, this method tracks all messages posted by financial experts and disentangles (value-relevant) signals from (irrelevant) noise. Second, we use our attention-based metric to characterize the effects of news events on stock market activity in the form of several reactions, consisting of abnormal returns, trading volume, price volatility, and price jumps.[1] We compare market reactions in the presence (absence) of attention and we show that the reaction of stocks to news depends on whether investors pay attention to news or not.

Compared to data from traditional newswires, data from a micro-blogging platform can provide several advantages. First, Twitter may *break the news* because existing value-relevant information can be shared on social media before being reported by mainstream newswires (Kwak et al., 2010). Second, Twitter, by itself, may even *create the news*. Companies can use the Twitter platform to announce key information to investors in compliance with Regulation Fair Disclosure, since the Security and Exchange Commission (SEC) reports on the use of social media by companies and markets participants.[2] Third, by utilizing a wider number of "content providers" instead of focusing on a specific newswire, Twitter enables the construction of a transparent measure of news relevance and investor attention. As "news stories are not all created equal" (Barber & Odean, 2008), exploiting social media activity associated with a specific news release can help researchers and practitioners divide market reaction into (attention-grabbing) value-relevant news and noisy signals.

At low trading frequencies (such as daily or weekly), prior research documented mixed evidence on the power of search engines, social media and Internet communication in predicting asset returns

[1] Although the high-frequency market reaction to intraday stock-specific news using data from traditional newswires (Reuters or Dow Jones Newswire Services) has already been studied in the literature (Groß-Klußmann & Hautsch, 2011; Boudt & Petitjean, 2014), this paper is, to the best of our knowledge, the first to use data from Twitter for this purpose.

[2] Regulation Fair Disclosure applies to social media and other emerging means of communication used by public companies in the same way it applies to company websites. See "SEC Says Social Media OK for Company Announcements If Investors Are Alerted" (April, 2, 2013).

(see, Nardo et al., 2016, for a survey). Although those findings could be explained through market efficiency, other factors could also justify the absence of predictability. First, information may be rapidly incorporated into asset prices, which requires an intraday analysis to better understand the causal relationship between online user-generated content and market dynamics. Second, value-relevant messages may, in fact, be lost in a massive flow of noisy content, leading to, on average, noisy signals. In this respect, accounting for users' credibility and reputation could help disentangle news from noise. Third, investor attention (based on online discussions) strongly correlates with traditional newswires stories. Therefore, combining online messages with traditional news flows could permit identifying "news that matters" from among noisy stories and routine news coverage. In this paper, we provide a new methodology and empirical analysis to explore these three competing explanations.

To disentangle news from noise, our methodology relies on network theory and identifies *experts in the crowd*. Specifically, we first start with a list of influential *Twitterers* (i.e., contraction of "Twitter" and "user") sharing opinions, views and news about the stock market on Twitter. Then, we implement an iterative algorithm based on directed relationships (friendships) between *Twitterers* to characterize a network of thousands of financial experts. After identifying experts in the crowd, we consider five listed U.S. companies consisting of Apple, General Electrics, Walmart, Johnson & Johnson, and IBM. We extract all messages sent on Twitter by our list of financial experts.

When implementing our approach, we combine Twitter data with firm-specific "Hot Headlines" (HH) from the Bloomberg Terminal, focusing on unscheduled news published during the market's opening hours.[3] For each of the HH, we utilize a similarity measure to examine whether or not the news was available on Twitter before being reported by Bloomberg. This measure also allows us to automatically compute a transparent proxy of investor attention by analyzing the similarity between HH content and messages published on Twitter around the release of the news.

Relying on our investor attention measurement, the empirical analysis reveals several distinct patterns and regularities in the data. First, for a large number of news events, we find that the content of the unscheduled news was already available on Twitter before being published on Bloomberg. The evidence—documented for other types of events, such as the death of Osama bin Laden or the plane

[3]For instance, our unscheduled news announcements were related to product announcements, events announcements, activist investors communication, and legal issues, among others.

crash in the Hudson River—confirms that Twitter can also "break" financial news. One implication of this result is that combining Twitter flow with traditional newswires' data can help locate the exact timestamp at which public information was available to market participants. This, in turn, allows for better assessment of the impact of unscheduled news on the financial markets' movements. Along these lines, we use one-minute intraday data and newswire-corrected timestamps to investigate market reactions to news that receive high-attention versus low-attention from *Twitterers*. We find evidence that attention-grabbing news are followed by large and persistent changes in trading volume, volatility, price jumps. When investors instead pay no attention to news, the impact on such measures of market reaction is very low and the effect of news vanishes very quickly at intraday levels. For both high-attention and low-attention news, we do not find any price predictability after news releases, consistent with the efficient market hypothesis.

Overall, our results suggest that Twitter can help trace news events that matter in a continuous flow of information, without relying on "black box" pre-processed data or subjectively picking seemingly relevant news. By combining Twitter flow with high-frequency news events from traditional newswires, researchers can disentangle the effects of pre-announcement private information from ambiguous timestamp identification and avoid underestimating the impact of unscheduled news cause by noisy stories or rumors.

The remainder of this paper is organized as follows. Section 2.2 discusses the related literature. Section 2.3 describes the identification algorithm we use to construct the tweets database. This section further presents the intraday stock data and Bloomberg data. Section 2.4 introduces our methodology to proxy investor attention in real time and outlines an event-study analysis. Section 2.5 discusses our empirical results. Section 2.6 concludes.

2.2 Related literature and hypothesis

Modern finance theory suggests that "news"—defined as textual information from traditional media—should not influence stock markets, unless the news events contain value-relevant information about the discounted value of future cash flows. Assuming that information revealed by traditional media is stale due to the publication lag, media should thus play no role in the price discovery process.

A few decades ago, before the advent of the Internet and the availability of high-frequency data, one could argue that information disseminated by traditional daily morning and afternoon newspapers was stale when made public. This conclusion is, however, questionable in today's financial markets, where an almost continuous flow of news can be exploited by fast-moving traders (Foucault et al., 2016) and by machine reading the news (Groß-Klußmann & Hautsch, 2011). Furthermore, news is no longer the monopoly of traditional media, that is, every user can now be a media outlet by publishing content on blogs, message boards, or social media (Shirky, 2008). U.S. companies can now directly use social media to disseminate key information to investors, in compliance with Regulation Fair Disclosure. Traditional media are still among the main news providers, but their business model has evolved from a daily newspaper to a continuous flow of online information, where breaking news often plays a significant role in developing online traffic. These recent technological, organizational, and regulatory changes reinforce the need for empirical research on the informational efficiency of financial markets.

The literature on the high-frequency market impact of scheduled macroeconomic releases (Andersen et al., 2007; Bollerslev et al., 2016) and Federal Open Market Committee announcements (Faust et al., 2007; Wongswan, 2009) is vast. High-frequency scheduled news (surprises) has a significant impact on asset prices and, typically, explains a substantial fraction of the increase in price volatility and trading volume following the news (Balduzzi et al., 2001). Recently, Bernile et al. (2016) also document substantial informed trading before the official release time of scheduled announcements, consistent with information leakage from the news media or from insiders. However, research on the impact of unscheduled news arrivals is relatively scarce. While macroeconomic events often affect the movements of individual stocks, sudden and unexpected firm-specific information can also impact asset pricing and market liquidity (Boudt & Petitjean, 2014). For example, analyzing the intraday market dynamics of firm specific unstructured news at the Paris Bourse, Ranaldo (2008) finds a significant increase in liquidity and higher adverse selection costs around news arrivals. Similar results have been found for the Toronto Stock Exchange by Riordan et al. (2013) using the Thomson Reuters newswire messages.

When analyzing the impact of unscheduled news on stock returns, liquidity, or volatility, researchers and financial economists reconstruct news databases by searching news events about a given company on Factiva, Thomson Reuters, Dow Jones News Services, Lexis-Nexis, or the Wall Street Journal (Das, 2014). However, contrary to precisely timed macroeconomic announcements, the identification of the exact timestamp of unscheduled news events is non-trivial. Moreover, as the number of news articles published steadily increases, tracing value-relevant news in an overwhelming number of articles published every day is a major challenge. Pre-processed news data such as the Reuters NewsScope Sentiment Engine (Riordan et al., 2013) or the RavenPack News Analytics (Smales, 2014) can help solve issues related to the identification of value-relevant news. Proprietary "black box" algorithms developed by a few private companies automatically assess novelty, relevancy, and sentiment scores for all articles from a list of news providers. For example, Groß-Klußmann & Hautsch (2011) use intraday pre-processed news from Reuters to analyze the high-frequency trading impact of unscheduled news releases by separating announcements into relevant and non-relevant news. They notice a significant positive (negative) price movements prior to positive (negative) relevant news releases, but only a weak return response thereafter. Volatility, liquidity, and trading volume typically start increasing about 60 minutes prior to the news release, peak at the exact time of the release, and decrease later in the day. Although pre-event movement could be explained by private information, the authors argue that the availability of other sources of information and an induced clustering of news items are mainly responsible for pre-announcement effects. This finding and the absence of transparency of the algorithms used to derive articles' scores (novelty, relevancy, and sentiment), encourage further research in the area, not only to improve the timestamp of news detection but also to disentangle value-relevant news from noisy content with a more robust methodology.

Building on progresses made in the area of natural language processing, named entity recognition, and topic classification, recent papers on computational science focus on methodologies to automatically detect "breaking stories" in a continuous flow of messages from social media (Mathioudakis & Koudas, 2010; Petrovic et al., 2013; Ifrim et al., 2014). The basic intuition behind event detection is as follows. When value-relevant breaking news arrives, users on social media will change their posting activity and start talking very actively about the event. By analyzing the tweet flow in real

time, looking for surge in absolute posting activity, a burst in the frequency of certain keywords, or the appearance of new topical clusters, practitioners can identify value-relevant news even in the absence of a specific news provider or another measure of relevancy. Historical databases of messages are available (albeit expensive), so that discovering the precise timestamp at which information was public is possible (at the ex-post level) by identifying the first mention of a news article on social media.

Because all news is not created equal, Twitter can also help disentangle value-relevant attention-grabbing news from noisy content. Theoretical models and empirical studies often suggest that investor attention to marketwide news plays a central role in determining asset prices and volatility (Li & Yu, 2012; Andrei & Hasler, 2015; Yuan, 2015). At the company level, Barber & Odean (2008) provide evidence that retail investors are net-buyers of attention-grabbing stocks, and Solomon et al. (2014) document that media coverage attracts investor attention and affects investors' capital allocations to mutual funds. Recently, Boulland et al. (2017) demonstrate that investor attention, proxied by the use of an English-language electronic wire service by European firms to disseminate company news, affect market reactions to earnings surprises.

Indirect proxies, such as 52-week high (Driessen et al., 2013), the day of the week (Friday effect) (DellaVigna & Pollet, 2009), or the level of media coverage (Barber & Odean, 2008), have been used in the literature to proxy investor attention. Despite the substantial progress in the research, using these proxies has certain important caveats. On the one hand, market data (such as price and volume) contain idiosyncratic components that are unrelated to attention. On the other hand, simply counting the number of news articles does not take into account the salience of news coverage and could be easily affected by routine company press releases.

To overcome issues related to proxy selection, recent studies examine the number of queries about a given company on Google to compute a more direct measure of investor attention (see, Da et al., 2011; Dimpfl & Jank, 2016). Although the Google search engine is of interest from a practical viewpoint, proxying investor attention through the evolution of online search behavior also has certain limits. For instance, the Google data are available only on a daily basis with no information on the absolute level of search (scaled on a range of 0 to 100 based on a topic's proportion to all searches

on all topics). Thus, the precise assessment of the evolution in the Search Volume Index remains elusive. In this respect, combining news published on traditional media with attention toward the news on Twitter could resolve issues related to partial identification. It also allows the construction of a transparent high-frequency proxy of investor attention. We attempt to bring this resolution to research on news reaction analysis.

Following the existing theoretical and empirical literature, we hypothesize that news arrivals cause price jumps and are followed by a persistent increase in volatility and trading volume. We add to the literature by examining two further hypotheses. The first hypothesis focuses on movements prior to the news release and on the importance of using newswire-corrected timestamps for intra-day studies, following results from Bradley et al. (2014) showing that timestamp delays could lead to incorrect inferences. We hypothesize that the pre-announcement effect is generally overestimated when newswires timestamps are considered, and that correcting for timestamp delays significantly reduces movements prior to news releases. The second hypothesis is closely related to the theoretical framework of Andrei & Hasler (2015): high attention should induce high return volatility if attentive investors immediately incorporate new information into prices. Conversely, when investors pay little attention to news, information should only be gradually incorporated into prices so volatility is low. We thus test empirically the hypothesis that market reaction (volatility, price jumps, trading volume and return) to news is much more pronounced when investor pay attention to news than when they are not.

2.3 Data

We proceed with the description of our databases. In Section 2.3.1, we first present the Twitter data and its adjustment. Section 2.3.2 details the data on stocks and news announcements extracted from the Bloomberg Data Analytics.

2.3.1 Twitter data as proxy for attention

On the micro-blogging platform Twitter, users can post short messages, called "tweets" (140-characters maximum), share messages sent by other users with their community of followers ("retweet"), or

simply read "tweets" of users they choose to follow. Compared to other social networks, such as Facebook or LinkedIn, relationships between users on the platform are public and can be accessed through the Twitter Application Programming Interface (TAPI). The public character of Twitter allows us to transform the Twitter network into an adjacency matrix and to identify specific clusters related to the domain of interest. In our analysis, we use Twitter directed relationships to identify a list of *financial influencers*. We focus on the tweets of important investors (through pre-filtering), financial journalists, and experts working in financial markets or institutions.

Specifically, we start with a list of 10 influential *Twitterers* (i.e, contraction of "Twitter" and "user") sharing news and ideas about the stock market on Twitter. We impose the following four criteria to include a user in our initial list: (1) the user has a verified Twitter account, (2) the user has a dedicated Wikipedia page, (3) the user has at least 100,000 followers, and (4) the user has a job related to financial markets.[4] Table 2.1 presents our initial list of 10 users. We denote this set as N_0.

[Insert Table 2.1 about here]

We conjecture that common friends of influential experts in finance should also be influential and tweet regularly about financial markets. We use the TAPI to extract the friends list of each user in N_0.[5] We then insert the unique identifier of all users followed by at least one user from N_0 into a MongoDB database, ending up with a list of 15,390 users. Finally, we construct a new list N_1 by augmenting N_0 with the 50 most commonly followed *Twitterers* from the list of 15,390 users.[6] Appendix A details our setup and implementation of the network algorithm used to characterize attention. Figure 2.4 shows our constructed network N_1 based on this setup. We notice that the information network is highly connected, with a total of 973 directed links between the 60 users from N_1.

[Insert Figure 2.4 about here]

[4]The final list of 3,010 users identified using our methodology is robust to the initial list of 10 users chosen. We find a similarity of 85-95% when considering other lists of 10 financial experts.

[5]https://dev.twitter.com/rest/reference/get/friends/list.

[6]These *Twitterers* added during the first iteration include Elon Musk (Tesla CEO), ZeroHedge (financial media), Citron Research (financial analyst), Blackhorse Analytics (equity research), Joe Weisenthal (Bloomberg editor), John Carney (Wall Street Journal market editor), Fred Wilson (venture capitalist), the New York Times Business section (media), Dan Primarck (journalist at Fortune), Chris Sacca (venture investor), Henry Blodget (former equity research analyst, CEO Business Insider), Horace Dediu (industry analyst), and T. Boones Pickens (hedge fund manager).

As in step 1, we iterate this algorithm by extracting the friends list of each user in N_1 and adding the 50 most commonly followed users to N_1.[7] Having generated 60 iterations, we obtain our final network N_{60} that consists of 3,010 users.[8] Our final list includes official media Twitter accounts (e.g., CNBC, Financial Times, Reuters, and Bloomberg), journalists' personal accounts (e.g., Jim Cramer, Carl Quintanilla, Maria Bartorimo, and David Faber), market participants and investors (e.g., Warren Buffet, Carl Icahn, Mark Cuban, and Marc Andreesen), CEO and insiders (e.g., Tim Cook, Elon Musk, Satya Nadella, and Marissa Mayer), institutions (CBOE, Federal Reserve, and Nasdaq) and several celebrities, such as Taylor Swift, Ellen DeGeneres, Barack Obama, and Oprah Winfrey.

Given this network, we are particularly interested in examining trading activity patterns around Twitter messages as reliable proxies for investor attention. To achieve this goal, we focus on five U.S. companies that are Apple (AAPL), Walmart (WMT), International Business Machine (IBM), Johnson & Johnson (JNJ) and General Electric (GE).[9] Those companies are amongst the 10 companies with the highest market capitalization in the U.S. as of January 1, 2013. This classification also helps us avoid a sectoral bias observed typically on high-tech companies, such as Google, Microsoft, or AT&T, or in the oil industry (Chevron, Exxon).[10]

[Insert Table 2.2 about here]

Based on this implementation, our Twitter database consists of 498,366 messages containing a keyword related to one of the five companies in the sample. For the messages of market participants from N_{60}, our adjusted data span the period between January 1, 2013, to December 31, 2015.

[7] In this case, we consider a new list of 60 influencers instead of the initial list of 10. *Twitterers* identified during the second iteration include official financial media (e.g., WSJ, Bloomberg, CNBC), financial journalists (e.g., Jenn Ablan, Dennis K. Berman, Charles Gasparino), hedge fund and portfolio managers (e.g., Doug Kass, Mark Dow, Anthony Scaramucci), and traders/venture capitalists (e.g., Paul Kedrosky, Jon Najarian, Bill Gurley).

[8] For brevity, we do not report the full list yet it is available available upon request.

[9] One challenge is that historical access to archives for keyword-related queries is rather limited. The TAPI allows registered applications to extract only the last 3,200 tweets sent by each user. To have a greater depth and retrieve all tweets since January 2013, we hence develop an application using Python. First, we rely on the new "advanced search tool" available on Twitter (since April 2014) and we extract the unique identifier of all tweets sent by users in N_{60} including keywords related to companies in our sample. For the Apple company, for instance, we extract the unique identifier of all tweets containing keywords "Apple," "$AAPL," "AAPL," "Tim Cook," "iPhone," "iPad," "iPod," "iTunes," and "Macbook" sent by experts from N_{60}. Then we use the Twitter "GET statuses/show/:id" function to retrieve detailed information about each message and insert all tweets into a MongoDB NoSQL database. Given the TAPI limits, the data collection process is limited to one message every five seconds. To collect all data, we ran our Python script during one month.

[10] In our study, we are particularly interested in investigating how markets respond to new information arrivals when investors pay attention to news and when they do not. Of course, future research can consider other asset classes with marketwise scheduled news announcements.

The most widely covered company is Apple (414,844 tweets), followed by Walmart (32,872), IBM (25,444), General Electric (17,915), and Johnson & Johnson (7,191).[11] Table 2.2 shows a sample of messages published on January 2, 2013 (the first trading day of our period).

2.3.2 News announcements and stock data

We use Bloomberg Professional Service (BPS) to extract company-specific news announcements.[12] When constructing our news database, we focus on Bloomberg "Hot Headlines" (HH) because HH are typically released very quickly by Bloomberg Analytics. In order to alert practitioners about the release of a (potentially) value-relevant announcement (e.g., a political event, a macroeconomic event or company-specific news), HH are very short (10 words on average). These hot headlines are further distinguishable in the flow of news available on Bloomberg by being capitalized and highlighted in red. The exact timestamp of the news release (up to the second) is available on Bloomberg.

We manually extract all HH relative to Apple, General Electric, IBM, Johnson & Johnson, and Walmart. Similar to the patterns of tweet flow, the most covered company is Apple, with a total of 1,528 HH between January 1, 2013, and December 31, 2015, followed by General Electric (977), Walmart (457), Johnson & Johnson (383) and IBM (323). Then, we manually filter all HH to remove duplicate events, irrelevant news, and announcements related only to variation in stock prices.[13] We further eliminate headlines about scheduled news or events. For example, a few times a year and during market trading hours, Apple organizes special events (keynote), where the company makes announcements about new products or developments. Price volatility is especially high during those scheduled events, and, given the high number of news announcements provided in a small amount of time, isolating the effect of a specific announcement is difficult.[14] As in Bollerslev et al. (2016),

[11] We conjecture the large difference between social activity about Apple and other companies in several respects. First, Apple was the company with the highest market capitalization in the world at this time. Second, Apple is the most covered company by media and a well-known company to the general public. Third, high-tech companies are, on average, more covered on social media than industrial companies. Lastly, every new product released by Apple is followed by a wave of euphoria in the real world—as fans, for instance, queuing in front of Apple stores— also visible on social media.

[12] BPS is a platform through which financial professionals can monitor and analyze real time data, news and analytics.

[13] For Apple, for instance, we remove the headline "Blackberry previews secure work space tech for Android, iOS" or the headline "Apple unchanged erasing gain of 1.4% at the open."

[14] When conducting robustness checks (unreported for brevity), we also include those news articles in our event-study. The results remain qualitatively the same.

we consider only day-trading sessions and, hence, neglect intentionally all news events published overnight and on nontrading days.

For all companies in our sample, we use one-minute data (transaction prices and volume) from January 1, 2013, to December 31, 2015.[15] As is standard in the literature, we omit trading days that have too many missing values or low trading activity. Because trading volume and volatility exhibit strong intraday patterns (due to opening and closing hours), we use the procedure of Erdemlioglu et al. (2015) and remove periodic patterns before conducting our empirical analysis.[16]

2.4 Methodology

In this Section, we describe our main methodology for examining the high-frequency response of stocks to investor attention. The next Subsection presents the underlying continuous-time model and shows how we characterize market fluctuations in various forms. In Section 2.4.2, we outline the identification of our attention measure and in Section 2.4.3, we present the event study methodology.

2.4.1 Reaction forms in continuous-time

As we attempt to examine the market impact of attention in real time, we describe the behavior of stocks at short-time scales in continuous-time. Therefore, we assume that the log-price of a stock $p(t)$ follows a standard diffusion process with jumps. While the former component helps us characterize the smooth/diffusive volatility reaction, jumps reflect abnormal uncertainty or shocks. That is

$$dp(t) = \underbrace{\mu(t)dt}_{\text{drift}} + \underbrace{\sigma(t)dW(t)}_{\text{volatility shocks}} + \underbrace{\kappa(t)dq(t) + h(t)dL(t)}_{\text{jump shocks}}, \tag{2.1}$$

where $dp(t)$ denotes the logarithmic price increment for $t \geq 0$, $\mu(t)$ is a continuous, locally bounded, variation process, $\sigma(t)$ is a strictly positive and càdlàg (right-continuous with left limits) stochastic volatility process, and $W(t)$ is a standard Brownian motion. In Equation (2.1), $q(t)$ further denotes a

[15] While a higher frequency analysis (tick-by-tick, one-second) could shed light on how market participants process information in a model with fast-moving (slow-moving) traders (Foucault et al., 2016), we restrain our analysis to one-minute data due to data availability.

[16] For brevity, we do not report the estimated periodicity factors and illustrate the intraday diurnal patterns. These results are available upon request.

counting process (e.g., compound Poisson process), $L(t)$ is a pure Lévy jump process (e.g., Cauchy process), $\kappa(t)$ and $h(t)$ denote the jump shock sizes of the counting and Lévy processes, respectively. Intuitively, the jump shocks of (2.1) potentially represent both finite- and infinite-activity. While finite-activity jumps capture rare and large abnormal reactions, the infinite-activity component tracks relatively small yet frequent jumps in asset prices.

Given this underlying model, we next estimate diffusive (spot) volatility and detect the arrivals of extreme price changes (i.e., intraday jumps). We use the truncation approach of Bollerslev et al. (2013) to identify the realized intraday jump shock increments of the assets. That is,

$$JV_{t,i} = \{i \in [0, T] : |r_{t,i}| \geq u\}, \tag{2.2}$$

where $r_{t,i}$ is the intraday price increment (return) at time t of a trading day i, $u = \alpha\Delta^{\varpi}$ is the truncation threshold and α (> 0) is expressed in units of standard deviations of the continuous part of the process for a constant $\varpi \in (0, 1/2)$. This truncation approach in (2.2) can be used to detect large price changes (i.e., jumps), and hence its reverse version retains the diffusive (or continuous) volatility shock component, such that

$$CV_{t,i} = \{i \in [0, T] : |r_{t,i}| < u\}, \tag{2.3}$$

where $CV_{t,i}$ is the estimated diffusive spot volatility of (2.1). As in Bollerslev et al. (2013), we set the truncation thresholds $\alpha = 3$ and $\varpi = 0.47$ for both jump and volatility estimations (i.e., (2.2) and (2.3), respectively). Finally, we follow Groß-Klußmann & Hautsch (2011) to estimate abnormal trading volume by standardizing the process by the yearly average of the corresponding underlying 1-minute interval, such that

$$V^*_{i,t} = \frac{V_{i,t}}{\frac{1}{250}\sum_{d=-250}^{-1} V_{d,i,t}}, \tag{2.4}$$

where $V^*_{i,t}$ denotes the abnormal trading volume on minutes t for company i, $V_{i,t}$ is the one-minute trading volume and $V_{d,i,t}$ is the trading volume of the corresponding underlying minute t on day d. To characterize the returns on high frequency, we assume a normal-return asset pricing model as in Groß-Klußmann & Hautsch (2011). Specifically, we define the abnormal return as the difference

between the actual return and the estimated normal return given by

$$R_{i,t} = \alpha_i + \beta_1 Rm_t + \beta_2 R_{i,t-1} + \epsilon_{i,t} \tag{2.5}$$

$$AR_{i,t} = R_{i,t} - \hat{R}_{i,t} \tag{2.6}$$

where $AR_{i,t}$ denotes the one-minute abnormal return of company i, $R_{i,t}$ is the one-minute return and Rm_t is the one-minute return of the S&P500 (SPY Exchange Traded Fund). In line with Fama (1998), and as we focus our analysis on a short [-30:+30] minutes event window, the model of normal returns considered barely affects the inference about abnormal returns (i.e., the expected returns on a short event window are close to zero).[17] To pin down the effects of unscheduled news on intraday returns, we implement a trading strategy of shorting stocks on negative news and of buying stocks on positive news. We define headline's sentiment manually since sentiment measurement based on standard dictionary-based approach (see, e.g., Loughran & McDonald, 2011; Jegadeesh & Wu, 2013) is likely to be biased due to the low number of words in hot headlines.

2.4.2 Linking attention to reaction

To proxy investor attention to news, we analyze all messages posted on Twitter in a [-15:0] minutes window before (and at the exact same minute) the release of each Bloomberg headline. We utilize a Term Frequency-Inverse Document Frequency (TF-IDF) cosine similarity measure to avoid considering messages posted around the release of the news but not related to the news.[18] Given the interactions between b_1 (a Bloomberg headline) and t_1 (a Twitter message) (collapsed into two TF-IDF vectors B and T), cosine similarity is given by

$$Cos^{sim}(b_1, t_1) = \frac{\sum_{i=1}^{n} B_i T_i}{\sqrt{\sum_{i=1}^{n} B_i^2} \sqrt{\sum_{i=1}^{n} T_i^2}}. \tag{2.7}$$

[17] In (unreported) robustness checks, we confirm that our results are insensitive to the choice of asset pricing model (i.e., constant-mean or market-return).

[18] Cosine similarity is a standard approach taken from natural language processing and information science literature to assess the similarity between two documents (see, Loughran & McDonald, 2016).

Because the TF-IDF value is always positive, cosine similarity ranges between 0 and 1.[19] Higher cosine similarity implies a closer similarity between a given message published on Twitter and the Bloomberg headline. Table 2.3 reports examples of cosine similarities between a Bloomberg headline and all messages sent on Twitter on a [-15:0] minutes window around the timestamp of the release of the headline on Bloomberg.[20]

[Insert Table 2.3 about here]

For each Bloomberg headline, we compute an attention variable by adding the cosine similarity between the headline and all messages sent on Twitter in a [-15:+0] minutes window around the exact timestamp of the news release from Bloomberg. Assuming there are n messages published on Twitter within this window, we finally define $NewsAttention_i$ as the level of attention to $HH\ i$

$$NewsAttention_i = \sum_{j=1}^{n} Cos^{sim}(b_i, t_{i,j}). \tag{2.8}$$

We define attention-grabbing (low-attention) news all news with a $NewsAttention_i$ score greater than (or equal to) 0. We also consider other threshold values (0.5 and 1) to separate attention-grabbing news from low-attention news and we find that results are robust to the threshold value considered (see Appendix B for an example on trading volume). For readability, we only report the results for a threshold value of 0.

2.4.3 Analyzing market reaction to news

We conduct an event-study to investigate the high-frequency impact of unscheduled news announcements. We consider four reaction forms: abnormal trading volume, abnormal returns, diffusive volatility, and sudden price jumps. We account for investor attention to Bloomberg news releases

[19] To improve the accuracy of the TF-IDF cosine similarity measure, we use a Porter stemmer to remove the commoner morphological and inflexional endings from the words in all messages and headlines. We also remove all stop-words, links, company names, and mentions from messages. For example, the headline "Apple PT cut to 530 from 660 at Nomura" became "pt cut 530 660 nomura." The tweet "It's one of the great conundrums of investing. What IS this stock? @JamesStewartNYT on whether $AAPL is growth or value. @CNBC" became "one great conundrum invest what stock whether growth valu."

[20] That is, "Einhorn drops suit against Apple over shareholder vote" (released on Bloomberg at 11:25:12 a.m. on March 1, 2013).

through the flow of tweet. Relying on this scheme, we investigate the characteristics of the market reaction to high-attention and low-attention news separately.

We set a [-30:+30] minutes event window. We follow Bollerslev et al. (2016) and eliminate news articles published during the first and the last 30 minutes of each trading day. Therefore, we derive all minutes in the event window from the same trading days, which allows us to cope with the identification issue due to overnight news and the sharp opening variation at 9:30 a.m. on each trading day. We also impose a minimum length of 30 minutes between two events for a given company to avoid problems related to overlapping or timing. Taken together, we examine the duration, news timing, and persistence of all news-implied reactions. We end up with a total of 547 events. To assess the significance of our variable of interest (return/volatility/jumps/volume), we compare the estimates on the event window with those obtained from the last trading days without any unscheduled news events during market opening hours. For example, for the Bloomberg *HH* "Apple gets 30M iPad deal from LA unified school district" published at 1:10:24 p.m. on June 19, 2013, we consider an event window from 12:40 p.m. to 1:40 p.m. (61 minutes) on that day, and we compare event window results by considering the last previous trading days without any unscheduled news as our estimation window (330 minutes of trading on June 18, 2013, from 10 a.m. to 3:30 p.m.). We carry out non-parametric Corrado (1989) rank tests for the statistical inferences.

2.5 Results

This Section presents and discusses our empirical results. In the next subsection, we propose a correction procedure to identify the timestamps of news releases. In Section 2.5.2, we decompose market news responses into high- versus low-attention components and compare the results.

2.5.1 Reaction timing: does Twitter break the news?

Twitter provides incentives for users to try "breaking the news". Indeed, publishing information on Twitter before the release on traditional newswires could increase users' credibility and reputation. Even official media Twitter accounts (e.g., CNBC, Reuters) and journalists associated with traditional media tend to publish "breaking news" on Twitter before reporting the news on their websites or

platforms. By increasing their reputation and their number of followers, the media (and journalists) can increase readership and maximize future revenues derived from traffic acquired through Twitter. Investors and market participants also have incentives to share breaking news on social media to increase their own reputation or to influence other investors.

For all 547 pre-identified Bloomberg news events, we start by comparing the exact timestamp of the *HH* (up to the second) with the first mention of the same news on Twitter. Manually analyzing all messages with a positive cosine similarity sent on a [-15:0] minutes window around the release of each *HH*, we find that Twitter effectively *breaks* 127 news events out of 547 (23.22%). In previous example shown in Table 2.3, we identify a mention of the news on Twitter two minutes before Bloomberg release. At 11:23:45 a.m., Kaja Whitehouse, a New York Times reporter covering crime and corruption, published the following message on Twitter "David Einhorn withdraws lawsuit against Apple. Manhattan federal court approves." Bloomberg headline was then posted at 11:25:12 a.m. "Einhorn drops suit against Apple over shareholder vote." Table 2.4 presents examples of cases for which Twitter breaks the news.[21]

[Insert Table 2.4 about here]

The delay between newswire-reported timestamps and the very first moment at which news arrive on social media (and hence becomes public) tend to support the conclusion of Groß-Klußmann & Hautsch (2011) on high-frequency news-implied market reactions. Price movements prior to news releases may not be solely attributed to private pre-release information, but could be explained by biased (or imperfect) timestamps. As also shown by Bradley et al. (2014) for analysts' recommendations, identifying the exact minute at which an unscheduled news event was made public is crucial for a high-frequency analysis, as a failure to do so can lead to incorrect inferences.

Before analyzing whether the degree of attention influences how markets respond to news announcements, we first conduct an event-study to assess if combining traditional newswire data with Twitter helps disentangling the effects of private information from misspecification of the exact timestamp of news releases. More precisely, we compare market reaction to news considering (1) all *HH*

[21] The duration of release time ranges between few seconds and a few minutes.

using Bloomberg reported timestamp as the event minute and (2) all HH considering the first mention of the news on Twitter (when social media "breaks the news") and Bloomberg reported timestamp otherwise. Table 2.5 reports the results for each 5-minutes interval around the release of HH.

[Insert Table 2.5 about here]

As expected, we find a strong and significant increase in volatility, price jump, trading volume and abnormal return around the release of unscheduled news announcements. More interestingly, we find that considering Bloomberg reported timestamp tends to overestimate the magnitude of the movements prior to the news release, especially for price jumps and trading volume. For example, considering the 5 minutes prior to the news release, trading volume and price jumps are overestimated by around 10% due to poor timestamp identification. Figure 2.5 illustrates the overestimation of trading volume prior to the news announcements when timestamp delays are not taken into account.

After correcting for timestamp delays using Twitter to identify (if any) the first mention of the news on social media, we find that the pre-announcement effect disappears for volatility and trading volume, and is significantly reduced for price jumps. While the improvement may seem limited, timestamp delays have serious implications for intraday studies. Therein, we provide evidence that combining the Twitter messages of financial experts with traditional media news stories allows a precise identification of the exact minute at which news was made public. This, in turn, allows academics and practitioners to better understand the role (if any) of private information in the price formation process and the magnitude and length of market reaction to unscheduled news. In the remainder of this paper, we use newswire-corrected timestamps to conduct intraday event-studies.

2.5.2 Decomposing responses into attention-grabbing news and low attention news

We now turn to assess whether the degree of attention influences how markets respond to news announcements. Table 2.6 reports the results for each minute surrounding the announcement releases. Table 2.7 reports the results for each 5-minute intervals. Figures 2.6, 2.7, 2.8, 2.9 and 2.10 illustrate the patterns for, respectively, volatility, price jumps, abnormal trading volume, abnormal return and cumulative abnormal return.

[Insert Tables 2.6 and 2.7 about here]

[Insert Figures 2.6-2.10 about here]

As in Andrei & Hasler (2015), we find that the volatility is significantly higher after the release of attention-grabbing news, while not significantly higher for low-attention news. The reaction is statistically significant up to 15 minutes after the release of news, peaking around 10 minutes after the release of news before slowly decreasing. Price volatility is, on average, 50% higher following high-attention news than following low-attention news. Turning to jump-type tail reactions, we also find a large impact differential between high- and low-attention news. Price jumps are significantly more frequent from two minutes before the release up to five minutes after the release of high-attention news. However, we do not find any significant increase in the number of price jumps for low-attention news, except at the exact minute of the release. Overall, when considering a [+1:+5] minutes interval after the release, the probability of having a jump is more than four times higher for high-attention news than that for low-attention news. Our results are consistent with those in Dewachter et al. (2014) on Euro area official communication impact on the foreign exchange market. Using the same measures of volatility and jumps, Dewachter et al. (2014) find that central bank unscheduled communication triggers large jumps and a significant increase in volatility for approximately an hour after the news release. We provide evidence that similar effects also exist on the equity market: unscheduled company-specific news triggers large and significant increase in market uncertainty. We show that this result particularly holds when investors give attention to news.

We also find similar patterns for trading volume. For instance, we notice that trading volume is, on average, two times higher following high-attention news than following low-attention news. Furthermore, trading volume remains statistically abnormal for up to 30 minutes after the release of high-attention news, whereas the effects die out very quickly (within five minutes) for news events that do not receive attention from market participants. These results are broadly consistent with Groß-Klußmann & Hautsch (2011), who utilize the same measure of abnormal volume and find that the money value traded is around 2.6 times higher following the release of high-relevance news and only 1.5 times higher for low-relevance news. Tweet flow and investor attention thus may help disentangle relevant news events from those having only noisy signals. We conjecture that the comparability of

our results could be due to a strong correlation between our measure of investor attention and the "relevancy" indicator provided by the Reuters Newscope Sentiment Engine used by Groß-Klußmann & Hautsch (2011). Regarding abnormal returns, we find an increase of 0.047% (0.013%) at the exact minute of the release of attention-grabbing (low-attention) unscheduled news releases, but no predictability after the release (no momentum nor price reversal). The market is efficient enough, in the sense of Jensen (1978), that a trader cannot generate profits based on widely disseminated news without acting almost immediately. This finding is consistent with the intraday event-study of Busse & Green (2002) related to the analysts' views broadcasted on CNBC TV.

Overall, we find that correcting timestamp delays by combining newswires-reported timestamps with social media content partially eliminates the pre-announcements effects and that the degree of attention strongly influences how markets respond to news announcements. According to Hirshleifer & Teoh (2003), the immediate reaction to news within a short event windows suggests that some investors turn their attention very rapidly to relevant announcements. In line with this argument, our findings are thus consistent with the theoretical model of Andrei & Hasler (2015) on the role played by investors' attention to news in determining volatility. In this regard, we provide empirical evidence that unscheduled attention-grabbing news are also significantly followed by large and persistent market impact on trading volume (for up to 30 minutes) and prices jumps (for up to 5 minutes). While a trading strategy based on unscheduled news releases is not profitable using one-minute data (as in Groß-Klußmann & Hautsch, 2011), even after correcting for timestamp delays and taking into account the level of attention, the returns trends might still be exploited by algorithmic (fast-moving) traders able to trade at the exact second of the release of the news. We encourage further research in this area, as well as future research analyzing the impact of attention to news for smaller companies for which the level of attention might have more impact on the price dynamics, as shown for example by Huberman & Regev (2001).

2.6 Conclusion

This paper develops a new measure of investor attention by combining the news flows from conventional newswires with the tweet flow of market participants and financial experts. We find evidence

that investor attention can help identify the exact time at which a news event becomes publicly available. Firm-specific announcements often break on Twitter before being reported on newswires, which in turn allows researchers to better understand the role of private information in affecting trading processes prior to the official news releases. Market reaction to news in pre-announcement spells could be related to biases in the timestamp of news releases rather than information leakages.

Our results also suggest that the degree of attention on Twitter about particular news changes the trading activity of stocks. While unscheduled attention-grabbing news are significantly followed by large and persistent market impact (on volume, volatility and price jumps), low-attention news flow fails to move trading. The price impact of new information is large only if investors give close attention to news. Studies on empirical asset pricing and market structure may hence incorporate the attention factor which captures how investors view and interpret the information content of news announcements.

Our study offers several lines for future research. One important direction is to examine the interaction between attention and trading patterns at ultra high-frequency (UHF) scales such milliseconds or nanoseconds. Studying UHF price dynamics may hence permit to uncover whether market reactions to news published on Twitter are driven by algorithmic traders, who can use textual analysis to automatically derive trading signals from tweet flows.[22] Another line would be to investigate the heterogeneity in market reaction depending on the credibility or reputation of the user who sends the tweet. For example, in the case of Twitter earnings leak by a FinTech company called "Selerity Corp", the impact on financial markets was partially muted as practitioners had been debating online about the veracity of the message and figures provided by the Fintech company.[23] When Reuters Twitter account confirms Selerity information, market reacts strongly, which reflects how opinions and credibility could affect the speed of adjustment to news events. Last but not least, we believe that an interesting path for future research would be to analyze the role of dissemination in market liquidity at the intraday level, extending previous findings from Blankespoor et al. (2013) on the relation between firm-initiated news via Twitter, bid-ask spreads and abnormal depths. In that vein, combining Twitter

[22] Unlike a machine, because an investor needs at least few seconds to read a tweet and pass an order, we believe that the speed of reaction to news could also help disentangle pure algorithmic trading from human trading.

[23] As we discuss in the introduction and illustrated on Figure 2.2.

data with firm-initiated traditional press releases could help understanding the relation between the level of attention and information asymmetry in financial markets.

2.7 Appendix A - Setup and implementation of the network algorithm

We consider all active users m in Twitter. As of January 1, 2016, m≈300 million. Relationships in Twitter are formalized on an $m \times m$ matrix, where $a_{i,j} = 1$ if user i follows user j, and $a_{i,j} = 0$ otherwise ($i \in m$, $j \in m$). We proceed as follows.

Step 1. We select 10 users i (i_1, i_2, ..., i_{10}) and we denote this list N_0. For each user $j \notin N_0$, we compute a variable of influence by defining $c_j = \sum_{i=i_1}^{i_{10}} a_{i,j}$.

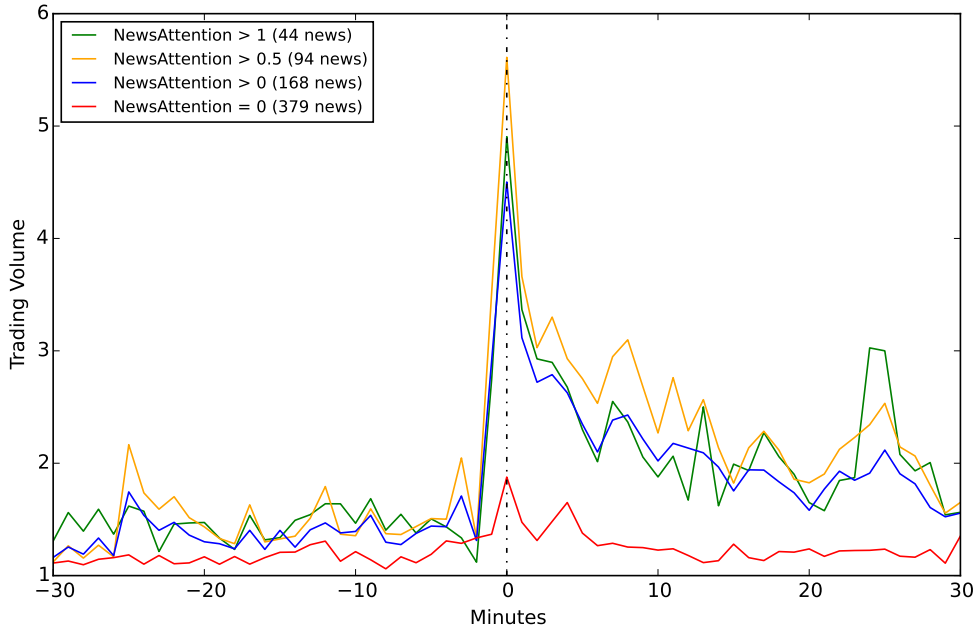
Step 2. We sort users in descending order of influence c'_j. We add the first 50 users to N_0, and we denote this list N_1.

Step 3. We select 60 users (i_1, i_2, ..., i_{60}) from N_1. For each user $j \notin N_1$, we compute a new variable of influence by defining $c'_j = \sum_{i=i_1}^{i_{60}} a_{i,j}$.

Step 4. We sort users in descending order of influence c_j. We add the first 50 users to N_1, and we denote this list N_2.

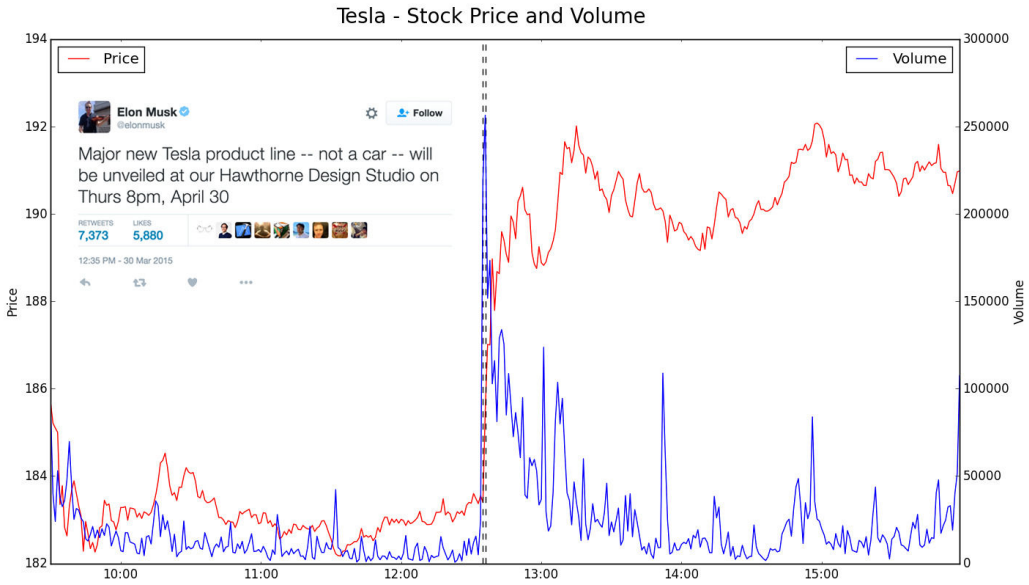
Step 5. We replicate step 3 and 4 until reaching network N_{60} composed of 3,010 users i (i_1, i_2,...i_{3010}).

2.8 Appendix B - High-frequency volume patterns around information for different threshold values of attention



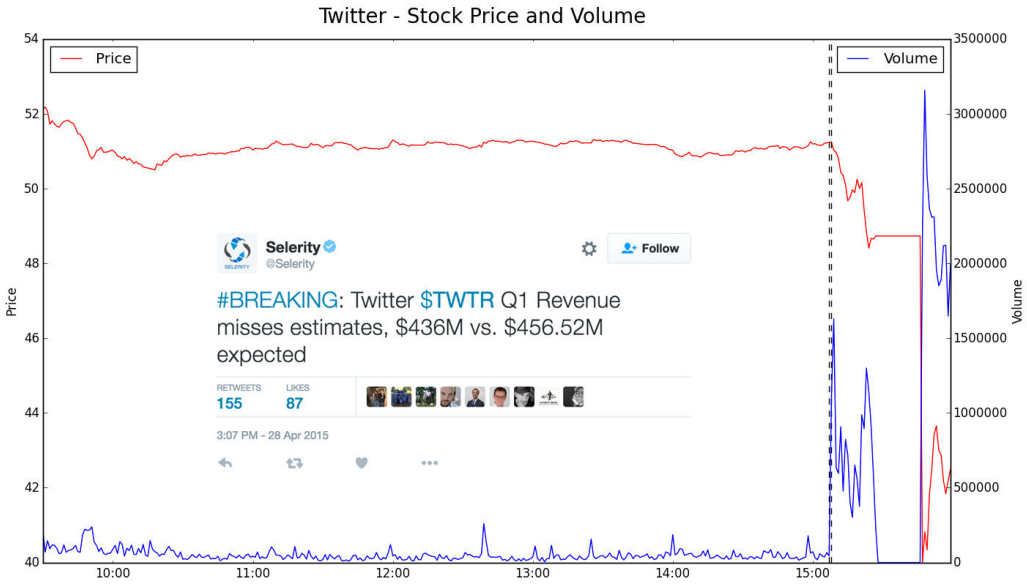
Notes: The figure presents the evolution of abnormal volume around the release of unscheduled HH with newswire-corrected timestamps. We consider HH when investors do not pay attention to news ($NewsAttention_i = 0$) and for various threshold values to define attention-grabbing-news ($NewsAttention_i > 0$; $NewsAttention_i > 0.5$; $NewsAttention_i > 1$).

FIGURE 2.1: Elon Musk tweet - Impact on Tesla stock price and trading volume



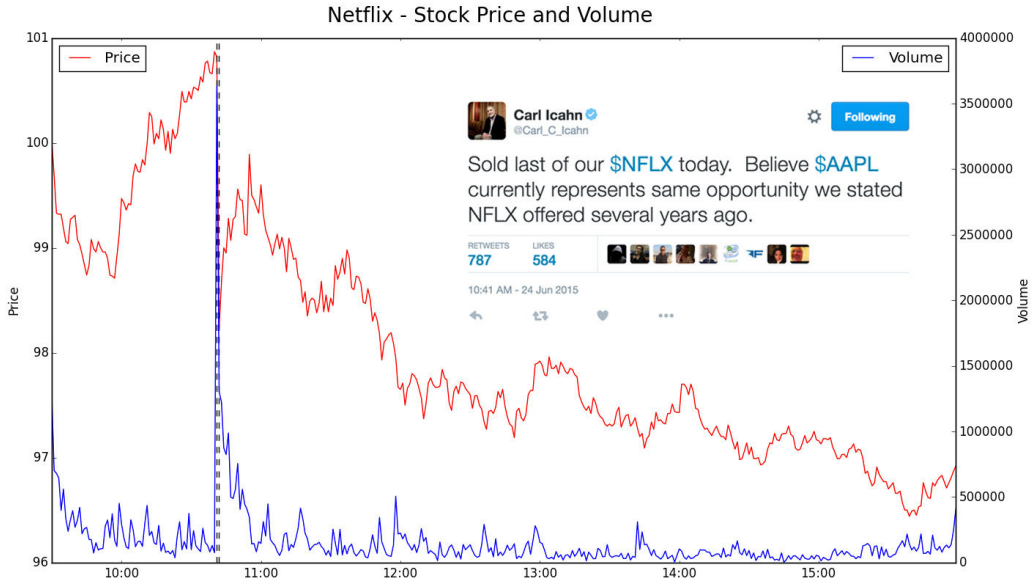
Notes: The figure illustrates the large increase in Tesla's stock price and trading volume following Elon Musk tweet announcing a new product line on March 30, 2015 at 12:35 p.m. For a complete story, read "Elon Musk tweet about new product line boosts Tesla shares" (MarketWatch, March 30, 2015).

FIGURE 2.2: Selerity tweet - Impact on Twitter stock price and trading volume



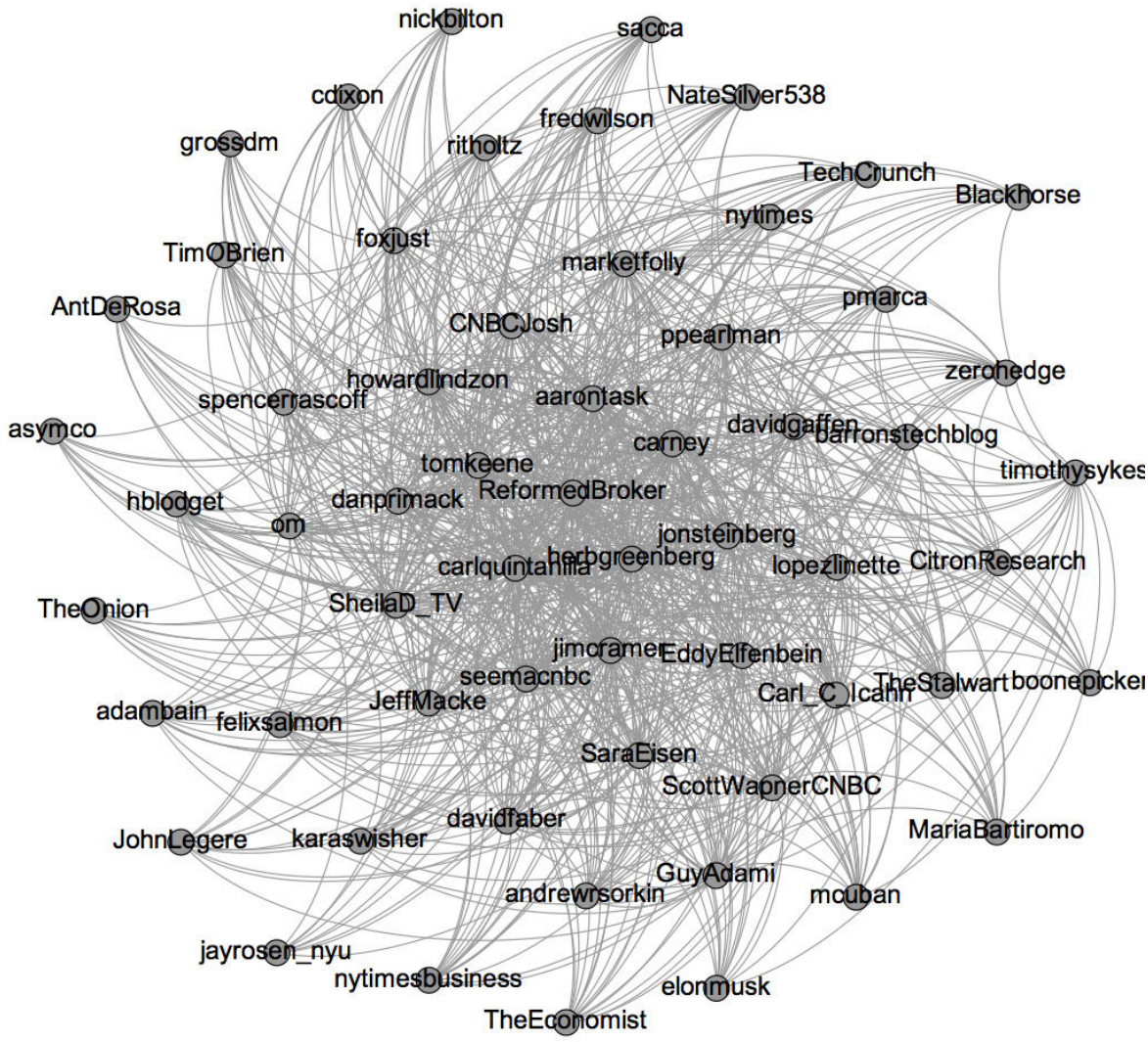
Notes: The figure shows the large decrease in Twitter's stock price and large increase in trading volume following Selerity tweet leaking Twitter earnings results on April 28, 2015 at 3:07 p.m. The leak prompted a NYSE trading halt for "news pending" starting at 3:27 p.m. For a complete story, read "The tweets that made Twitter stock crash" (MarketWatch, April 28, 2015).

FIGURE 2.3: Carl Icahn tweet - Impact on Netflix stock price and trading volume



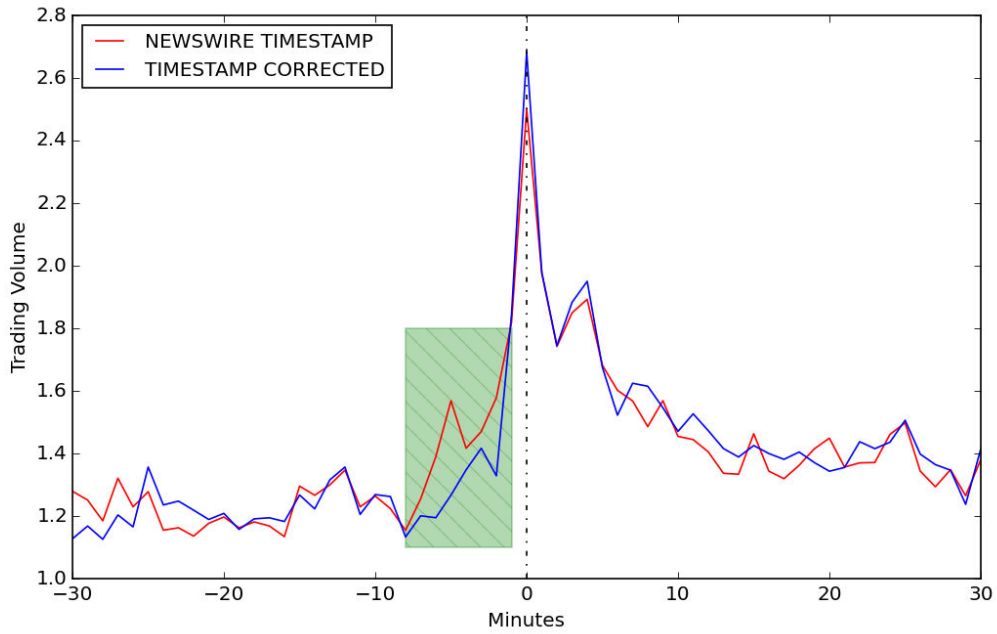
Notes: The figure illustrates the large decrease in Netflix stock price and large increase in trading volume following Carl Icahn tweet announcing that he sold his stake in Netflix on June 24, 2015 at 10:41 a.m. For a complete story, read "Carl Icahn sells his Netflix stock near record highs" (MarketWatch, June 24, 2015).

FIGURE 2.4: *Network N_1 - Twitter financial influencers*



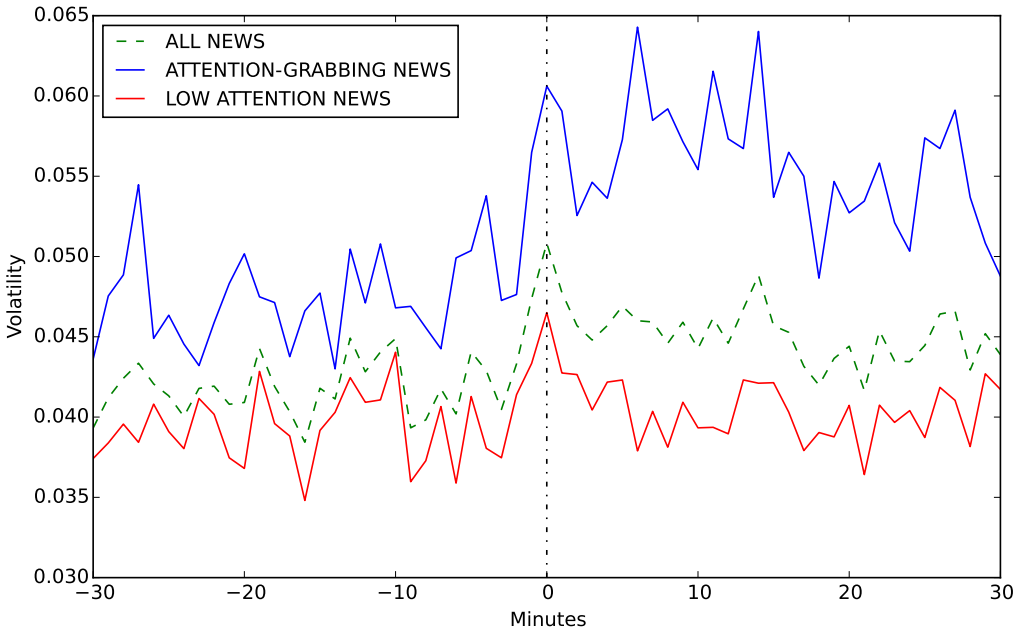
Notes: The figure shows the network structure N_1 (60 users). Each node represents a user and each link a directed friendships between two users. The graph is generated using Gephi, an open-source network analysis and visualization software.

FIGURE 2.5: High-frequency volume patterns with (without) newswire-corrected timestamps



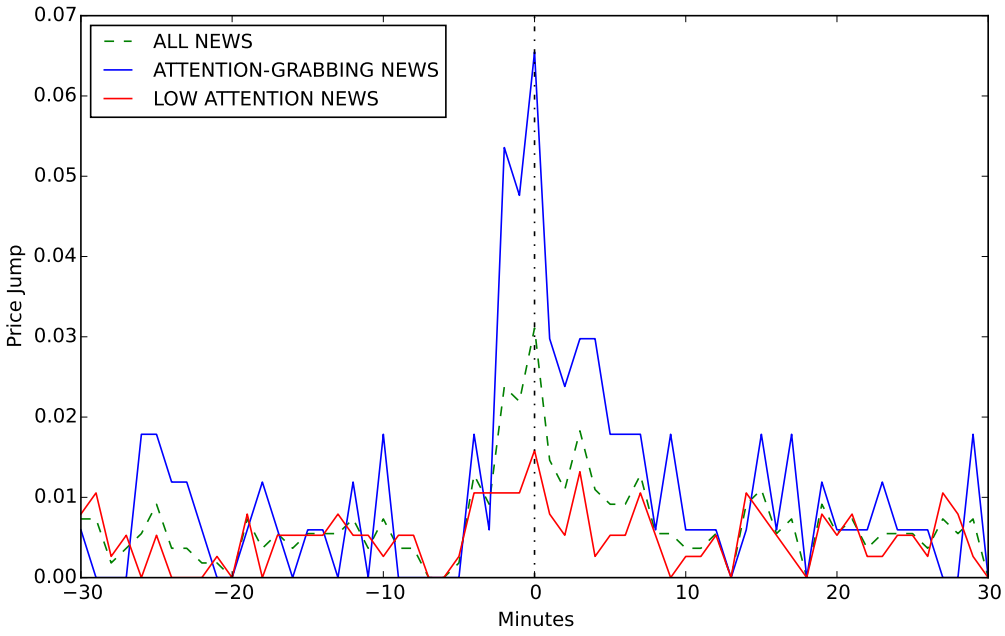
Notes: The figure presents the evolution of abnormal trading volume on a [-30:+30] minutes event-window around the release of unscheduled *HH* for "All news" (547 news stories) with and without corrected newswire timestamps. In green, we highlight the period during which abnormal trading volume is overestimated if timestamp delays are not corrected.

FIGURE 2.6: High-frequency volatility patterns around information and attention



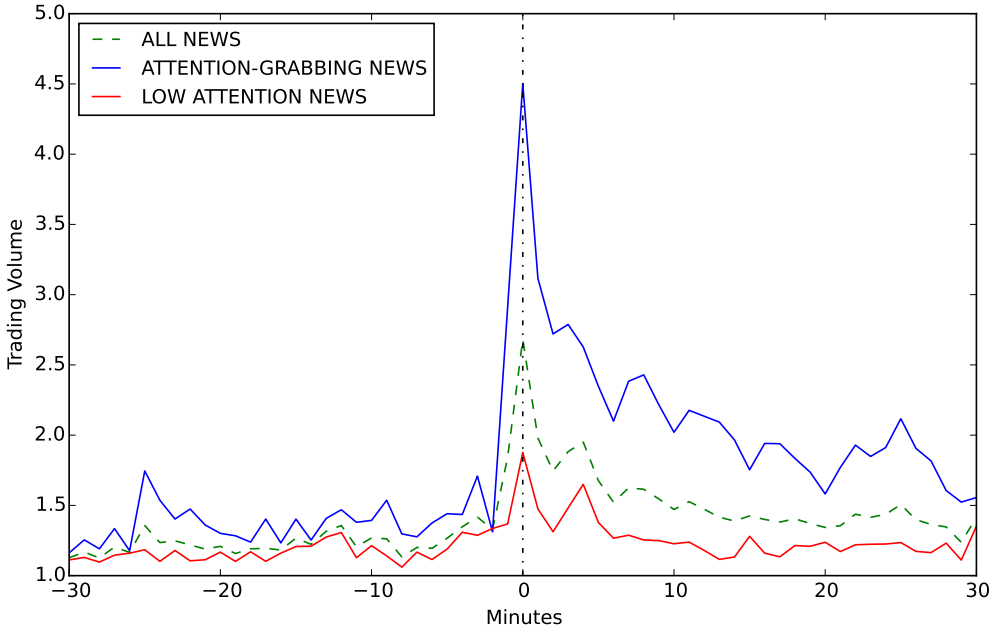
Notes: The figure presents the evolution of volatility on a [-30:+30] minutes event-window around the release of unscheduled *HH* with newswire-corrected timestamps, for "All news" (547 news stories), "Attention-grabbing news" ($NewsAttention_i > 0$; 168 news stories) and "Low-attention news" ($NewsAttention_i = 0$; 379 news stories).

FIGURE 2.7: High-frequency jump patterns around information and attention



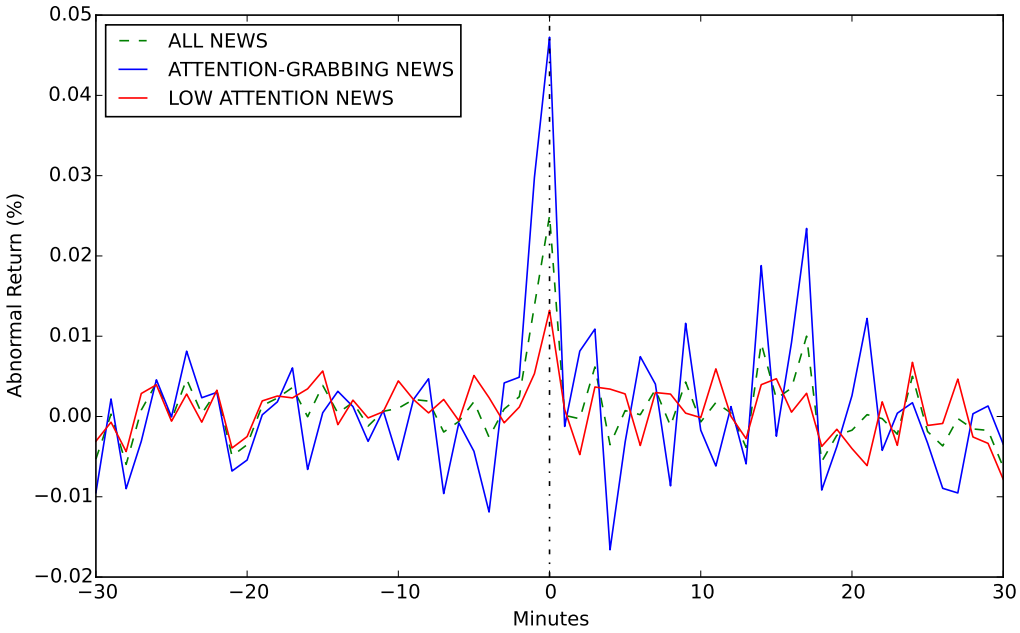
Notes: The figure presents the evolution of the average number of jumps on a [-30:+30] minutes event-window around the release of unscheduled HH with newswire-corrected timestamps, for "All news" (547 news stories), "Attention-grabbing news" ($NewsAttention_i > 0$; 168 news stories) and "Low-attention news" ($NewsAttention_i = 0$; 379 news stories).

FIGURE 2.8: High-frequency volume patterns around information and attention



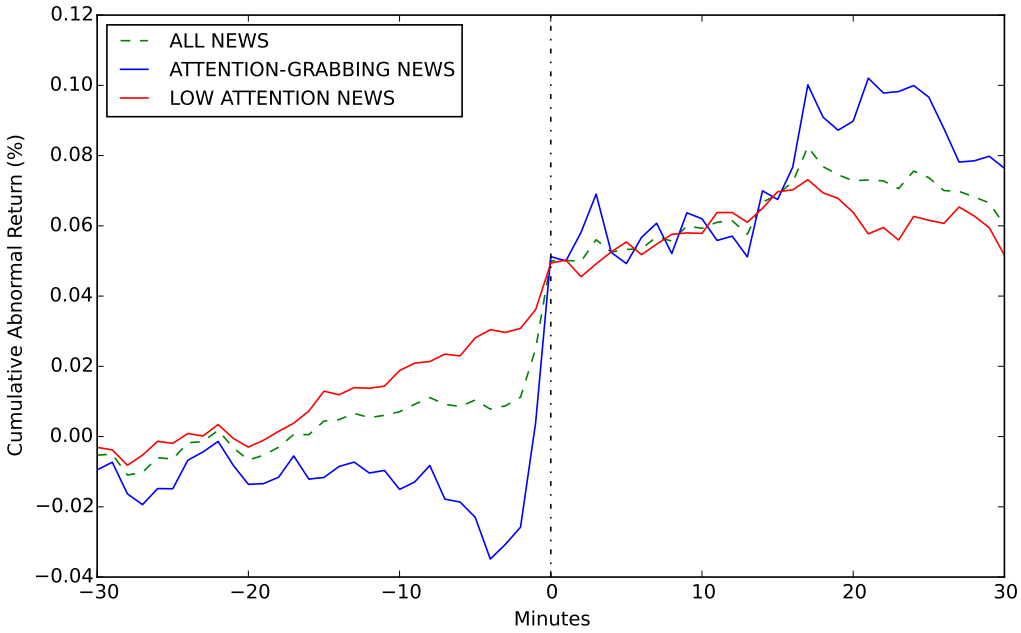
Notes: The figure presents the evolution of abnormal trading volume on a [-30:+30] minutes event-window around the release of unscheduled *HH* with newswire-corrected timestamps, for "All news" (547 news stories), "Attention-grabbing news" ($NewsAttention_i > 0$; 168 news stories) and "Low-attention news" ($NewsAttention_i = 0$; 379 news stories).

FIGURE 2.9: High-frequency abnormal returns patterns around information and attention



Notes: The figure presents the evolution of abnormal returns on a [-30:+30] minutes event-window around the release of positive (long) and negative (short) unscheduled *HH* with newswire-corrected timestamps, for "All news" (358 news stories), "Attention-grabbing news" ($NewsAttention_i > 0$; 124 news stories) and "Low-attention news" ($NewsAttention_i = 0$; 234 news stories). *HH* with a neutral sentiment are removed from the analysis.

FIGURE 2.10: High-frequency cumulative return patterns around information and attention



Notes: The figure presents the evolution of cumulative abnormal returns on a [-30:+30] minutes event-window around the release of positive (long) and negative (short) unscheduled *HH* with newswire-corrected timestamps, for "All news" (358 news stories), "Attention-grabbing news" ($NewsAttention_i > 0$; 124 news stories) and "Low-attention news" ($NewsAttention_i = 0$; 234 news stories). *HH* with a neutral sentiment are removed from the analysis.

TABLE 2.1: The initial list of 10 influential users in N_0

user	username	description	follower	friend	message
@carlquintanilla	Carl Quintanilla	Fin. journalist, CNBC	119,306	3,799	4,696
@Carl_C_Icahn	Carl Icahn	Activist investor	285,997	107	267
@herbgreenberg	Herb Greenberg	Fin. journalist, The Street	240,000	702	25,903
@howardlindzon	Howard Lindzon	Fin. analyst, fmr hedge fund manager	253,405	1,682	104,860
@jimcramer	Jim Cramer	TV journalist, fmr hedge fund manager	948,924	428	65,442
@MariaBartiromo	Maria Bartiromo	Fin. journalist, FOX	186,602	1,048	12,011
@om	Om Malik	Venture capitalist, entrepreneur	1,532,690	1,242	44,494
@pmarca	Marc Andreesen	Investor, entrepreneur	546,562	7,465	96,096
@ReformedBroker	Joshua Brown	Fin. advisor	143,361	3,296	102,210
@timothysykes	Timothy Sykes	Stock trader, penny-stock expert	122,404	6,408	63,864

Notes: The table reports descriptive statistics for the 10 *Twitterers* in N_0 (initial list). *"Follower"* represents the number of *Twitterers* who opt in to see tweets of each expert. *"Friend"* represents the number of *Twitterers* who have opted in to follow each expert. *"Message"* represents the total number of messages sent by each expert since the creation of its Twitter account.

TABLE 2.2: A sample of Twitter messages published on January 2, 2013

Date	Twitter	Content
2013-01-02 12:00:55	KeeneOnMarket	Reversals to the downside $AAPL $NFLX $CRM
2013-01-02 12:02:15	business	FLASH: Amazon wins dismissal of Apple s false advertising claim
2013-01-02 12:02:17	jyarow	AMAZON WINS DISMISSAL OF APPLE S FALSE ADVERTISING CLAIM over use of App Store.
2013-01-02 12:04:59	bespokeinvest	A number of triple digit priced growth stocks are already down 1%+ from the open. $AAPL $CMG $CRM $ISRG $LNKD $PCLN $PNRA
2013-01-02 12:06:08	sspencer_smb	pleased w/ how smb handled $AAPL today. focused on long to resist @555. then played reversal below 548. flexible
2013-01-02 12:06:13	bespokeinvest	$AAPL tested its 50-DMA at the open but is now at its lows for the day.
2013-01-02 12:10:26	SAI	Amazon Gets Apple s Lawsuit Over App Store Tossed Out $AMZN $AAPL by @jyarow http://t.co/Mrmv8R3l
2013-01-02 12:20:18	daveweigel	RT @keshasuxx: ANIMALS!! i m on the cover of @seventeenmag! check it amp get a copy on 1/8 [...]
2013-01-02 12:20:33	tlmontana	@bespokeinvest: $AAPL tested its 50-DMA at the open but is now at its lows for the day.
2013-01-02 12:20:35	Techmeme	Court dismisses Apple s case against Amazon for use of App Store name @markgurman / 9to5Mac [...]
2013-01-02 12:21:02	talkingbiznews	New York Times deputy investigations editor discussed Wal-Mart coverage: http://t.co/dLY53m39
2013-01-02 12:21:37	Reuters	Judge rejects Apple false advertising claims vs. Amazon http://t.co/f3O1DDky $AAPL $AMZN
2013-01-02 12:21:47	FastMoneyLydia	Cooperman: Sold down $AAPL position - didn t like the way it was acting: doesn t like cash policy. Prefers Qualcomm. @cnbcfastmoney
2013-01-02 12:23:13	MarketCurrents	Johnson amp Johnson JNJ declares $0.61/share quarterly dividend in line with previous. Forward yield ... http://t.co/k9DA4m5M $JNJ
2013-01-02 12:26:00	verge	Court rules Amazon s App Store isn t false advertising but full trademark lawsuit goes on http://t.co/8JvdLpdw
2013-01-02 12:27:25	arohan	@jtruman5 JJ tried but it requires your email address. Could you send me a request at http://t.co/Bas1X6zS
2013-01-02 12:27:42	murphyrosecliff	@FastMoneyLydia @cnbcfastmoney I m hoping those comments from Leon Cooperman re $AAPL cash policy on @cnbcfastmoney will reach [...]
2013-01-02 12:29:53	business	Amazon wins dismissal of Apple s false advertising claims http://t.co/VvD5shw2
2013-01-02 12:31:17	ReutersBiz	Judge rejects Apple false advertising claims vs. Amazon http://t.co/170gupcv
2013-01-02 12:33:31	RedDogT3	@TrueChartTrader @ep_capital no one is 100%. I try and be consistent. As far as $aapl. I have a lot of happy followers there
2013-01-02 12:35:13	TechCrunch	Court Rejects Apple s False Advertising Claim In App Store Trademark Lawsuit http://t.co/0WDYOUV8 by @sarahintampa
2013-01-02 12:37:39	sharkbiotech	just went through my trades from last year 56% winners avg gain $1.09 avg loss .88 20% of my gains were $AAPL related 1 down month April
2013-01-02 12:38:26	jonnajarian	fine with the call sold $AAPL $VMW amp $NFLX this am 4 a trade RT @Pete_Romano: Najarian just cant admit he was wrong to go all cash.
2013-01-02 12:39:21	Benzinga	Smartphone Preference: 38% of Gamers Choose Google s Android Over Apple s iOS: http://t.co/WkJQfr37 $AAPL $GOOG $YHOO
2013-01-02 12:40:20	business	Samsung loses bid to seal sales data in Apple dispute http://t.co/SzS5TLwh
2013-01-02 12:45:18	bizrpt	Judge rejects Apple false advertising claims vs. Amazon http://t.co/pdhJko6A @bizrpt $AAPL $AMZN
2013-01-02 12:46:49	savitz	Acacia Research In Settlement With Apple In Patent Case: The patent licensing firm Acacia Research this morning ... http://t.co/mavLxG0B
2013-01-02 12:53:02	Chris_Ciaccia	How many of you will care less about the iPhone 6 than you do Ray Lewis or Ed Reed? http://t.co/JzbtrAfr
2013-01-02 12:56:21	CNBC	U.S. judge granted Amazon s bid to end part of Apple lawsuit over Amazon s use of the term APP STORE - http://t.co/udu5fAWj $AMZN $AAPL
2013-01-02 12:59:04	CNBCTopStories	Judge Rejects Apple s False Advertising Claims Against Amazon http://t.co/QOC1dRTZ

Notes: The table provides all messages sent on Twitter by experts from N_{60} on January 2, 2013 between 12 p.m. and 1 p.m.

TABLE 2.3: Cosine similarity example between a Bloomberg news and tweet flow

Source	Time	Content	Similarity
BBG	11:25:12	EINHORN DROPS SUIT AGAINST APPLE OVER SHAREHOLDER VOTE	
TWT	11:21:40	Ex-Apple marketing guru Guy Kawasaki now advising Motorola http://t.co/NWc1tpoZ4F	0.0
TWT	11:21:50	I'd like to get up to about 400 contracts on the sell side of these $AAPL calls assuming she stays below $435 for rest of the day basically	0.0
TWT	11:22:54	BREAKING NEWS: Government sources confirm that horsemeat has been found in a significant percentage of $AAPL products.	0.0
TWT	11:23:32	See also http://t.co/FfDdSFj4pL MT @juliakmarsh: Thank you PO Walke and Sgt Whitley at Transit Dt 32 for finding my iPhone this morning!	0.0
TWT	11:23:36	And then of course if I had short 400 contracts I d be poppin molly and sweatin ... WOOOOOO! $AAPL	0.0
TWT	11:23:45	David Einhorn withdraws lawsuit against Apple. Manhattan federal court approves	0.14
TWT	11:24:05	Can Apple and Google Win This New Market? http://t.co/h0wEPStwIe	0.0
TWT	11:25:01	Risk/reward of $AAPL settlement at any time IMHO heavily favors $VHC and that doesn't even include that I think they win against $CSCO.	0.05
TWT	11:25:27	Einhorn Drops Suit Against Apple Over Shareholder Vote $AAPL	1.0

Notes: The table presents an example of cosine similarity between the Bloomberg *HH* "Einhorn drops suit against Apple over shareholder vote" and all messages published on Twitter about Apple on a [-15:0] minutes around the release of the news event. We mark the tweets related to Bloomberg *HH* in red.

TABLE 2.4: News release times for Twitter versus Bloomberg

Date	News Provider	Content
2013-08-13 14:22:26	Bloomberg	ICAHN HAS LARGE POSITION IN APPLE ICAHN SAYS ON TWITTER
2013-08-13 14:21:29	Carl_C_Icahn	We currently have a large position in APPLE. We believe the company to be extremely undervalued. Spoke to Tim Cook today [...]
2014-10-08 12:04:55	Bloomberg	APPLE SENDS OUT INVITATIONS FOR OCT 16 EVENT IN CUPERTINO WSJ
2014-10-08 12:01:56	geoffreyfowler	Invites out for Apple event in Cupertino on October 16. http://t.co/gQRSp7YsMY
2015-02-13 13:23:55	Bloomberg	APPLE SAID TO HIRE AUTO EXPERTS TO WORK IN RESEARCH LAB FT
2015-02-13 13:22:26	tim	About those Apple car rumours... Apple is hiring automotive experts to work in a secret research lab @FT sources say [...]
2013-07-12 14:28:36	Bloomberg	GE SAYS ENGINES NOT TIED TO BOEING FIRE AT HEATHROW REUTERS
2013-07-12 14:24:41	firstadopter	It s NOT US!!! *GEsays*BA fire doesn t involve engines
2015-04-20 12:04:35	Bloomberg	GE SAID TO BE IN TALKS TO SELL COMMERCIAL LENDING BUSINESS DJ
2015-04-20 12:02:26	d4ytrad3	$GE in discussion to sell commercial lending biz
2015-05-06 15:30:08	Bloomberg	BLACKSTONE SAID TO BE AMONG BIDDERS FOR GE LENDING UNIT FT
2015-05-06 15:29:01	ftfinancenews	Blackstone joins race for GE lending unit http://t.co/1p2vviyuzH
2014-07-18 10:13:54	Bloomberg	IBM CUT TO SELL VS HOLD AT SOCIETE GENERALE EARLIER
2014-07-18 09:59:48	zozotrader	IBM cut at SocGen
2014-11-24 14:16:29	Bloomberg	ICAHN HAS ABSOLUTELY NO INVOLVEMENT IN IBM CNBC S WAPNER
2014-11-24 14:13:43	ScottWapnerCNBC	Sources tell me @Carl_C_Icahn has absolutely no involvement in $IBM. Stock had moved earlier on rumor that he did.
2015-10-27 13:51:24	Bloomberg	IBM LEARNED IN AUG SEC CONDUCTING PROBE ON REVENUE RECOGNITION
2015-10-27 13:47:24	Livesquawk	IBM says that the SEC is investigating the company in relation to revenue recognition - Rtrs $IBM
2014-01-15 12:53:34	Bloomberg	CARLYLE PAYING 4B 4 2B FOR J J BLOOD TESTING UNIT WSJ
2014-01-15 12:40:36	MikeSpectorWSJ	@OneCarlyle paying between $4 billion and $4.2 billion for J amp J blood-testing unit. @WSJ scoop: http://t.co/Uho1t5AUC6
2014-03-20 10:21:00	Bloomberg	JNJ 1.2B JUDGMENT OVERTURNED BY ARKANSAS SUPREME COURT AP
2014-03-20 10:18:50	YahooNews	Arkansas Supreme Court overturns $1.2B judgment against Johnson & Johnson over drug marketing @AP
2015-04-23 11:00:12	Bloomberg	J&J BOOSTS QTRLY DIV TO 75C SHR FROM 70C EST 74C
2015-04-23 10:59:43	OpenOutcrier	$JNJ Increasing div
2013-03-08 10:25:30	Bloomberg	LESLIE DACH OF WAL MART TO LEAVE COMPANY POLITICO REPORTS
2013-03-08 10:18:30	mikeallen	Wal-Mart CEO Mike Duke tells associates: Leslie Dach executive vice president corporate affairs leaving in June after 7 yrs [...]
2013-10-15 12:15:23	Bloomberg	WAL MART TO CLOSE UNDER PERFORMING STORES IN BRAZIL AND CHINA
2013-10-15 12:15:19	sbanjo	Walmart closing stores in Brazil China that aren't profitable revises down square footage from 20-22 million to 14 million
2015-12-11 12:48:08	Bloomberg	Walmart COM BEGINS SELLING APPLE WATCH TECHCRUNCH
2015-12-11 12:46:09	TechCrunch	Walmart com Begins Selling The Apple Watch https://t.co/KLZ3gbnuep by @sarahintampa

Notes: The table presents 15 (selected) cases where Twitter effectively "breaks the news". For each news, the first line represents Bloomberg reported timestamp with the associated Bloomberg headline. The second line presents the first mention of the news on Twitter, the user who "breaks the news" and the tweet content.

TABLE 2.5: Event-study results without (with) timestamp correction

	Volatility		Jump		Volume		Return (%)	
TIME	BLOOMBERG TIME	CORRECTED TIME	BLOOMBERG TIME	CORRECTED TIME	BLOOMBERG TIME	CORRECTED TIME	BLOOMBERG TIME	CORRECTED TIME
[−30 : −26]	0.417	0.417	0.006	0.005	1.253	1.158	-0.002*	-0.001
[−29 : −25]	0.421	0.421	0.006	0.005	1.253	1.204	-0.001	-0.000
[−28 : −24]	0.420	0.418	0.005	0.005	1.233	1.217	0.000	0.001
[−27 : −23]	0.419	0.417	0.004	0.005	1.229	1.242	0.001	0.002
[−26 : −22]	0.417	0.414	0.004	0.005	1.192	1.245	0.001	0.002**
[−25 : −21]	0.416	0.412	0.003	0.004	1.182	1.249	0.000	0.001
[−24 : −20]	0.409	0.411	0.002	0.002	1.165	1.220	0.000	-0.000
[−23 : −19]	0.417	0.419	0.003	0.003	1.167	1.204	-0.000	-0.001
[−22 : −18]	0.421	0.420	0.003	0.003	1.171	1.193	0.001	-0.000
[−21 : −17]	0.419	0.416	0.003	0.004	1.177	1.188	0.000	-0.000
[−20 : −16]	0.415	0.412	0.003	0.004	1.169	1.187	0.001	0.001
[−19 : −15]	0.418	0.413	0.004	0.005	1.188	1.198	0.002**	0.002**
[−18 : −14]	0.410	0.407	0.004	0.005	1.209	1.212	0.002	0.002
[−17 : −13]	0.412	0.413	0.004	0.005	1.233	1.236	0.002	0.002
[−16 : −12]	0.417	0.418	0.006	0.005	1.268	1.269	0.002	0.001
[−15 : −11]	0.428	0.429	0.006	0.005	1.288	1.273	0.001	0.001
[−14 : −10]	0.435	0.436	0.005	0.006	1.281	1.274	0.001	0.001
[−13 : −9]	0.429	0.432	0.005	0.005	1.273	1.282	0.002	0.001
[−12 : −8]	0.422	0.422	0.005	0.005	1.244	1.245	0.002	0.001
[−11 : −7]	0.421	0.420	0.004	0.004	1.225	1.214	0.002	0.001
[−10 : −6]	0.414	0.412	0.004	0.003	1.257	1.212	0.002	0.001
[−9 : −5]	0.415	0.410	0.004	0.002	1.318	1.211	0.001	0.001
[−8 : −4]	0.423	0.417	0.007	0.004	1.357	1.228	0.001	-0.000
[−7 : −3]	0.424	0.419	0.009*	0.005	1.420	1.285	0.001	-0.000
[−6 : −2]	0.431	0.422	0.012***	0.010**	1.485	1.311	0.002	0.000
[−5 : −1]	0.444	0.436	0.016***	0.014***	1.570*	1.440	0.004	0.003
[−4 : 0]	0.457*	0.450	0.020***	0.020***	1.757**	1.724**	0.008***	0.008***
[−3 : 1]	0.466**	0.460*	0.020***	0.020***	1.869***	1.850**	0.007**	0.008***
[−2 : 2]	0.470**	0.470**	0.019***	0.020***	1.923***	1.916**	0.006***	0.008***
[−1 : 3]	0.467**	0.473**	0.018***	0.019***	1.977***	2.027***	0.006***	0.009***
[0 : 4]	0.469**	0.470**	0.016***	0.017***	1.828***	1.847***	0.004	0.006**
[1 : 5]	0.463**	0.462**	0.012***	0.013***	1.754**	1.755***	-0.000	0.001
[2 : 6]	0.457**	0.458*	0.010**	0.012***	1.719**	1.731**	0.000	0.001
[3 : 7]	0.462**	0.459**	0.010**	0.012***	1.646**	1.678**	0.001	0.001*
[4 : 8]	0.459*	0.458*	0.008	0.010**	1.581**	1.597**	0.002	-0.000
[5 : 9]	0.459**	0.459*	0.007	0.008	1.536**	1.556**	0.001	0.001
[6 : 10]	0.453*	0.453*	0.006	0.007	1.504*	1.556*	0.000	0.001
[7 : 11]	0.454*	0.454*	0.005	0.006	1.472*	1.526*	0.001*	0.002
[8 : 12]	0.451	0.451	0.004	0.005	1.442*	1.486*	0.001*	0.001
[9 : 13]	0.460**	0.455*	0.002	0.004	1.395	1.455*	0.001	0.000
[10 : 14]	0.457*	0.461**	0.005	0.004	1.396*	1.446*	0.002**	0.001
[11 : 15]	0.462**	0.464**	0.006	0.006	1.376*	1.420*	0.004**	0.002**
[12 : 16]	0.466**	0.462**	0.007	0.006	1.359*	1.402*	0.003*	0.002**
[13 : 17]	0.458**	0.459**	0.007	0.007	1.364*	1.400*	0.003	0.004**
[14 : 18]	0.450*	0.450*	0.007	0.007	1.380*	1.396*	0.003	0.004*
[15 : 19]	0.443	0.440	0.006	0.007	1.378*	1.380*	0.002	0.002*
[16 : 20]	0.439	0.437	0.005	0.005	1.380*	1.371*	-0.001	0.001
[17 : 21]	0.430	0.430	0.006	0.006	1.390*	1.382*	-0.001	0.000
[18 : 22]	0.438	0.434	0.006	0.005	1.392*	1.384*	-0.001	-0.002
[19 : 23]	0.439	0.437	0.007	0.006	1.401*	1.397*	-0.001	-0.001
[20 : 24]	0.441	0.437	0.005	0.005	1.411*	1.430*	-0.001	0.000
[21 : 25]	0.435	0.437	0.005	0.005	1.409*	1.439*	0.000	0.000
[22 : 26]	0.440	0.446	0.004	0.005	1.393*	1.424*	-0.000	-0.001
[23 : 27]	0.443	0.449*	0.005	0.005	1.389*	1.410**	-0.000	-0.001
[24 : 28]	0.445*	0.448*	0.005	0.005	1.350*	1.370*	0.000	-0.000
[25 : 29]	0.443*	0.451*	0.005	0.006	1.326*	1.352*	-0.001	-0.002
[26 : 30]	0.446*	0.450**	0.004	0.005	1.324*	1.322*	-0.001	-0.003
Event	547	547	547	547	547	547	547	547

Notes: The table reports the significance of volatility, jump, abnormal volume, and abnormal return for each 5 minutes interval during a [-30:+30] event window around the release of unscheduled news announcements. We compare results when considering Bloomberg reported timestamp as the event minute ("Bloomberg Time") and when considering the first mention of the news on Twitter (when social media "breaks the news") and Bloomberg reported timestamp otherwise ("Corrected Time"). *, ** and *** denote significance respectively at the 10% level, 5% level and 1% level. Significance is assessed using non-parametric Corrado rank test.

TABLE 2.6: Event-study results with high- versus low-attention (1-minute intervals)

	Volatility		Jump		Volume		Return (%)	
	HIGH ATTENTION	LOW ATTENTION	HIGH ATTENTION	LOW ATTENTION	HIGH ATTENTION	LOW ATTENTION	HIGH ATTENTION	LOW ATTENTION
-30	0.044	0.037	0.006	0.008	1.162	1.112	-0.009	-0.003
-29	0.048	0.038	0.000	0.011	1.255	1.130	0.002	-0.001
-28	0.049	0.040	0.000	0.003	1.191	1.097	-0.009	-0.004
-27	0.054	0.038	0.000	0.005	1.335	1.145	-0.003	0.003
-26	0.045	0.041	0.018	0.000	1.178	1.160	0.005	0.004
-25	0.046	0.039	0.018	0.005	1.745	1.184	-0.000	-0.001
-24	0.045	0.038	0.012	0.000	1.537	1.102	0.008	0.003
-23	0.043	0.041	0.012	0.000	1.403	1.179	0.002	-0.001
-22	0.046	0.040	0.006	0.000	1.474	1.105	0.003	0.003
-21	0.048	0.037	0.000	0.003	1.360	1.113	-0.007	-0.004
-20	0.050	0.037	0.000	0.000	1.301	1.167	-0.005	-0.003
-19	0.047	0.043*	0.006	0.008	1.284	1.101	0.000	0.002
-18	0.047	0.040	0.012	0.000	1.238	1.170	0.002	0.003
-17	0.044	0.039	0.006	0.005	1.403	1.102	0.006*	0.002
-16	0.047	0.035	0.000	0.005	1.233	1.160	-0.007	0.003
-15	0.048	0.039	0.006	0.005	1.402	1.207	0.000	0.006
-14	0.043	0.040	0.006	0.005	1.253	1.209	0.003	-0.001
-13	0.050	0.042	0.000	0.008	1.407*	1.275	0.001	0.002
-12	0.047	0.041	0.012	0.005	1.469	1.307*	-0.003	-0.000
-11	0.051	0.041	0.000	0.005	1.380	1.128	0.001	0.001
-10	0.047	0.044	0.018	0.003	1.393	1.213	-0.005	0.004
-9	0.047	0.036	0.000	0.005	1.537*	1.141	0.002	0.002
-8	0.046	0.037	0.000	0.005	1.297	1.060	0.005	0.000
-7	0.044	0.041	0.000	0.000	1.276	1.167	-0.010	0.002
-6	0.050	0.036	0.000	0.000	1.375*	1.115	-0.001	-0.001
-5	0.050	0.041	0.000	0.003	1.441	1.189	-0.004	0.005*
-4	0.054	0.038	0.018	0.011	1.435*	1.309	-0.012	0.002
-3	0.047	0.037	0.006	0.011	1.709	1.287	0.004	-0.001
-2	0.048	0.041	0.054***	0.011	1.313	1.335	0.005	0.001
-1	0.056	0.043	0.048***	0.011	2.914**	1.368	0.030*	0.005*
0	0.061	0.047	0.065***	0.016**	4.502***	1.877***	0.047***	0.013**
1	0.059	0.043	0.030***	0.008	3.116***	1.475*	-0.001	0.001
2	0.053	0.043	0.024*	0.005	2.720***	1.312	0.008	-0.005
3	0.055	0.040	0.030***	0.013*	2.788***	1.482*	0.011	0.004
4	0.054	0.042	0.030***	0.003	2.627***	1.650	-0.017	0.003
5	0.057	0.042	0.018	0.005	2.348***	1.378**	-0.003	0.003
6	0.064	0.038	0.018	0.005	2.100***	1.266	0.007	-0.004
7	0.058**	0.040	0.018	0.011	2.384**	1.288	0.004	0.003
8	0.059	0.038	0.006	0.005	2.429**	1.254*	-0.009	0.003
9	0.057	0.041	0.018	0.000	2.216**	1.249	0.012	0.000
10	0.055	0.039	0.006	0.003	2.020**	1.227	-0.002	-0.000
11	0.062*	0.039	0.006	0.003	2.176*	1.239	-0.006	0.006
12	0.057	0.039	0.006	0.005	2.135*	1.179	0.001	0.000
13	0.057*	0.042	0.000	0.000	2.092*	1.116	-0.006	-0.003
14	0.064*	0.042	0.006	0.011	1.966**	1.133	0.019	0.004
15	0.054	0.042	0.018	0.008	1.753	1.280**	-0.002	0.005
16	0.056*	0.040	0.006	0.005	1.941**	1.160	0.009*	0.001
17	0.055	0.038	0.018	0.003	1.939*	1.134	0.023	0.003
18	0.049	0.039	0.000	0.000	1.834**	1.214	-0.009	-0.004
19	0.055	0.039	0.012	0.008	1.737*	1.209	-0.004	-0.002
20	0.053	0.041	0.006	0.005	1.581*	1.238*	0.003	-0.004*
21	0.053	0.036	0.006	0.008	1.770**	1.171	0.012**	-0.006
22	0.056	0.041	0.006	0.003	1.929**	1.220	-0.004	0.002
23	0.052	0.040	0.012	0.003	1.848**	1.223	0.000	-0.004
24	0.050	0.040	0.006	0.005	1.912	1.225	0.002	0.007
25	0.057	0.039	0.006	0.005	2.117**	1.235	-0.003	-0.001
26	0.057	0.042	0.006	0.003	1.906***	1.173*	-0.009	-0.001
27	0.059*	0.041	0.000	0.011	1.817**	1.164	-0.010	0.005
28	0.054	0.038	0.000	0.008	1.606*	1.231	0.000	-0.003
29	0.051	0.043	0.018	0.003	1.523	1.111	0.001	-0.003
30	0.049	0.042	0.000	0.000	1.556	1.352*	-0.003	-0.008
Event	168	379	168	379	168	379	124	368

Notes: The table reports the significance of volatility, jump, abnormal volume, and abnormal return for each minute during a [-30:+30] event window around the release of high-attention and low-attention news events. We consider as minute 0 the newswire corrected timestamp. *, ** and *** denote significance respectively at the 10% level, 5% level and 1% level. Significance is assessed using non-parametric Corrado rank test.

TABLE 2.7: Event-study results with high- versus low-attention (5-minute intervals)

	Volatility		Jump		Volume		Return (%)	
	HIGH ATTENTION	LOW ATTENTION	HIGH ATTENTION	LOW ATTENTION	HIGH ATTENTION	LOW ATTENTION	HIGH ATTENTION	LOW ATTENTION
[−30 : −26]	0.048	0.039	0.005	0.005	1.224	1.129	-0.003	-0.000
[−29 : −25]	0.048	0.039	0.007	0.005	1.341	1.143	-0.001	0.000
[−28 : −24]	0.048	0.039	0.010	0.003	1.397	1.138	0.000	0.001
[−27 : −23]	0.047	0.040	0.012	0.002	1.439	1.154	0.002	0.002
[−26 : −22]	0.045	0.040	0.013	0.001	1.467	1.146	0.004	0.002
[−25 : −21]	0.046	0.039	0.010	0.002	1.504	1.137	0.001	0.000
[−24 : −20]	0.046	0.039	0.006	0.001	1.415	1.133	0.000	-0.000
[−23 : −19]	0.047	0.040	0.005	0.002	1.364	1.133	-0.001	-0.000
[−22 : −18]	0.048	0.039	0.005	0.002	1.331	1.131	-0.001	0.000
[−21 : −17]	0.047	0.039	0.005	0.003	1.317	1.131	-0.001	0.000
[−20 : −16]	0.047	0.039	0.005	0.004	1.292	1.140	-0.001	0.002
[−19 : −15]	0.047	0.039	0.006	0.005	1.312	1.148	0.000	0.003
[−18 : −14]	0.046	0.039	0.006	0.004	1.306	1.170	0.001	0.003
[−17 : −13]	0.046	0.039	0.004	0.006	1.340	1.191	0.001	0.002
[−16 : −12]	0.047	0.040	0.005	0.006	1.353	1.232	-0.001	0.002
[−15 : −11]	0.048	0.041	0.005	0.006	1.382	1.225	0.000	0.001
[−14 : −10]	0.048	0.042	0.007	0.005	1.380	1.226	-0.001	0.001
[−13 : −9]	0.048	0.041	0.006	0.005	1.437*	1.213	-0.001	0.002
[−12 : −8]	0.047	0.040	0.006	0.005	1.415	1.170	-0.000	0.001
[−11 : −7]	0.047	0.040	0.004	0.004	1.377	1.142	-0.001	0.002
[−10 : −6]	0.047	0.039	0.004	0.003	1.376	1.139	-0.002	0.002
[−9 : −5]	0.047	0.038	0.000	0.003	1.385	1.134	-0.002	0.002
[−8 : −4]	0.049	0.039	0.004	0.004	1.365	1.168	-0.004	0.002
[−7 : −3]	0.049	0.039	0.005	0.005	1.447*	1.213	-0.004	0.002
[−6 : −2]	0.050	0.039	0.015*	0.007	1.454*	1.247	-0.002	0.001
[−5 : −1]	0.051	0.040	0.025***	0.009*	1.762*	1.298	0.005	0.003
[−4 : 0]	0.053	0.041	0.038***	0.012***	2.374**	1.435*	0.015*	0.004
[−3 : 1]	0.054	0.042	0.040***	0.011**	2.711**	1.468**	0.017**	0.004
[−2 : 2]	0.055	0.043*	0.044***	0.010**	2.913***	1.473**	0.018***	0.003
[−1 : 3]	0.057	0.043*	0.039***	0.011**	3.208***	1.503**	0.019***	0.004*
[0 : 4]	0.055*	0.042	0.036***	0.009*	3.151***	1.559**	-0.000	0.001
[1 : 5]	0.056*	0.041	0.026***	0.007	2.720***	1.460**	0.001	0.000
[2 : 6]	0.058**	0.041	0.024***	0.006	2.517***	1.418*	0.001	0.002
[3 : 7]	0.059**	0.040	0.023***	0.007	2.449***	1.413*	-0.003	0.002
[4 : 8]	0.059**	0.040	0.018*	0.006	2.377***	1.367*	0.002	0.001
[5 : 9]	0.059*	0.039	0.015*	0.005	2.295***	1.287*	0.003	0.000
[6 : 10]	0.058**	0.040	0.013	0.005	2.230***	1.257	-0.000	0.002
[7 : 11]	0.058*	0.039	0.011	0.004	2.245**	1.251	-0.001	0.002
[8 : 12]	0.058**	0.040	0.008	0.003	2.195**	1.230	-0.000	0.001
[9 : 13]	0.059**	0.040	0.007	0.002	2.128**	1.202	0.001*	0.001
[10 : 14]	0.059**	0.041	0.005	0.004	2.078**	1.179	0.001	0.002
[11 : 15]	0.058**	0.041	0.007	0.005	2.025**	1.189	0.004**	0.001
[12 : 16]	0.057**	0.041	0.007	0.006	1.978**	1.173	0.009*	0.002
[13 : 17]	0.056*	0.040	0.010	0.005	1.938**	1.164	0.008*	0.002
[14 : 18]	0.054*	0.040	0.010	0.005	1.887**	1.184	0.003	0.001
[15 : 19]	0.054*	0.039	0.011	0.005	1.841**	1.199*	0.004	-0.001
[16 : 20]	0.053	0.039	0.008	0.004	1.807**	1.191	0.005	-0.003
[17 : 21]	0.053	0.039	0.008	0.005	1.772**	1.193	-0.000	-0.003
[18 : 22]	0.054*	0.039	0.006	0.005	1.770**	1.210	0.001	-0.003
[19 : 23]	0.053	0.040	0.008	0.005	1.773**	1.212	0.003	-0.001
[20 : 24]	0.054*	0.039	0.007	0.005	1.808**	1.215	0.001	-0.000
[21 : 25]	0.054*	0.040	0.007	0.005	1.915**	1.215	-0.003	0.001
[22 : 26]	0.055*	0.040	0.007	0.004	1.942**	1.215	-0.004	0.001
[23 : 27]	0.055*	0.040	0.006	0.005	1.920**	1.204	-0.004	0.001
[24 : 28]	0.056*	0.040	0.004	0.006	1.871**	1.206	-0.004	-0.001
[25 : 29]	0.054*	0.041	0.006	0.006	1.794**	1.183	-0.004	-0.002
[26 : 30]	0.446*	0.450**	0.005	0.005	1.682**	1.206	-0.001	-0.003
Event	168	379	168	379	168	379	124	368

Notes: The table reports the significance of volatility, jump, abnormal volume, and abnormal return for each 5-minute interval during a [-30:+30] event window around the release of high-attention and low-attention news events. We consider as minute 0 the newswire corrected timestamp. *, ** and *** denote significance respectively at the 10% level, 5% level and 1% level. Significance is assessed using non-parametric Corrado rank test.

Chapter 3

Market manipulation and suspicious stock recommendations on social media

Abstract

Social media can help investors gather and share information about stock markets. However, it also presents opportunities for fraudsters to spread false or misleading statements in the marketplace. Analyzing millions of messages sent on the social media platform Twitter about small capitalization firms, we find that an abnormally high message activity on social media is associated with a large price increase on the event day and followed by a sharp price reversal over the next week. Our findings are consistent with the patterns of a pump-and-dump scheme, where fraudsters use social media to temporarily inflate the price of small capitalization stocks. To differentiate between the effects of overoptimism by noise traders and the illegal gains of a pump-and-dump scheme, we investigate social interactions between Twitter users through the use of network theory. We identify several clusters of users with suspicious online activity (stock promoters, fake accounts, automatic postings), favoring the manipulation/promotion hypothesis over the behavioral hypothesis.

Keywords: Asset pricing, Market manipulation, Fraud detection, Network analysis, Social media

JEL classification: C18, D80, G12, G14.

"*John, one thing I can promise you, even in this market, is that I never ask my clients to judge me on my winners. I ask them to judge me on my losers because I have so few. And in the case of Aerotyne, based on every technical factor out there, John, we are looking at a grand slam home run.*"

The Wolf of Wall Street, Dir. Martin Scorcese. Paramount Pictures, 2013. Movie.

3.1 Introduction

Market manipulation is as old as trading on organized exchanges (Putniņš, 2012). However, despite their long prevalance and considerable academic research on the topic, our understanding of the phenomenon is far from adequate. While theoretical models have been developed to address trade-based manipulation (Allen & Gale, 1992) or information-based manipulation (Bommel, 2003), empirical studies continue to be very scarce. This paper contributes to the emerging empirical literature on market manipulation by focusing on a specific type of illegal price manipulation: pump-and-dump schemes.

Pump-and-dump schemes involve touting a company's stock through false or misleading statements in the marketplace in order to artificially inflate (pump) the price of a stock. Once fraudsters stop hyping the stock and sell their shares (dump), the price typically falls. Although pump-and-dump schemes have existed for many decades, the emergence of the Internet and social media has provided a fertile new ground for fraudsters. False or misleading information can now be disseminated to a large number of potential investors with minimum effort, anonymously, and at a relatively low cost.[1] According to the Security and Exchange Commission (SEC)[2], "investors who learn of investing opportunities from social media should always be on the lookout for fraud."

To gain a better understanding of pump-and-dump schemes, we first focus on reported manipulation cases by analyzing all SEC litigation releases published between 1996 and 2015. We construct a database of pump-and-dump frauds, extending previous findings from Aggarwal & Wu (2006). We find that pump-and-dump schemes mainly target small capitalization stocks with low liquidity, also known as "micro-cap" or "penny stocks", traded in the Over-The-Counter (OTC) market. Market manipulators involved in such schemes often combine a false or misleading press release with a touting of the stock on spam e-mails, websites, bulletin boards, and fax blast. In two cases, fraudsters specifically use Twitter to manipulate stock prices.

While empirical proofs of market manipulation on small capitalization stocks have been identified using data from stock spam (e-mails) recommendations (Böhme & Holz, 2006; Frieder & Zittrain,

[1] "Investor Alert: Social Media and Investing - Avoiding Fraud" - Security and Exchange Commission, January 2012

[2] "Updated Investor Alert: Social Media and Investing - Stock Rumors" - Security and Exchange Commission, November 2015

2007; Hanke & Hauser, 2008; Nelson et al., 2013) and messages boards (Sabherwal et al., 2011), pump-and-dump schemes on social media have, to the best of our knowledge, never been empirically studied. We extend the literature on indirect empirical evidence of market manipulation by analyzing data from one of the largest worldwide social media platforms: Twitter. Analyzing data from Twitter could provide new insights as the interactions between users are directly observable. This feature allows researchers to cluster users based on common characteristic and identify those with suspicious behaviors. Further, as data are collected in real time, the analysis is not affected by the survivorship bias that occurs when data are collected "ex-post". As with any illegal activity, we should expect market manipulators to delete their messages or accounts after committing a fraud in order to decrease the probability of being caught by the Security and Exchange Commission. However, collecting data in real time will minimize this problem.

We conduct event studies to analyze the impact of a spike in posting activity on Twitter on the returns of small capitalization stocks. We find that an abnormally high message activity on social media about a company is associated with a large price increase on the event day, followed by a sharp price reversal over the next five days. This price reversal pattern is consistent with a pump-and-dump scheme (manipulation hypothesis) but it could also simply be caused by overoptimistic noise traders (behavioral hypothesis). While judicial inquiries would be needed to assess precisely if a large increase/decrease in stock prices is caused by fraudsters or by irrational unsophisticated traders, we investigate social interactions using network theory to identify suspicious online behaviors. Clustering users by Twitter mentions (a mention is a Tweet that contains another user's username anywhere in the body of the Tweet), we identify few groups of users with behaviors that could be fraudulent (multi-account posting, automatic posting, scheduled posting activity), favoring the manipulation hypothesis over the behavioral hypothesis. Overall, our findings shed light on the need for a higher control of the information published on social media and better education for investors seeking trading opportunities on the Internet.

Our paper is organized as follows. Section 3.2 briefly discusses the theoretical literature on market manipulation and reviews the empirical literature using data from the Internet. Section 3.3 describes the database we construct by analyzing SEC litigation releases and justifies our focus on OTC stocks

and Twitter. Section 3.4 covers the OTC Markets Group and data extracted from Twitter. Section 3.5 shows the results of the event studies. Section 3.6 proposes a methodology to identify potential fraudsters by analyzing interactions between users and discusses how to avoid frauds on social media. Section 3.7 presents our conclusions.

3.2 Related literature and hypothesis

Market manipulation undermines economic efficiency both by making prices less accurate as signals for efficient resource allocation and by making markets less liquid for risk transfer (Kyle & Viswanathan, 2008). Despite the importance of fair and transparent markets, little is know about the prevalence and impact of market manipulation (Putniņš, 2012). Theoretical studies have shown that traders can generate profits through trade-based manipulation (Allen & Gale, 1992) or information-based manipulation (Bommel, 2003). However, like any illegal behavior, market manipulation is not directly observable, and empirical studies remain very scarce. Owing to this lack of available data, our first strand of the literature focus on reported manipulation cases.

Studying all cases pursued by the Security and Exchange Commission from January 1990 to October 2001, Aggarwal & Wu (2006) present an extensive review of stock market manipulation in the United States. They find that around 50% of the manipulated stocks are small capitalization stocks (penny stocks) quoted in the OTC markets, such as the OTC Bulletin Board and the Pink Sheets.[3] With regard to techniques used by fraudsters, more than 55% of cases involve spreading rumors or false information. Manipulators also frequently use wash trades and nominee accounts to create artificial trading activity.

However, only a small fraction of manipulation is detected and prosecuted (Comerton-Forde & Putniņš, 2014). Further, focusing on reported cases tends to create a selection bias toward poor manipulation and is affected by the regulators' agenda (Bonner et al., 1998). Hence, another strand of the literature focuses on indirect evidences by studying abnormal market behaviors (for trade-based manipulation) or by detecting suspicious behaviors outside the market (for information-based manipulation).

[3]Only 17% of the reported cases occurs in the NYSE, the AMEX, or the NASDAQ.

Analyzing intraday volume and order imbalance, Ben-Davis et al. (2013) show evidence suggesting that some hedge funds manipulate stock prices on critical reporting dates. Their findings are consistent with those of Carhart et al. (2002) on end-of-quarter manipulation by mutual funds. In line with this study, a nascent strand of the literature focuses on information-based manipulation by analyzing new datasets of stock spams (newsletters) sent by fraudsters trying to pump the value of a stock. Böhme & Holz (2006), Frieder & Zittrain (2007), Hanke & Hauser (2008), and Nelson et al. (2013) all find a significant positive short-run price impact after a stock spam touting, followed by a price reversal over the following days. Similar patterns have been observed when Internet message board activity is used to identify pump-and-dump scheme on small stocks without fundamental news by Sabherwal et al. (2011).

In this paper, we follow both approaches. We first start by analyzing reported manipulation cases before conducting an empirical investigation of pump-and-dump schemes. Then, focusing on data from Twitter, we make two hypotheses. First, we hypothesize that a high number of positive messages about a company on Twitter is associated with a contemporaneous increase in the price of the stock, followed by a price reversal over the next trading days. Second, we hypothesize that if the price reversal is related to market manipulation or to stock promotion (pump-and-dump), we should identify abnormal behaviors and suspicious stock recommendations on social media (at least in some cases).

3.3 SEC litigation

Before extending the literature on indirect empirical evidence by analyzing information-based market manipulation on Twitter, we construct an updated database of SEC civil enforcement actions. Our work is closely related to Aggarwal & Wu (2006) who collect all SEC litigation releases containing keywords related to market manipulation published between 1990 and 2001[4] and then manually classify all cases by the type of stocks targeted (listed on NYSE, AMEX, NASDAQ, OTC Markets...) and the type of people involved (insiders, brokers, shareholders...). We complement their findings by (1) extending the sample period, (2) using the new SEC classification, and (3) examining specifically

[4]More precisely, they search for the keywords "manipulation" and "9(a)" or "10(b)" (which refer to the two articles of the Securities and Exchange Act of 1934). Mei, Wu, & Zhou (2004) use the same list plus the keyword "pump-and-dump".

pump-and-dump schemes to identify those involved in frauds (insiders, promoters, traders...) and the tools used to send false or misleading information in the marketplace (press releases, spam e-mails, websites, message boards, social media...).

Since 1996, each enforcement action is classified by the SEC into a unique category and the classification is shared via the "SEC annual reports" and the "Select SEC and Market Data reports". Our database of SEC litigation releases contains 4,918 civil actions from 1996 to 2015, of which 471 are related to market manipulation, a slightly higher number than in Aggarwal & Wu (2006) for comparable years. Table 3.1 shows the distribution of SEC civil actions by category and by fiscal year. In the remainder of this paper, we focus our analysis on the category "market manipulation", which includes the subcategory "newsletter/touting". Each case is included in only one category following SEC classification, even though many cases involve multiple allegations and may fall under more than one category.

[Insert Table 3.1 about here]

Overall, market manipulations account for 9.60% of all civil actions initiated by the SEC between 1996 and 2015. During our sample period, the SEC demonstrated its commitment to prosecuting market manipulation occurring in cyberspace on numerous occasions. For example in October 1998 (fiscal year 1999), the SEC launched a nationwide "sweep" for purveyors of fraudulent spam, online newsletters, message board postings and websites caught in an "effort to clean up the Internet", which led to 23 enforcement actions against 44 individuals and companies.[5] In 2000, the fourth nationwide Internet fraud sweep led to 15 enforcement actions against 33 companies and individuals who used the Internet to defraud investors. More recently, in July 2013, the SEC launched the Microcap Fraud Task Force to target abusive trading and fraudulent conduct in securities issued by microcap companies. This announcement was followed by a steep increase in the number of cases related to market manipulation on microcap companies between July and September 2013 (fiscal year 2013).

Even if the absolute number of reported cases is affected by the SEC agenda and may be biased toward poor manipulation, studying the different civil actions can still help us understand which type

[5] "SEC Conducts First Ever Nationwide Internet Securities Fraud Sweep, Charges 44 Stock Promoters in 23 Enforcement Actions"

of stocks are manipulated, who are the people involved and which tools, and techniques are used by fraudsters. Since 2002, the SEC has been releasing detailed complaints about a great majority of civil actions they initiate. While litigation releases only summarize the enforcement case in approximately one page, complaint reports provide more details about the fraudulent scheme and the exact role of each defendant, in ten to thirty pages. Using this new report, we manually analyze all complaints classified as "market manipulation" or "newsletter/touting". From the 362 "market manipulation" civil actions initiated by the SEC between 2002 and 2015, we managed to collect detailed complaint reports for 273 cases, of which 150 are related to pump-and-dump schemes. Table 3.2 summarizes, year by year, the type of stocks targeted by fraudsters, the type of people involved in the manipulation schemes, and the tools used to disseminate false or misleading information in the marketplace.

[Insert Table 3.2 about here]

We find that 86% of pump-and-dump schemes target stocks traded on OTC markets. The most common channel of communication used by fraudsters to send false or misleading information in the marketplace is press releases (73.3%), followed by spam e-mails/newsletters (34%), websites (32%), fax blast (12.6%), and message boards (10.6%).[6] Prosecuted cases mostly target frauds performed by company insiders (CEO, CFO) (60.7%), by stock promoters paid in cash or in shares to pump the price of a stock (49.3%)[7], and by traders/shareholders (37.3%).

In two cases, fraudsters specifically use Twitter to manipulate stock prices. In the first case, a Canadian couple used their website (PennyStockChaser), Facebook, and Twitter to pump up the stock of microcap companies and sold their shares after the pump (see Appendix A). In the second case, a Scottish trader falsely tweeted that two companies were being investigated, which caused sharp drops in the stock prices of the targeted companies (see Appendix B). Given these recent cases

[6]The sum is not equal to 100% as fraudsters often combine multiple channels of communication to increase the outreach and visibility of their messages.

[7]Stock promotion (investor relation) are not illegal per se. If promoters provide full disclosure of their compensation (type, amount, person paying the compensation) in all their communication, and if the information provided is neither false nor misleading, stock promotion can be legal. The Securities Act of 1933, Section 17(b) states the following: "It shall be unlawful for any person, by the use of any means or instruments of transportation or communication in interstate commerce or by the use of the mails, to publish, give publicity to, or circulate any notice, circular, advertisement, newspaper, article, letter, investment service, or communication which, though not purporting to offer a security for sale, describes such security for a consideration received or to be received, directly or indirectly, from an issuer, underwriter, or dealer, without fully disclosing the receipt, whether past or prospective, of such consideration and the amount thereof".

and the SEC's renewed attention toward risks created by social media communication, we believe that analyzing data from Twitter could provide new insights into the empirical literature on stock market manipulation.

3.4 Data

3.4.1 The OTC Markets Group

On the basis our preliminary analysis into SEC litigation, we choose to focus on stocks quoted by the OTC Markets Group. The OTC Markets Group is an electronic inter-dealer quotation and trading system providing marketplaces for around 10,000 OTC securities. The OTC Markets Group organizes securities into three tiered marketplaces: OTCQX, OTCQB, and OTC Pink. The marketplace on which a company trades reflects the integrity of its operations, its level of disclosure, and its degree of investor engagement.

1. OTCQX marketplace: Companies must meet high financial standards, be current in their disclosure, and receive third party advisory.

2. OTCQB marketplace: Companies must be current in their reporting, meet a minimum bid test of $0.01, and undergo an annual verification and management certification process.

3. OTC Pink marketplace: Open to all companies. The OTC Pink is divided into three subcategories based on the quantity and quality of information provided to investors: current information, limited information, and no information.

We download the list of all Common Stock and Ordinary Shares of companies incorporated in the United States, excluding American Depository Receipts, ETF, Funds, and Warrants. Our sample consists of 5,087 companies: 61 (1.20%) are quoted on OTCQX, 1,858 (36.52%) on OTCQB, and 3,168 (62.28%) on OTC Pink. Among the companies listed on OTC Pink, 814 provide current information, 403 provide limited information, and 1,951 provide no information. Companies in the last category should, according to the OTC Markets Group "be treated with suspicion and their securities should be considered highly risky."

We use Bloomberg to download daily price data, traded volume data, and market capitalization for all 5,087 stocks. During the sample period, the vast majority of the stocks experienced a sharp decrease in price with a number of stocks losing nearly all their value. This finding is consistent with the finding reported by Ang et al. (2013) that over a long period, comparable listed-stocks tend to overperform OTC stocks by nearly 9% per year. However, a few stocks also showed impressive returns over the sample period. For example, the price of Micro Imaging Technology increased from $0.0229 to $0.45 between October 2014 and October 2015 (+1,865%). As documented by Eraker & Ready (2015), the returns of OTC stocks are negative on average and highly positively skewed, with a few "lottery-like" stocks doing extremely well while many of the stocks become worthless.

3.4.2 Twitter data

Twitter is a micro-blogging platform that enables users to send and read short 140-character messages called "tweets". Every day, more than 500 million messages are posted on Twitter. We develop a computer program in the Python programming language to collect data in real time using Twitter Search and Stream Application Programming Interface (API). Following Da et al. (2011) and Drake et al. (2012), we identify a stock using its ticker symbol. Precisely, using the Twitter "cashtag" feature, introduced in 2012, we extract all the messages containing a "$" sign followed by the ticker name, as in Sprenger et al. (2014).

In the course of our sample period from October 5, 2014, to September 1, 2015, we collected a total of of 7,196,307 tweets. Among the 5,087 companies, around 50% received a very low level of attention (between 0 and 20 tweets). On the other hand, four companies featured in more than 100,000 tweets: Tykhe Corp (*$ HALB*), Cardinal Energy Group (*$ CEGX*), Sterling Consolidated (*$ STCC*) and Arrayit Corp (*$ ARYC*). Table 3.3 presents descriptive statistics for the top 10 most discussed companies in the sample period. Overall, we find that Twitter activity is higher for companies listed on the OTC Pink marketplace, with a low stock price (penny stocks) and a small market capitalization.

[Insert Table 3.3 about here]

By analyzing the Twitter messages for the ten most discussed companies in our sample, we identified a list of 255 fake Twitter accounts posting exactly the same type of messages at different periods,

simply by replacing a ticker with another and changing a few keywords over time. After a certain period of abnormally high posting activity, the number of tweets reduced to a level close to zero. While it is difficult to ascertain if those bursts in social media activity are directly linked with attempts to manipulate the market, the use of multiple fake accounts to recommend buying a stock is at least suspicious. Details about the users' behavior are presented later in this paper.

The case of Wholehealth Products, Inc. (*$ GWPC*), the eight most-discussed stock in our sample, is especially interesting. On November, 20, 2014, the Security Exchange Commission suspended trading on GWPC because of concerns regarding the accuracy and adequacy of publicly disseminated information by the company, including information about the relationship between the company's business prospects and the current Ebola crisis.[8] By examining the number of messages containing the ticker *$ GWPC* posted on Twitter before the SEC halt, we identify a sharp increase in posting activity starting on October 26th (Figure 3.1). A total of 2,768 tweets were sent on that day, compared to an average of less than 30 messages per day on the week before. The spike in posting activity on Twitter was followed by a one-week increase in stock price and a sharp price reversal afterward.

[Insert Figure 3.1 about here]

This anecdotal example is typical of a pump-and-dump scheme. A false piece of information is shared on Twitter to generate a spike in the social media activity about a given company. Stock price increases (pump) over a short period, and decreases sharply (dump) afterward. In the next section, we conduct an event study to analyze if the price reversal pattern identified anecdotally in the *$ GWPC* case can be generalized. We do so by analyzing the link between an abnormally high activity on social media and OTC stocks returns.

3.5 Event study

Following Tumarkin & Whitelaw (2001) and Leung & Ton (2015), we define an event as follows: when the number of messages posted on Twitter about company i during a given day t exceeds the average of the previous 7 days plus two standard deviations. To account for the regular operational

[8] "SEC Suspends Trading in Companies Touting Operations Related to Prevention or Treatment of Ebola", November 20, 2016

hours of the exchanges, we consider all messages sent between 4 p.m. on day *t-1* and 4 p.m. on day *t* as pertaining to day *t*. As event criteria, we impose a minimum of 20 tweets from 20 distinct users to avoid having our results driven by a few active users. If an event is detected on a non-trading days, we consider the next trading day as the event day. To include an event in our event study, we impose a minimum stock price of $0.1 and a market capitalization greater than $1,000,000 at the beginning of the event window. As in Ang et al. (2013) and Eraker & Ready (2015), we test various thresholds for minimum price, minimum market capitalization, and minimum percentage of non-trading days to avoid having our results driven by illiquid or non-tradable stocks. Our results are robust to a minimum trading price of $0.01 to $1, a minimum market capitalization of $100,000 to $10,000,000, and minimum no-trading percentage of 25% to 75% days.

The following example illustrates our methodology using a specific company: SinglePoint Inc, (*$ SING*). During the sample period, a total of 15,188 messages containing the ticker *$ SING* were posted on Twitter. Figure 3.2 shows the daily number of messages on Twitter, and the threshold level we use for event detection. Using this methodology, we identify six events for *$ SING* company: on October 14, 2014; November 12, 2014; January 24, 2015; April 1, 2015; July 13, 2015; and August 7, 2015. Table 3.4 shows a sample of tweets related to October 14, 2014, event. The activity on Twitter on that day is typical of a stock promotion scheme, where tweets are sent by bots at a regular schedule and through multiple accounts. All the accounts promoting the stock are owned by "Stock Talk 101", a firm "engaged in the business of marketing and advertising companies for monetary compensation". As a disclaimer is clearly visible on the stock promoter's Twitter accounts, the scheme is not per se illegal. However, this example illustrates how Twitter can be used by stock promoters as a new channel of communication.

[Insert Figure 3.2 and Table 3.4 about here]

For each event, we analyze all the tweets sent on that day, using a domain-specific sentiment lexicon. We convert each tweet into a quantitative sentiment variable, and we aggregate the individual message sentiments to derive an event sentiment. We find that 82.41% of event days are associated with a positive sentiment. As already documented in the literature (see, e.g., Kim & Kim, 2014; Avery et al., 2016), online investors are mostly bullish when sharing information about stock market on the

Internet. Individual investors do not (typically) sell short, hold small portfolios and are net-buyer of attention-grabbing stocks (Barber & Odean, 2008). Thus, when individual investors talk about a stock on the Internet, they tend to post messages mainly about the stock they hold or the stock they want to buy using a bullish (positive) vocabulary. In this investigation, the bullishness bias can also be viewed as fraudsters trying to pump the price of a stock by sharing (false) positive information about a given company on social media. Applying our earlier method to all the stocks listed on the OTC markets with a minimum stock price of $0.1 and a market capitalization greater than $1,000,000, and examining event days with a positive sentiment, we identified 567 events. The distribution of events over time does not exhibit any significant clustering around a specific period or day of the week.

To compute abnormal return, one has first to define a model for expected returns. However, choosing a daily model for normal returns of OTC stocks is tricky, as even a five-factor model explains only 57.3% of the variation of the returns of OTC stocks with monthly data (Ang et al., 2013). To define abnormal return, we thus consider three models of normal return: a constant mean return model, a market return model and a capital asset pricing model. We use the NASDAQ MicroCap Index as a benchmark of market return. We test the significance of abnormal return during the event window by conducting a non-parametric Corrado (1989) rank test, making no assumption about the normality of the underlying data. We present our results using a 6-month estimation window and a 21-day event window. On unreported robustness check, we find that our results are robust to a 12-month estimation window and a 11-day event window.

For each event detected previously, we compute abnormal return for the estimation window [-130:-11] (L1 = 120 days) and event window [-10:+10] (L2 = 21 trading days). We transform each abnormal return $AR_{i,t}$ to a rank variable $K_{i,t}$, by assigning to the day with the highest return over the complete window (estimation and event window) a rank of +141, to the day with the second highest return a rank of +140, and so on until we assign the lowest return a rank of 1. Tied ranks are treated by the method of midranks. To allow for missing returns, ranks are standardized by dividing by one plus the number of non-missing returns in each firm's excess returns time series.

$$K_{i,t} = \frac{rank(AR_{i,t})}{(1 + M_i)} \tag{3.1}$$

where M_i is the number of non-missing values for security i in L1 and L2. This yields order statistics for the uniform distribution with an expected value of one-half. The rank test statistic for day t (T_t) is equal to

$$T_t = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (U_{i,t} - 0.5)/S(U) \tag{3.2}$$

where N is equal to the number of events. The estimated standard deviation S(U) is defined on the estimation (L1) and event (L2) window as[9]

$$S(U) = \sqrt{\frac{1}{L_1 + L_2} \sum_t \left[\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} (U_{i,t} - 0.5) \right]^2} \tag{3.3}$$

where N_t represents the number of non-missing returns in the cross-section of N-firms on day t.

We conduct event studies using various thresholds for stock price, market capitalization, and percentage of non-trading days to determine whether or not a stock should be included. Precisely, we use four filtering methods: [1] all stocks with a minimum price at the beginning of the event window of \$0.1 and a market capitalization greater than \$1,000,000 (as defined previously), [2] all stocks with a minimum price of \$0.01 and a market capitalization greater than \$100,000, [3] all stocks with a minimum price of \$1 and a market capitalization greater than \$10,000,000, and [4] all stocks listed on the OTC Pink marketplace with a price greater than \$0.00001. We test the statistical significance of abnormal return on each day of the event window and on each 5-day rolling interval to identify a price reversal over a one week period. Table 3.5 summarizes the results based on different filtering methods. Figure 3.3 presents abnormal return (AR) and cumulative abnormal return (CAR) during the [-10:+10] event window, where day 0 is defined as a day of abnormally high activity on Twitter. Figure 3.4 shows the value of the non-parametric statistic computed by converting abnormal return to ranks on both the estimation and the event window. Results are presented when using a market return model to compute abnormal returns. Appendix C summarizes the results from a constant mean return model and a capital asset pricing model model.

[Insert Table 3.5, Figure 3.3 and Figure 3.4 about here]

[9] We also consider a multi-day version by multiplying by the inverse of the square root of the period's length.

As in Kim & Kim (2014), we identify a strong contemporaneous relationship between Twitter activity and stock price on the event day (*t0*). On analyzing stocks with a minimum price at the beginning of the event window of 0.1$ and a market capitalization greater than 1 million, we find an abnormal return of +6.49% on the event day. This finding is consistent with that of Sabherwal et al. (2011) who reported an increase of +13.93% on the event day when an event was defined as an abnormal number of messages on the financial message board "TheLion.com". When including all stocks listed on the OTC Pink marketplace, we find a significant increase of +5.80% on the day before the event and +22.68% on the event day.[10]

More interestingly, we find a significant post-event price reversal. Cumulative abnormal return is statistically significant and negative on an [+1:+5] window, with a post-event cumulative decrease in stock price between 2.5% and 3%. Again, this finding is consistent with that of Sabherwal et al. (2011) who observed a significant post-event decrease in stock price of -5.4% over over the 5 trading days following the event day. Two non-exclusive hypotheses can explain the price reversal pattern and the deviation from the efficient market hypothesis. First, we could conjecture that social media can be used as a proxy of investor overoptimism. In a market driven by unsophisticated traders with limits to arbitrage, price can deviate temporarily from its fundamental values in the presence of irrational sentiment-driven noise traders. In such a case, the price reversal identified on OTC stocks is simply caused by "standard" investor sentiment, as explained by Tetlock (2007). Another explanation could be that the sharp increase on the event day is caused by fraudsters or stock promoters pumping the price of targeted stocks, before dumping it on the following days after having made an illegal profit. This hypothesis would also be consistent with the price reversal pattern identified in our event studies.

To partially isolate one hypothesis from the other, we conduct a network analysis by examining the interactions between users in order to identify (if any) suspicious behaviors on Twitter.

[10] On unreported robustness checks, we find that results are robust when we remove events with return on the event day greater or equal to +50%.

3.6 Network analysis and suspicious behaviors

Differentiating an overoptimistic effect from manipulation is a challenging issue, as the goal of a pump-and-dump scheme is precisely to intensify positive sentiment. However, analyzing directed interactions between users on a network can help in identifying suspicious behaviors, as shown by Diesner et al. (2005) using the Enron email corpus. In that regard, Twitter offers an interesting framework, as interactions are directly observable through the function "mention" and "retweet"

Twitter's "retweet" function allows any user to share among their own list of followers any message created by another user. The "mention" function allows users to "tag" other members on a tweet to start a conversation with this user. The action of "retweeting" or "mentioning" can be considered as an interaction between two users. When user A chooses to retweet the original message posted by user B or to mention user B in a tweet, we can represent this interaction in a graph as a directed link between node A and node B. Then, as in any directed network, we can cluster users based on interaction similarities in order to identify potential suspicious behaviors. For example, if user A retweets all messages posted by user B, those two users will be clustered together. While clustering can also be caused by natural interactions or real friendships, an automatic approach helps identify suspicious patterns before we manually analyze interactions to confirm (or invalidate) our hypothesis.

From the 7,196,307 tweets in our database, we identify a total of 2,011,315 users' interactions (retweets or mentions). After excluding all users with less than 50 directed entrant or outbound links, we have a network of 8,961 users and 205,093 directed links. Figure 3.5 shows the Twitter network, where each nodes represents a user from our database, and each directed link is a retweet or a mention from one user by another. Clustering is based on directed links similarity, and node size depends on the number of entrant links. Colors depend on modularity, an optimization method for detecting community structure in networks (see, Brandes et al., 2008).

[Insert Figure 3.5 about here]

We identify five clusters characterized by a very high level of interactions between themselves and a low level of interactions with all the other clusters. The first cluster, in blue on the top right corner of Figure 3.5 is composed of 481 users. This group is organized around stock promoters from

the website http://stockmarketnews.co/ and include the official Twitter account of a company quoted on OTC market (*$ GPDB*, The Green PolkaDot Box). Analyzing all tweets containing the ticker *$ GPDB* during our sample period shows clear evidences of stock promotion.[11] For example, on July, 21, 2015, 9,533 tweets containing the ticker *$ GPDB* were published on Twitter by a total of 1,162 distinct users, without any specific news on that day. On analyzing the tweet content, we find only 31 distinct messages, suggesting that a promotion scheme involving fake accounts and automatic posting was underway. Activity on Twitter was abnormally high on the next day (2,176 tweets) before collapsing afterward (no tweets containing the ticker *$ GPDB* were sent between July 23 and July 29). Another promotion campaign started on July 30, with a total of 601 tweets.

The second cluster, in pink on the right side of Figure 3.5, is composed of accounts sharing an interest in cryptocurrencies. The user with the highest number of entrant links from this cluster is CannabisCoins, a "medical marijuana-backed digital currency" quoted on OTC markets (*$ CANN*). By studying all tweets related to *$ CANN* company, we identify different suspicious posting behaviors. For example, on October 9, 2014, between 3:13 a.m. and 3:22 a.m., a tweet by CannabisCoin announcing a future event was retweeted 735 times. The large peak in social activity was caused by a list of fake accounts retweeting automatically CannabisCoin's message to increase the message's outreach and visibility.[12]

Analyzing other clusters, we find similar patterns. The most common anomaly is a very large peak in volume on social media caused by a large number of (fake) accounts posting or retweeting a message on Twitter about a given company. A high number of users also declare themselves as stock promoters in their Twitter descriptions. Some groups of promoters tend to act together on various occasions to tout a stock with a spamming method resembling a spam blast or fax blast.

While it is true that promoting a tweet can also help a firm increase its sales or improve its brand awareness (without any manipulation), we believe that investors should always be very cautious of any information about OTC stocks posted on social media. According to the Security and Exchange Commission, "fraudsters can set up new accounts specifically designed to carry out their scam while

[11] The vast majority of the accounts from this cluster have now been suspended by Twitter for spam or inappropriate behaviors.

[12] Generating fake attention by buying followers or retweeters is very easy on Twitter as some companies offer, without respecting Twitter Terms of Services, to use fake accounts to automatically retweet a message at a cost around $5 for 1,000 retweets.

concealing their true identities" and investors should "be skeptical of information from social media accounts that lack a history of prior postings or sending messages". To the SEC's recommendation, we would like to add that investors should be skeptical of information published by any non-verified accounts and should carefully examine previous tweets from the user to detect any of the anomalies highlighted above (scheduled automatic postings, previous tweets not related to the financial market, abnormal followers/retweets ratio...).

At the same time, users should also be aware of those who are committed to exposing pump-and-dump schemes through their tweets, often using proprietary algorithm to detect anomalies and inform market participants. For example, a user named "ThePumpTracker" (now "theOTCtoday") used to publish alerts on Twitter on detecting a stock that was under promotion. By checking alerts from this Twitter account on the days on which we identified abnormal Twitter activity, we find that around 10% of our events where also identified by "ThePumpTracker" as being related to a stock promotion. We recommend that investors actively find users tracking pump-and-dump schemes and stock promotions before investing in OTC stocks.

While further research is needed to better understand how information is disseminated on Twitter, we believe that our analysis provides better insights into the techniques that could be used by potential fraudsters on social media. It could also help investors avoid penny stock scams on the Internet. Our finding reinforces the SEC recommendation that "investors who learn of investing opportunities from social media should always be on the lookout for fraud".

3.7 Conclusion

Social media can help investors gather and share information about stock markets. However, it also presents opportunities for fraudsters to send false or misleading statements in the marketplace. In that regard, the social media platform Twitter is a very attractive channel for manipulators or stock promoters as it allows them to target a wide unsophisticated audience, more prone to being scammed than sophisticated investors. The anonymity of Twitter and the ease with which fake accounts and/or bots can be used to spam the network also facilitate fraudsters' activities.

In this paper, we first analyze all SEC litigation releases by focusing on pump-and-dump schemes. We find that information-based stock market manipulation mainly targets small capitalization stocks traded over the counter. Market manipulators use various channels of communication to send false or misleading information in the marketplace, such as press releases, spam e-mails, and websites. In two cases prosecuted by the SEC (Litigation Release No. 21580 and Litigation Release No. 23401), fraudsters specifically use Twitter to manipulate stock prices.

Then, we complement the literature on indirect empirical evidence of market manipulation by analyzing a novel dataset of more than seven million messages published on Twitter during a one-year period. We provide empirical evidence showing that fraudsters can use social media to artificially inflate the price of a stock. Defining an event as an abnormally high posting activity on Twitter about a given company, we identify a large increase in stock price on the event day, followed by a sharp price reversal over the next five trading days. Examining interactions between users (retweets and mentions), we identify suspicious clusters of Twitter users, using fake accounts, automatic postings, or scheduled retweets to recommend buying a stock. While a judicial inquiry would be needed to assess if the promotion is legal or not, our findings shed light on the need for higher control over the information published on social media and better education for investors seeking trading opportunities on the Internet. Given the risk of manipulation and the average negative return of OTC stocks, we reaffirm that individual investors should be very cautious when choosing to invest on risky and illiquid small capitalization stocks.

3.8 Appendix A - Litigation release No. 21580

U.S. SECURITIES AND EXCHANGE COMMISSION - Litigation Release No. 21580 / June 29, 2010

Securities and Exchange Commission v. Carol McKeown, Daniel F. Ryan, Meadow Vista Financial Corp., and Downshire Capital, Inc., Civil Action 10-80748-CIV-COHN (S.D. Fla. June 23, 2010)

The Securities and Exchange Commission announced today that it has obtained an emergency asset freeze against a Canadian couple who fraudulently touted penny stocks through their website, Facebook and Twitter. The SEC also charged two companies the couple control and obtained an asset freeze against them. According to the SEC's complaint, the defendants profited by selling penny stocks at or around the same time that they were touting them on www.pennystockchaser.com. The website invites investors to sign up for daily stock alerts through email, text messages, Facebook and Twitter.

The SEC alleges that since at least April 2009, Carol McKeown and Daniel F. Ryan, a couple residing in Montreal, Canada, have touted U.S. microcap companies. According to the SEC's complaint, McKeown and Ryan received millions of shares of touted companies through their two corporations, defendants Downshire Capital Inc., and Meadow Vista Financial Corp., as compensation for their touting. McKeown and Ryan sold the shares on the open market while PennyStockChaser simultaneously predicted massive price increases for the issuers, a practice known as "scalping." The SEC's complaint, filed in the U.S. District Court for the Southern District of Florida, also alleges McKeown, Ryan and one of their corporations failed to disclose the full amount of the compensation they received for touting stocks on PennyStockChaser. The SEC alleges that McKeown, Ryan and their corporations have realized at least $2.4 million in sales proceeds from their scalping scheme.

The SEC's complaint charges McKeown, Ryan, Downshire Capital Inc. and Meadow Vista Financial Corp. with violating Section 17(a) of the Securities Act of 1933, Section 10(b) of the Securities Exchange Act of 1934, and Rule 10b-5 thereunder. The SEC's complaint also charges McKeown, Ryan and Meadow Vista Financial Corp. with violating Section 17(b) of the Securities Act of 1933. In addition to the emergency relief already granted by the U.S. District Court the Commission also seeks a preliminary injunction and permanent injunction, along with disgorgement of ill-gotten gains plus prejudgment interest and the imposition of a financial penalty, penny stock bars against the individuals and the repatriation of assets to the United States.

In the course of its investigation, the SEC worked with the Quebec Autorité des marchés financiers (AMF), which was also investigating this matter. As a result of both ongoing investigations, the AMF obtained an emergency order freezing assets and a cease trade order against McKeown, Ryan, Downshire Capital Inc. and Meadow Vista Financial Corp. The SEC appreciates the collaboration with the AMF. The SEC's case was investigated by Michael L. Riedlinger, Timothy J. Galdencio and Eric R. Busto of the Miami Regional Office. The SEC's litigation effort will be led by Christine Nestor, Amie R. Berlin and Robert K. Levenson. The SEC's investigation is continuing.

3.9 Appendix B - Litigation release No. 23401

U.S. SECURITIES AND EXCHANGE COMMISSION - Litigation Release No. 23401 / November 6, 2015

Securities and Exchange Commission v. James Alan Craig, Civil Action No. 3:15-cv-05076) (N.D. Cal.)

On November 5, 2015, the Securities and Exchange Commission filed securities fraud charges against a Scottish trader whose false tweets caused sharp drops in the stock prices of two companies and triggered a trading halt in one of them.

According to the SEC's complaint filed in federal court in the Northern District of California, James Alan Craig of Dunragit, Scotland, tweeted multiple false statements about the two companies on Twitter accounts that he deceptively created to look like the real Twitter accounts of well-known securities research firms. Also yesterday, the U.S. Attorney's Office for the Northern District of California filed criminal charges.

The SEC's complaint alleges that Craig's first false tweets caused one company's share price to fall 28 percent before Nasdaq temporarily halted trading. The next day, Craig's false tweets about a different company caused a 16 percent decline in that company's share price. On each occasion, Craig bought and sold shares of the target companies in a largely unsuccessful effort to profit from the sharp price swings.

The SEC's investigation also determined that Craig later used aliases to tweet that it would be difficult for the SEC to determine who sent the false tweets because real names weren't used. According to the SEC's complaint:

- On Jan. 29, 2013, Craig used a Twitter account he created to send a series of tweets that falsely said Audience, Inc. was under investigation. Craig purposely made the account look like it belonged to the securities research firm Muddy Waters by using the actual firm's logo and a similar Twitter handle. Audience's share price plunged and trading was halted before the fraud was revealed and the company's stock price recovered.
- On Jan. 30, 2013, Craig used another Twitter account he created to send tweets that falsely said Sarepta Therapeutics, Inc. was under investigation. In this case Craig deliberately made the Twitter account seem like it belonged to the securities research firm Citron Research, again using the real firm's logo and a similar Twitter handle. Sarepta's share price dropped 16 percent before recovering when the fraud was exposed.

The Commission's complaint charges that Craig committed securities fraud in violation of Section 10(b) of the Securities Exchange Act of 1934 and Rule 10b-5 thereunder. The complaint seeks a permanent injunction against future violations, disgorgement and a monetary penalty from Craig.

The SEC has issued an Investor Alert titled "Social Media and Investing - Stock Rumors" prepared by the Office of Investor Education and Advocacy. The alert aims to warn investors about fraudsters who may attempt to manipulate share prices by using social media to spread false or misleading information about stocks, and provides tips for checking for red flags of investment fraud.

The SEC's investigation was conducted by staff in the Market Abuse Unit including Elena Ro, John Rymas, and Steven D. Buchholz. The case was supervised by Joseph G. Sansone, Co-Chief of the Market Abuse Unit. The SEC's litigation will be led by Ms. Ro and John S. Yun of the SEC's San Francisco Regional Office. The SEC acknowledges the assistance of the U.S. Department of Justice and the Federal Bureau of Investigation.

3.10 Appendix C - Constant mean return model and capital asset pricing model

Abnormal returns and cumulative abnormal returns (5-day) - Constant mean return model

	[1]		[2]		[3]		[4]	
	AR	5-day CAR	AR	5-day CAR	AR	5-day CAR	AR	5-day CAR
t-10	0.0035	0.0151	0.0059	0.0166	0.0045	0.0007	-0.0042	-0.0041
t-9	-0.0002	0.0148	-0.0020	0.0129	-0.0046	-0.0048	0.0022	0.0078
t-8	-0.0005	0.0090	-0.0004	0.0113	-0.0011	-0.0060	0.0004	0.0078
t-7	-0.0010	0.0058	0.0010	0.0069	-0.0013	-0.0081	0.0038	0.0084
t-6	-0.0059	-0.0041	-0.0060	-0.0015	-0.0069	-0.0093	0.0003	0.0025
t-5	-0.0029	-0.0105	-0.0060	-0.0134	-0.0065	-0.0204	-0.0018	0.0049
t-4	-0.0027	-0.0130	0.0002	-0.0112	-0.0061*	-0.0220**	-0.0003	0.0024
t-3	0.0009	-0.0116	0.0072*	-0.0036	0.0028	-0.0180*	0.0034	0.0054
t-2	-0.0028	-0.0134	-0.0008	-0.0054	-0.0090	-0.0258	0.0051	0.0067
t-1	0.0310***	0.0234**	0.0379***	0.0385***	0.0185***	-0.0003	0.0531***	0.0595**
t0	0.0663***	0.0926***	0.0842***	0.1288***	0.0574***	0.0636***	0.2341***	0.2954***
t1	-0.0062	0.0891***	-0.0043	0.1243***	-0.0002	0.0696***	-0.0152	0.2805***
t2	-0.0048	0.0835***	-0.0035	0.1135***	-0.0115*	0.0553***	-0.0103	0.2668***
t3	-0.0066	0.0797***	-0.0060	0.1083***	-0.0038	0.0605***	-0.0159	0.2458***
t4	-0.0071*	0.0415	-0.0040	0.0664	-0.0063	0.0357	-0.0026	0.1902
t5	-0.0001	-0.0249***	-0.0042	-0.0221**	-0.0096*	-0.0314***	0.0071	-0.0369
t6	-0.0130	-0.0317***	-0.0077	-0.0254**	-0.0082	-0.0394***	-0.0175*	-0.0392
t7	-0.0014	-0.0283**	-0.0072	-0.0291**	-0.0087*	-0.0365***	-0.0037	-0.0325
t8	0.0004	-0.0212**	0.0009	-0.0222*	-0.0056	-0.0384***	-0.0076	-0.0243
t9	-0.0085	-0.0225*	0.0001	-0.0181	-0.0039	-0.0361***	0.0072	-0.0146
t10	-0.0087	-0.0311*	-0.0066	-0.0205	-0.0038	-0.0302**	0.0012	-0.0204
Event	561		925		257		877	

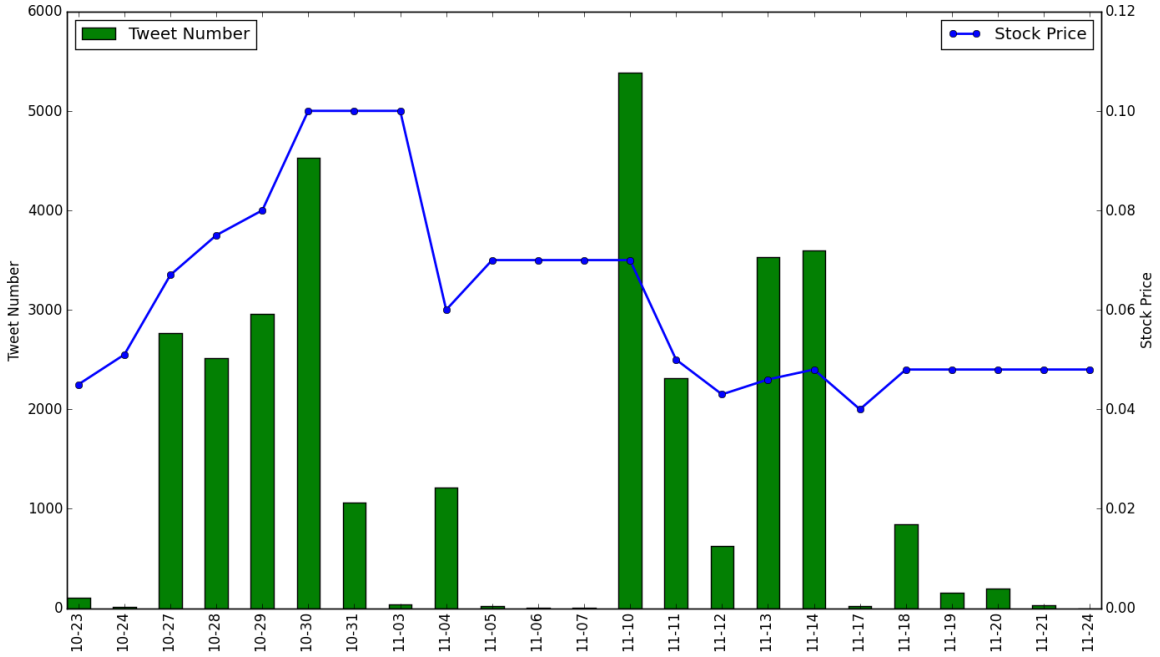
Notes: This table shows the abnormal returns, relative to the event day *t0*, on a [-10:+10] days event window. Cumulative abnormal returns on day t are equal to the sum of abnormal returns from day t-4 to day t. ***, ** and * represent abnormal returns significance respectively at the 1%, 5%, and 10% level using a Corrado rank test. Results are presented for [1] stocks with a price greater than $0.10 and a market capitalization greater than $1,000,000, [2] stocks with a price greater than $0.01 and a market capitalization greater than $100,000, [3] stocks with a price greater than $1 and a market capitalization greater than $10,000,000, [4] all stocks listed on the OTC Pink marketplace with a price greater than $0.00001. Normal returns are computed using a constant mean return model.

Abnormal returns and cumulative abnormal returns (5-day) - Capital asset pricing model

	[1]		[2]		[3]		[4]	
	AR	5-day CAR	AR	5-day CAR	AR	5-day CAR	AR	5-day CAR
t-10	0.0046	0.0202	0.0054	0.0159	0.0074	0.0149	-0.0024	-0.0024
t-9	0.0014	0.0217	-0.0015	0.0129	-0.0005	0.0110	0.0044	0.0078
t-8	0.0018	0.0170	-0.0009	0.0103	0.0038	0.0106	-0.0002	0.0085
t-7	-0.0010	0.0121	0.0023	0.0067	-0.0009	0.0058	0.0057	0.0130
t-6	-0.0048	0.0020	-0.0060	-0.0006	-0.0020	0.0079	0.0043	0.0119
t-5	-0.0013	-0.0039	-0.0053	-0.0114	-0.0027	-0.0023	0.0022	0.0165
t-4	-0.0052	-0.0105	-0.0020	-0.0119	-0.0066*	-0.0084	-0.0011	0.0110
t-3	-0.0015	-0.0138	0.0053	-0.0057	-0.0006	-0.0129	0.0036	0.0147
t-2	0.0001	-0.0127	-0.0010	-0.0090	-0.0038	-0.0158	0.0077	0.0167
t-1	0.0339***	0.0260**	0.0388***	0.0358***	0.0212***	0.0074	0.0536***	0.0660*
t0	0.0664***	0.0937***	0.0819***	0.1230***	0.0560***	0.0662***	0.2300***	0.2938***
t1	-0.0060	0.0929***	-0.0062	0.1188***	0.0040	0.0767***	-0.0096	0.2853***
t2	-0.0040	0.0903***	-0.0035*	0.1099***	-0.0083	0.0691***	-0.0085	0.2732***
t3	-0.0068	0.0834***	-0.0071	0.1039***	-0.0013	0.0716***	-0.0155	0.2500***
t4	-0.0067*	0.0428	-0.0031	0.0619	-0.0026	0.0478*	0.0016	0.1979
t5	0.0034	-0.0201**	-0.0033	-0.0233**	-0.0060	-0.0142*	0.0079	-0.0241**
t6	-0.0136*	-0.0277**	-0.0092	-0.0263***	-0.0064	-0.0246**	-0.0174**	-0.0319**
t7	0.0000	-0.0236**	-0.0071	-0.0298**	-0.0048	-0.0211*	-0.0025	-0.0259**
t8	0.0051	-0.0118**	0.0040	-0.0187**	-0.0001	-0.0199**	-0.0075	-0.0178
t9	-0.0056	-0.0106	0.0009	-0.0147	-0.0009	-0.0181**	0.0066	-0.0128
t10	-0.0051	-0.0191	-0.0033	-0.0147	0.0023	-0.0098	0.0101	-0.0106
Event	523		882		234		868	

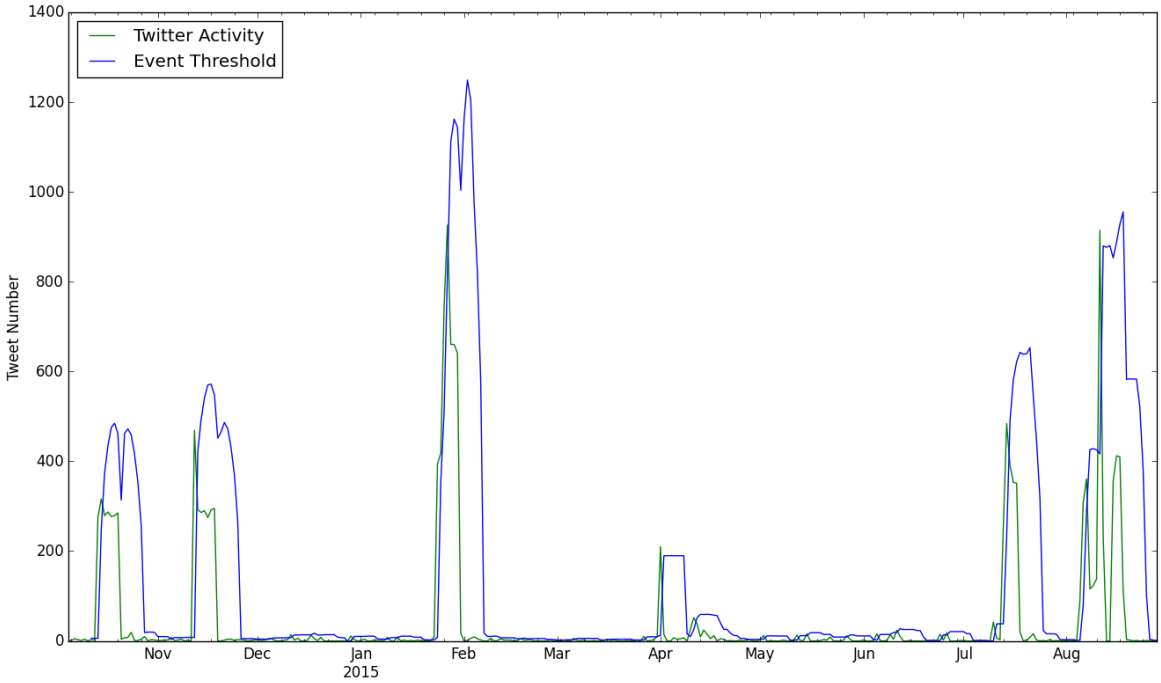
This table shows the abnormal returns, relative to the event day *t0*, on a [-10:+10] days event window. Cumulative abnormal returns on day t are equal to the sum of abnormal returns from day t-4 to day t. ***, ** and * represent abnormal returns significance respectively at the 1%, 5%, and 10% level using a Corrado rank test. Results are presented for [1] stocks with a price greater than $0.10 and a market capitalization greater than $1,000,000, [2] stocks with a price greater than $0.01 and a market capitalization greater than $100,000, [3] stocks with a price greater than $1 and a market capitalization greater than $10,000,000, [4] all stocks listed on the OTC Pink marketplace with a price greater than $0.00001. Normal returns are computed using a capital asset pricing model.

FIGURE 3.1: Wholehealth Products, Inc (*$ GWPC*) - Stock price and Twitter activity



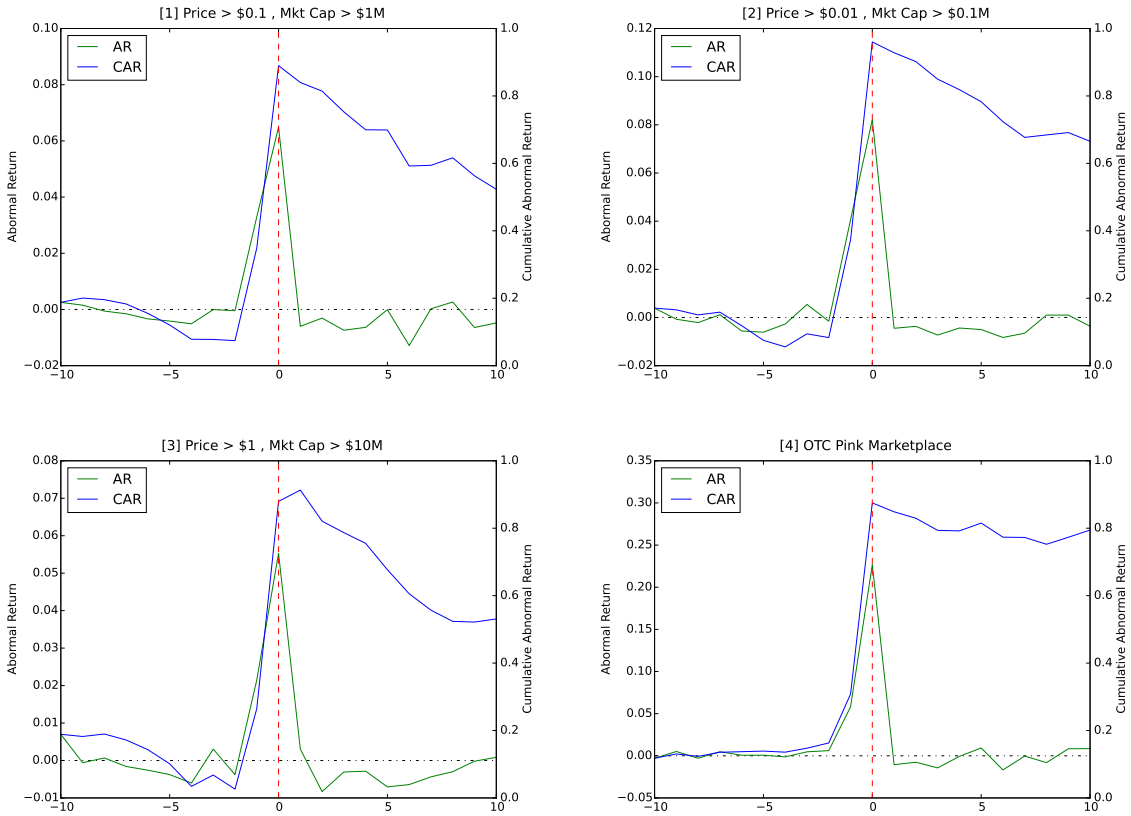
Notes: This figure shows the price of the Wholehealth Products (*$ GWPC*) shares (right-axis) and the daily number of messages containing the cashtag *$ GWPC* posted on Twitter between October 23, 2014, and November 24, 2014 (left-axis). Due to the SEC investigation, $GWPC stock price is flat at $0.048 between November 20, and November 24. *$ GWPC* stock price drops to $0.0001 when trading resumes on December 23, 2014.

FIGURE 3.2: SinglePoint, Inc (*$ SING*) - Twitter activity and event detection



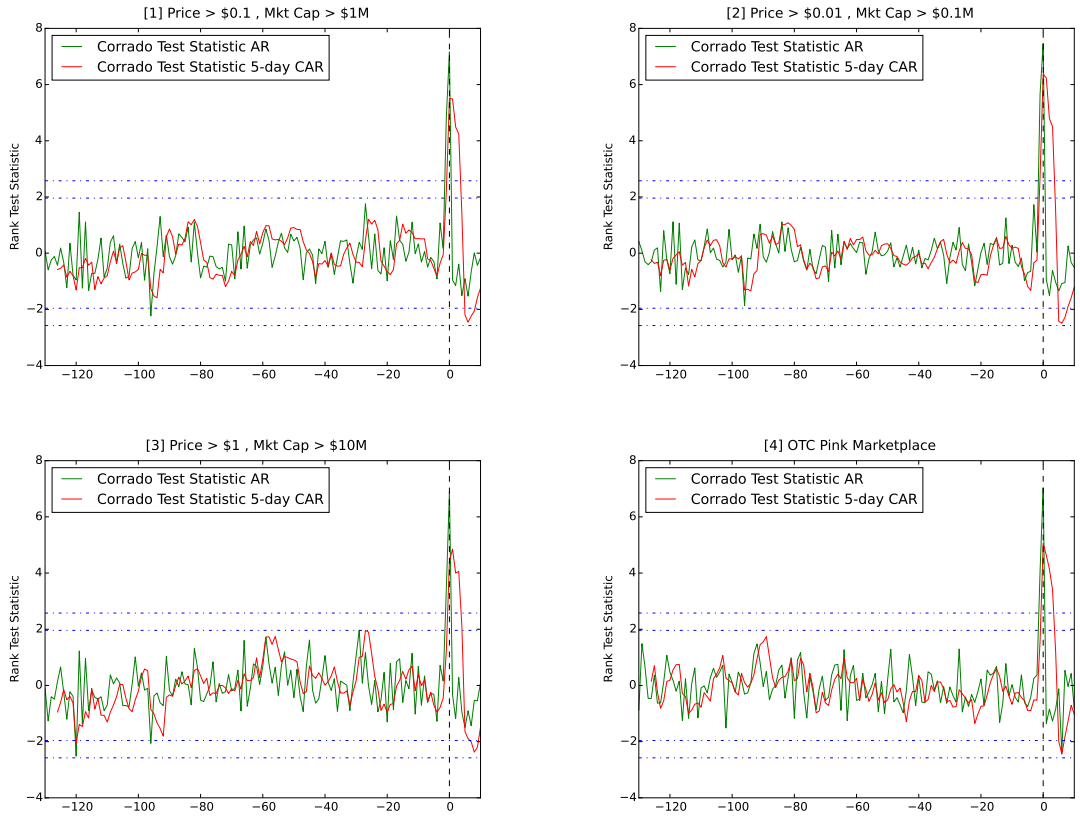
Notes: This figure shows the daily number of messages containing the cashtag *$ SING* posted on Twitter during the sample period (in blue) and the threshold level we use for event detection (in green, average daily number of messages of the previous 7 days plus two standard deviations).

FIGURE 3.3: Event Study - Abnormal returns and cumulative abnormal returns



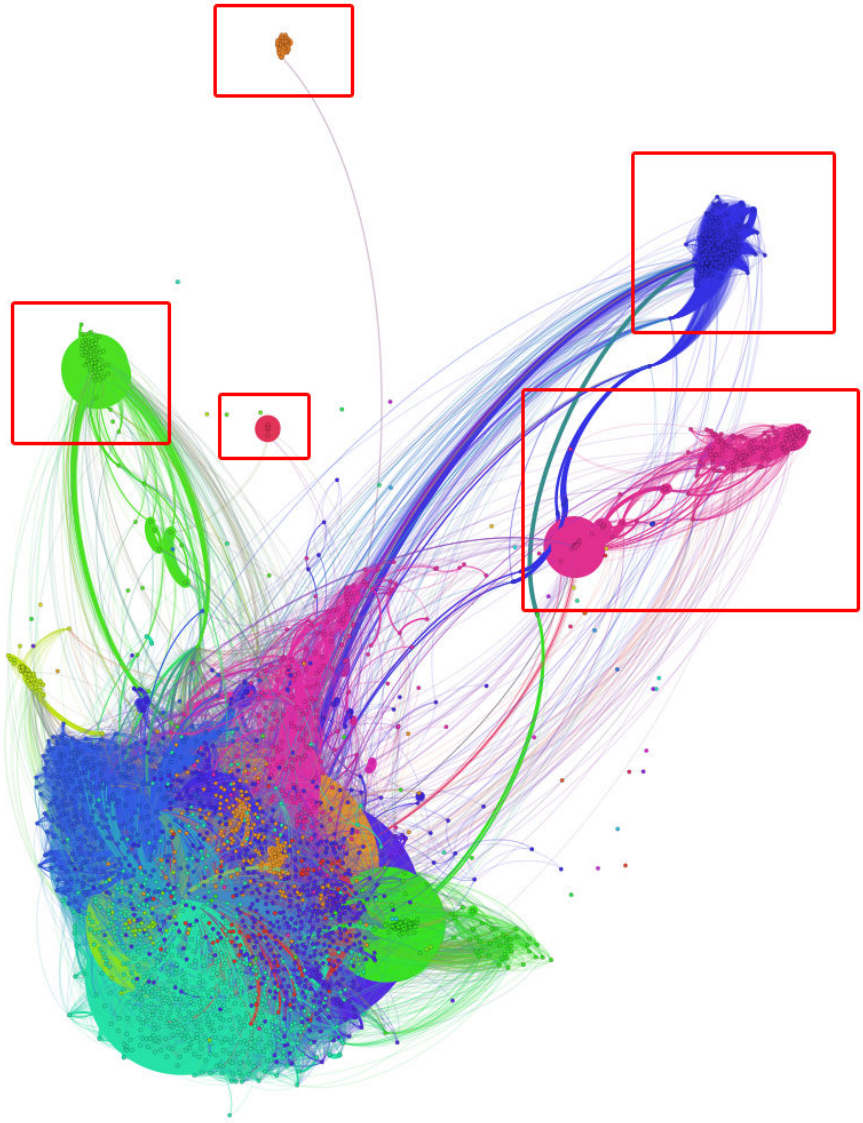
Notes: This figure shows the abnormal returns and the cumulative abnormal returns on a [-10:+10] days event window. Results are presented for [1] stocks with a price greater than $0.10 and a market capitalization greater than $1,000,000, [2] stocks with a price greater than $0.01 and a market capitalization greater than $100,000, [3] stocks with a price greater than $1 and a market capitalization greater than $10,000,000, [4] all stocks listed on the OTC Pink marketplace with a price greater than $0.00001.

FIGURE 3.4: Event study - Abnormal returns and cumulative abnormal returns - Rank test



Notes: This figure shows the one-day standardized average rank (green) and the 5-day rolling average rank (red) for both the estimation window [-130:-11] and the event window [-10:+10]. Horizontal dashed blue lines represent significance thresholds at the 5% level and 1% level. Results are presented for [1] stocks with a price greater than $0.10 and a market capitalization greater than $1,000,000, [2] stocks with a price greater than $0.01 and a market capitalization greater than $100,000, [3] stocks with a price greater than $1 and a market capitalization greater than $10,000,000, [4] all stocks listed on the OTC Pink marketplace with a price greater than $0.00001.

FIGURE 3.5: Network analysis of the Twittersphere based on retweets and mentions



Notes: This figure shows interactions between users on Twitter. The graph is generated using Gephi, an open-source network analysis and visualization software. Each node represents a user and each link (edge) an interaction between two users. Clustering is based on directed links similarities using Force Atlas, a force-directed layout algorithm. Colors depend on modularity, an optimization method for detecting community structure in networks. Suspicious clusters are framed in red.

TABLE 3.1: Number of SEC civil actions by category and by fiscal year

	Broker-Dealer	Insider Trading	Securities Offering	Market Manipulation	Other Civil Actions	Total
1996	23	29	76	4	81	213
1997	19	36	66	11	72	204
1998	20	38	82	18	71	229
1999	15	51	67	26	78	237
2000	20	36	70	34	99	259
2001	13	47	56	17	104	237
2002	16	52	80	26	143	317
2003	32	37	70	18	156	313
2004	17	32	59	17	139	264
2005	20	42	34	30	138	264
2006	6	37	45	22	107	217
2007	59	31	44	27	101	262
2008	67	37	67	39	75	285
2009	26	42	106	34	104	312
2010	7	34	73	24	117	255
2011	21	48	82	29	86	266
2012	16	52	73	34	95	270
2013	7	43	76	23	58	207
2014	7	40	52	11	35	145
2015	4	26	58	28	46	162
Total	415	790	1,336	472	1,905	4,918
Total (%)	8.44%	16.06%	27.16%	9.60%	38.74%	100%

Notes: This table reports the number of SEC civil actions by category and by fiscal year. The category "Other Civil Actions" includes: "Investment Advisors/Companies", "Delinquent Fillings", "Civil Contempt", "Transfer Agents" and "Miscellaneous". The category "Market Manipulation" includes "Newsletter/Touting", a category initiated by the SEC in 1999 and re-integrated into "Market Manipulation" in 2003.

TABLE 3.2: Distribution of pump-and-dump manipulation cases

	Stock targeted		People involved				Tools used								
	OTC	Other markets	Insider	Promoter	Trader Shareholder	Other People	Press release	E-mail	Website	Fax	Telephone	Mailer	Message board	Social media	Other tools
2002	12	4	8	9	5	5	11	8	10	3	2	0	1	0	4
2003	8	1	6	4	3	2	6	4	4	2	0	0	2	0	4
2004	2	3	3	2	2	1	4	2	0	1	0	0	2	0	1
2005	11	7	12	8	4	2	12	5	5	3	3	0	1	0	5
2006	10	1	8	7	2	3	7	3	5	4	1	0	1	0	3
2007	7	2	6	3	4	1	8	3	2	2	0	0	2	0	3
2008	16	0	14	3	4	0	15	2	2	0	0	3	1	0	3
2009	12	0	6	7	5	2	8	3	3	3	0	2	0	0	4
2010	11	2	9	5	5	0	11	4	6	0	1	0	1	2	6
2011	7	0	3	5	4	2	5	4	3	1	1	0	0	0	0
2012	11	0	7	7	4	2	7	3	2	0	1	1	2	0	3
2013	3	0	2	3	0	1	3	3	0	0	1	0	1	1	1
2014	6	1	2	4	6	1	5	3	3	0	0	0	1	1	0
2015	13	0	5	7	8	2	8	4	3	0	0	3	1	0	2
Total	129	21	91	74	56	24	110	51	48	19	10	9	16	4	39
(%)	86.00	14.00	60.67	49.33	37.33	16.00	73.33	34.00	32.00	12.67	6.67	6.00	10.67	2.67	26.00

Notes: This table reports the distribution of pump-and-dump schemes from 2002 to 2015. Each case may involve multiple stocks, multiple people and multiple tools. "OTC" includes both the OTC Bulletin Board and the Pink Sheets. "Other markets" includes NASDAQ, NYSE, AMEX. "Insiders" are in great majority CEOs and CFOs. "Promoters" are paid investor relationship companies. "Other people" includes broker-dealers, attorneys and analysts. "Press releases" are disseminated through online wires like "PR Newswires" or "Business Wire". "E-mail" includes specialized newsletters and blast unsolicited e-mail. "Website" includes both companies' websites and promoters' websites. "Messages board" includes Yahoo! Finance, the Raging Bull, Investor Hub. "Social media" includes Twitter, Facebook and LinkedIn. "Other tools" includes mainly fake analyst reports and false filings sent to the SEC or to the FINRA.

TABLE 3.3: Top 10 most discussed OTC Markets stocks on Twitter

Ticker	Company	Market	Disclosure	Tweet Number	Stock Price	Market Cap
$HALB	Tykhe Corp	OTC Pink	Current	397,098	0.01	#NA
$CEGX	Cardinal Energy Group	OTC Pink	Current	169,263	0.8	28.14
$STCC	Sterling Consolidated	OTC Pink	Limited	143,572	0.045	1.81
$ARYC	Arrayit Corp	OTCQB	#NA	104,683	0.1624	6.43
$GPDB	Green Polkadot Box	OTC Pink	Current	93,352	1.85	19.75
$MINE	Minerco Resources	OTC Pink	Current	80,330	0.7813	19.04
$MYEC	MyEcheck	OTC Pink	Current	49,940	0.0202	81.00
$GWPC	Wholehealth Products	OTC Pink	Limited	36,500	0.25	19.92
$PUGE	Puget Technologies	OTCQB	#NA	32,797	0.0556	2.36
$CELH	Celsius Holdings	OTC Pink	Current	31,041	0.5283	9.78

Notes: This table presents the number of messages published on Twitter between October 5, 2014, and September 1, 2015, for the 10 most discussed stocks in our sample. Stock price (in USD) and market capitalization (in million USD) as of October, 1, 2014. #NA is used to indicate when information is not provided by Bloomberg.

TABLE 3.4: Messages containing the cashtag $SING posted on Twitter on October 13, 2014.

Date	User	Message content
2014-10-13 17:12:05	ckelly3	RT @majegivudys: $SING - SinglePoint's product suite will provide Medical Cannabis dispensaries a user-friendly [...]
2014-10-13 17:36:19	_Singlepoint_	$SING working to finish acquiring GreenStar Payment Solutions in short order
2014-10-13 17:44:57	badnewsbruno	RT @_Singlepoint_: $SING working to finish acquiring GreenStar Payment Solutions in short order
2014-10-13 18:54:40	_Singlepoint_	$SING increasing number of terminals every week on track to hit sales targets
2014-10-13 20:00:01	JayBugster	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:02	BoardwalkPennyS	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:02	Micro_Cap_Pro	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:02	MicroCapUnivers	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:03	PennyStockMach	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:03	StockShocks	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:03	PennyStockExcel	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:03	DaddyHotStocks	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:03	StockUltraman	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:03	HotStockCafe	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:03	Penny_Hotsocks	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:07	PlatinumPennys	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:11	Virmmac	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 20:00:13	IonPennyStocks	When completed GreenStar Payment Solutions, Inc. will be a wholly owned subsidiary of SinglePoint. - $SING
2014-10-13 21:00:03	JayBugster	Did you read the $SING LOI news? http://t.co/gTEykD6oSz
2014-10-13 21:00:03	Virmmac	Did you read the $SING LOI news? http://t.co/AxEUNqLoOk
2014-10-13 21:00:04	BoardwalkPennyS	Did you read the $SING LOI news? http://t.co/arilkU2OIX
2014-10-13 21:00:04	StockShocks	Did you read the $SING LOI news? http://t.co/ixKf03Rm9C
2014-10-13 21:00:04	PlatinumPennys	Did you read the $SING LOI news? http://t.co/SUg48Rf392
2014-10-13 21:00:04	StockUltraman	Did you read the $SING LOI news? http://t.co/WDy1STjZ2B
2014-10-13 21:00:04	Micro_Cap_Pro	Did you read the $SING LOI news? http://t.co/8LlCPwiBLV
2014-10-13 21:00:06	HotStockCafe	Did you read the $SING LOI news? http://t.co/f3EKncSwQq
2014-10-13 21:00:07	Penny_Hotsocks	Did you read the $SING LOI news? http://t.co/NrH2obv7ti http://t.co/Ln1HqzOnok
2014-10-13 21:00:11	PennyStockMach	Did you read the $SING LOI news? http://t.co/Hx4mynnUDr
2014-10-13 21:00:12	JayBugster	Did you read the $SING LOI news? http://t.co/b6brmQtl5t http://t.co/eqgcRTP9NG
2014-10-13 21:00:12	Penny_Hotsocks	Did you read the $SING LOI news? http://t.co/r8Au3hcEaL
2014-10-13 21:00:13	IonPennyStocks	Did you read the $SING LOI news? http://t.co/WJ5vFU7CHB
2014-10-13 21:04:54	aheadsupotc	$SING received a new alert.See why at http://t.co/FwU0sYdHLW $MSEZ #Penny #pennystocks [...]
2014-10-13 21:06:11	ckelly3	RT @aheadsupotc: $SING received a new alert.See why at http://t.co/FwU0sYdHLW $MSEZ #Penny #pennystocks [...]
2014-10-13 22:00:00	PennyStockExcel	SinglePoint, Inc. Signs LOI to Acquire 100% of GreenStar Payment Solutions - $SING
2014-10-13 22:00:00	StockUltraman	SinglePoint, Inc. Signs LOI to Acquire 100% of GreenStar Payment Solutions - $SING
2014-10-13 22:00:01	Micro_Cap_Pro	SinglePoint, Inc. Signs LOI to Acquire 100% of GreenStar Payment Solutions - $SING
2014-10-13 22:00:01	IonPennyStocks	SinglePoint, Inc. Signs LOI to Acquire 100% of GreenStar Payment Solutions - $SING
2014-10-13 22:00:02	HotStockCafe	SinglePoint, Inc. Signs LOI to Acquire 100% of GreenStar Payment Solutions - $SING

Notes: This table presents a sample of messages containing the cashtag *$ SING* posted on Twitter after the stock market closes on October 13, 2014, illustrating how stock promoters use automatic posting to create fake online activity. For example, in 12 seconds between 20:00:01 and 20:00:13, the same tweet was sent by 14 different Twitter accounts belonging to the same paid advertiser (multiple accounts spamming). The exact same pattern appears one hour later, from 21:00:03 to 21:00:13 (scheduled posts).

TABLE 3.5: Abnormal returns and cumulative abnormal returns (5-day) - Market return model

	[1]		[2]		[3]		[4]	
	AR	5-day CAR	AR	5-day CAR	AR	5-day CAR	AR	5-day CAR
t-10	0.0025	0.0156	0.0039	0.0121	0.0070	0.0133	-0.0029	-0.0064
t-9	0.0015	0.0164	-0.0007	0.0101	-0.0006	0.0100	0.0051	0.0056
t-8	-0.0006	0.0107	-0.0021	0.0065	0.0007	0.0076	-0.0029	0.0032
t-7	-0.0015	0.0052	0.0012	0.0028	-0.0016	0.0020	0.0049	0.0077
t-6	-0.0034	-0.0014	-0.0056	-0.0034	-0.0026	0.0029	0.0007	0.0049
t-5	-0.0041	-0.0081	-0.0061	-0.0134	-0.0038	-0.0079	0.0007	0.0085
t-4	-0.0051	-0.0147	-0.0027	-0.0154	-0.0060	-0.0133	-0.0013	0.0021
t-3	-0.0001	-0.0142	0.0054*	-0.0079	0.0030	-0.0110	0.0047	0.0097
t-2	-0.0004	-0.0130	-0.0015	-0.0106	-0.0038	-0.0131	0.0062	0.0110
t-1	0.0330***	0.0233**	0.0406***	0.0356***	0.0215***	0.0110	0.0580***	0.0683
t0	0.0649***	0.0923***	0.0822***	0.1239***	0.0553***	0.0700***	0.2268***	0.2944***
t1	-0.0060	0.0914***	-0.0045	0.1221***	0.0030	0.0791***	-0.0105	0.2851***
t2	-0.0031	0.0884***	-0.0037	0.1130***	-0.0083	0.0678***	-0.0076	0.2728***
t3	-0.0074	0.0814***	-0.0073	0.1073***	-0.0031	0.0685***	-0.0144	0.2523***
t4	-0.0064	0.0421	-0.0044	0.0624	-0.0029	0.0441**	-0.0007	0.1936
t5	-0.0001	-0.0229**	-0.0050	-0.0248**	-0.0070	-0.0182	0.0093	-0.0239**
t6	-0.0128	-0.0297**	-0.0083	-0.0286**	-0.0064	-0.0277*	-0.0167**	-0.0301**
t7	0.0002	-0.0264**	-0.0065	-0.0315**	-0.0044	-0.0238**	-0.0003	-0.0228*
t8	0.0027	-0.0163**	0.0010	-0.0232*	-0.0030	-0.0237**	-0.0081	-0.0165
t9	-0.0064	-0.0164	0.0010	-0.0178	-0.0002	-0.0210**	0.0083	-0.0076
t10	-0.0048	-0.0211	-0.0036	-0.0164	0.0008	-0.0131	0.0086	-0.0083
Event	567		929		260		892	

Notes: This table shows the abnormal returns, relative to the event day *t0*, on a [-10:+10] days event window. Cumulative abnormal returns on day t are equal to the sum of abnormal returns from day t-4 to day t. ***, ** and * represent abnormal returns significance respectively at the 1%, 5%, and 10% level using a Corrado rank test. Results are presented for [1] stocks with a price greater than $0.10 and a market capitalization greater than $1,000,000, [2] stocks with a price greater than $0.01 and a market capitalization greater than $100,000, [3] stocks with a price greater than $1 and a market capitalization greater than $10,000,000, [4] all stocks listed on the OTC Pink marketplace with a price greater than $0.00001. Normal returns are computed using a market return model (benchmark NASDAQ MicroCap Index).

Conclusion générale

Le développement des nouvelles technologies entraîne une augmentation considérable du volume, de la vitesse, de la variété et de la véracité des données disponibles pour les chercheurs. Depuis près de deux décennies, les universitaires et les praticiens tentent de tirer profit de la "révolution Big Data" pour apporter de nouvelles perspectives sur des problématiques très diverses, comme la propagation de maladies grâce à des données géolocalisées de téléphone mobile (Wesolowski et al., 2012), la mesure de l'activité économique via la variation du volume de requêtes sur Google (Choi & Varian, 2012), la prévision de l'inflation grâce à l'évolution des prix sur les boutiques en ligne (Cavallo & Rigobon, 2016), ou bien encore la prédiction des résultats des élections à partir des messages envoyés sur les réseaux sociaux (Tumasjan et al., 2010).

Bien que des résultats encourageants aient été observés dans de nombreux domaines de recherche, il n'existe toujours pas de consensus au sein de la profession académique en finance quant à la valeur ajoutée de ces approches Big Data pour comprendre ou prévoir les marchés financiers. En réalité, ce débat prend racine dans l'une des principales hypothèses de la théorie classique: l'efficience informationnelle des marchés. Si les prix reflètent pleinement toute l'information disponible sur le marché, il est donc par définition impossible de battre le marché, même en utilisant des stratégies complexes d'analyse Big Data. Bien que les résultats empiriques sur ce sujet soient globalement mitigés (voir Nardo et al., 2016, pour une revue de la littérature), la disponibilité croissante de données et le développement de nouvelles techniques (voir Varian, 2014, pour une discussion sur les outils et méthodes) donnent aux chercheurs en finance la possibilité de réévaluer constamment la question de l'efficience informationnelle des marchés de manière empirique. Le développement de cette littérature est également soutenu par trois recherches récentes de Da et al. (2011), Chen et al. (2014) et Avery et al. (2016) montrant qu'il est possible de prévoir l'évolution des marchés financiers à partir de données publiées sur Internet.

À la lumière de ces résultats, cette thèse apporte de nouvelles perspectives pour mieux comprendre le processus de formation des prix sur les marchés financiers grâce à une approche Big Data. Au cours de trois essais, nous avons abordé trois problèmes distincts liés à l'efficience informationnelle des marchés financiers, en nous intéressant: (1) au rôle du sentiment des investisseurs; (2) à l'impact de l'attention des investisseurs; et (3) aux effets des manipulations informationnelles de marché. Dans chaque essai, nous avons fourni des preuves empiriques montrant que l'analyse du contenu publié sur les réseaux sociaux permet d'améliorer notre compréhension de la manière dont les marchés financiers traitent l'information et, que dans certaines circonstances, une approche Big Data permet de prévoir l'évolution des marchés financiers.

Les principales contributions empiriques de cette thèse sont les suivantes. Tout d'abord, nous avons démontré que l'évolution du sentiment des investisseurs permet de prévoir les rendements agrégés du marché boursier à l'échelle intra-journalière, en accord avec les théories de la finance comportementale. Deuxièmement, nous avons mis en évidence le fait que le degré d'attention affecte la vitesse d'intégration de l'information et l'ampleur des variations de prix, en accord avec les théories sur l'attention des investisseurs. Troisièmement, nous avons fourni des preuves empiriques montrant qu'un niveau anormal d'activité sur les médias sociaux à propos d'une entreprise à faible capitalisation est suivi d'un retournement des cours boursiers lors de la semaine suivante, en accord avec une manipulation de type informationnelle.

Dans chaque essai, nous avons également fourni des contributions méthodologiques permettant d'explorer la relation entre le contenu publié sur les réseaux sociaux et les mouvements sur les marchés financiers. Tout d'abord, nous avons proposé une méthodologie permettant de construire un indicateur du sentiment des investisseurs à l'échelle intra-journalière de manière totalement transparente, en agrégeant le sentiment des messages individuels publiés sur la plateforme StockTwits. Deuxièmement, nous avons examiné le contenu publié par les experts sur les réseaux sociaux autour de la publication d'informations économiques et financières, et nous avons proposé une mesure de similarité textuelle afin de construire de manière automatique un indicateur quantitatif de l'attention des investisseurs. Troisièmement, en nous inspirant des résultats de la théorie des réseaux, nous

avons montré qu'analyser les interactions entre les utilisateurs sur la plateforme Twitter peut permettre d'identifier automatiquement les comportements suspects pouvant se rapprocher de tentatives de manipulation de marché.

Cependant, bien que la révolution des données offre de nombreuses possibilités aux praticiens et aux universitaires, elle pose également plusieurs défis qui ne doivent pas être sous-estimés. Tout d'abord, nous pensons qu'une attention toute particulière doit être accordée afin de ne pas tomber dans l'hubris (excès de confiance) du Big Data, tel que défini par Lazer et al. (2014). L'analyse Big Data n'est pas un substitut au bon sens, à l'utilisation des théories traditionnelles, ou à la nécessité de construire soigneusement un cadre de recherche (Einav & Levin, 2014). Le nombre exponentiel de données et le large éventail de méthodologies qui peuvent être utilisées pour convertir ces données en indicateurs structurés permettent en effet aux universitaires de générer un nombre presque illimité de "séries Big Data", dont certaines, par pur hasard, aideront sûrement à la prévision d'une variable donnée sur une période spécifique. Cette situation est en fait semblable au concept de "dragage de données" (data dredging) relatif à l'utilisation d'un grand nombre de variables et d'un grand nombre de règles à des fins d'inférence ou de sélection de modèle, puis à la publication des échantillons ou périodes favorables à l'hypothèse testée. En finance, cette critique de la "recherche des anomalies" a été discutée extensivement par Fama (1998) mais semble plus que jamais d'actualité à l'heure du Big Data. Deuxièmement, et en relation avec la remarque précédente, nous pensons que la transparence des résultats est une condition *sine qua none* au développement de la recherche en Big Data. La réplication scientifique revêt bien évidemment une importance capitale dans tous les domaines de recherche, mais cette question est d'autant plus complexe dans l'environnement Big Data, étant donné le volume de données et les questions liées à la confidentialité des données. A cet égard, nous réaffirmons et étendons les propositions de King (2011), à savoir: (1) encourager le partage des données, (2) mettre à disposition tous les codes informatiques utilisés, (3) élaborer des protocoles de partage des données permettant d'assurer la protection de la vie privée tout en facilitant le travail des chercheurs sur les données sensibles, (4) construire une infrastructure ouverte (open-source) commune pour rendre l'analyse et le partage des données faciles et (5) assouplir les règles légales relatives à la collecte, au partage et à la publication de données pouvant être pertinentes pour résoudre certains problèmes

sociétaux.

Bien que les pièges et les défis ne doivent pas être sous-estimés, nous croyons néanmoins que la révolution des données est en train de créer des opportunités sans précédent et va profondément affecter la recherche en finance. De notre côté, nous allons donc poursuivre, dans les années à venir, nos recherches sur ce sujet fascinant. Nous espérons à cet égard que nos travaux permettront de mieux comprendre et de mieux mesurer les comportements humains, et pourront aider les décideurs à faire face aux grands problèmes économiques et financiers de notre siècle.

General conclusion

New technologies are leading to a tremendous increase in the volume, velocity, variety and veracity of data available for researchers. For nearly two decades, academics and practitioners have been trying to take advantage of the "Big Data revolution" to bring new perspectives on a wide range of issues, such as the propagation of disease through mobile phone location data (Wesolowski et al., 2012), the measurement of the economic activity through queries on search engines or prices on online stores (Choi & Varian, 2012; Cavallo & Rigobon, 2016), or the prediction of election results through messages sent on social media (Tumasjan et al., 2010).

While encouraging results have been observed in many fields of research, there is still no consensus in the finance academic profession about the value-added of those approaches for understanding or forecasting financial markets. In fact, this debate takes root in one of the main hypothesis of financial markets: the efficient market hypothesis. If prices fully reflect all available information in the market, it is therefore impossible to forecast financial markets, even with complex Big Data analytics strategies. Although empirical results on this topic are for now rather disappointing (see, Nardo et al., 2016, for a survey of the literature), the increasing availability of data and the development of new techniques (see, Varian, 2014, for a discussion on tools and methods) give opportunities for researchers in finance to reassess empirically the efficient market hypothesis. The development of this burgeoning literature is also supported by three recent research from Da et al. (2011), Chen et al. (2014) and Avery et al. (2016) who find that value-relevant information can be extracted from data on the Internet and helps forecasting financial markets.

In light of these recent findings, this dissertation gives new insights on the price formation process in financial markets through the use of Big Data analytics. Throughout three essays, we tackle three distinct problems related to the informational efficiency of financial markets, namely (1) the role of investor sentiment, (2) the impact of investor attention, and (3) the effect of information-based market

manipulation. In each essay, we provide empirical evidence that analyzing the content published on social media can improve our understanding of how financial markets process information, and, under certain circumstances, improve market forecasts.

The main empirical contributions of this thesis are the following. First, we demonstrate that online investor sentiment helps forecasting aggregate stock market returns at the intraday level, consistent with sentiment-driven noise trading and behavioral finance hypothesis. Second, we prove that the degree of attention to news affect the speed and the magnitude of the integration of the information into stock prices, consistent with the investor attention theory. Third, we provide empirical evidence that an abnormal level of activity on social media about a small capitalization stock is contemporaneously correlated with a large increase in stock prices and is followed by a price reversal over the next trading week, consistent with a pump-and-dump scheme.

In each essay, we also provide methodological contributions to explore the relation between the content published on social media and the patterns on financial markets. First, we propose a methodology to construct a transparent intraday investor sentiment indicator by aggregating the sentiment of individual messages posted on the platform StockTwits with explicitly revealed sentiment. Second, we examine the content published by experts on social media around the release of unscheduled news and we propose a cosine similarity measure to automatically construct a quantitative indicator of attention to news. Third, extending previous findings from the network theory literature, we recommend analyzing interactions between users on the social media platform Twitter to identify automatically suspicious behaviors.

While the data revolution offers many opportunities for the financial and the academic profession, it also poses several challenges that should not be underestimated. First, we think that researchers should take care of not falling into the big data hubris (overconfidence), as defined by Lazer et al. (2014). Big data is not a substitute for common sense, traditional theory, or the need for careful research designs (Einav & Levin, 2014). The exponential number of data and the wide range of methodology that can be used to convert those data into structured indicators allow academics to generate a nearly unlimited number of "Big Data time series", of which, by pure chance, some will help forecasting a given variable on a specific period. This situation is actually similar to Sullivan

et al. (1999) critics of data-snooping bias of technical trading rules when a given set of data is used more than once for purposes of inference or model selection, and the "dredging for anomalies" critics formulated by Fama (1998). Second, and related to the previous remark, we think that a specific attention should be granted to ensure the transparency of findings. While ensuring scientific replication is of utmost importance in all areas, the volume of data and issues related to data privacy could render the replication process more complicated in the Big Data environment. Therein, we reaffirm and extend the propositions of King (2011) to ensure the data-rich future of social sciences: (1) encourage data sharing, (2) make available all the computer codes used to construct indicators from unstructured data, (3) develop privacy-enhanced data sharing protocols to facilitate the work of researchers on sensitive data, (4) build a common open-source infrastructure that makes data analysis and sharing easy and (5) relaxing the legal rules that prevents academics to collect, share and publish data that could be relevant to solve major societal problems.

While pitfalls and challenges should not be underrated, we nonetheless believe that the data revolution is creating unprecedented opportunities for academics and will profoundly affect the research in finance. In that regards, we will continue our research in this fascinating topic in the following years, and we hope to provide several contributions to improve our understanding of human behaviors and to help decision makers tackle majors problems in economics and in finance.

Bibliography

Aggarwal, R. K. & G. Wu (2006). Stock market manipulations. *The Journal of Business 79*(4), 1915–1953.

Allen, F. & D. Gale (1992). Stock-price manipulation. *Review of Financial Studies 5*(3), 503–529.

Andersen, T. G., T. Bollerslev, F. X. Diebold, & C. Vega (2007). Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics 73*(2), 251–277.

Andrei, D. & M. Hasler (2015). Investor attention and stock market volatility. *The Review of Financial Studies 28*(1), 33.

Ang, A., A. A. Shtauber, & P. C. Tetlock (2013). Asset pricing in the dark: The cross-section of OTC stocks. *Review of Financial Studies 26*(12), 2985–3028.

Antweiler, W. & M. Z. Frank (2004). Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance 59*(3), 1259–1294.

Antweiler, W. & M. Z. Frank (2006). Do US stock markets typically overreact to corporate news stories? *Working Paper, University of British Columbia and University of Minnesota*.

Avery, C. N., J. A. Chevalier, & R. J. Zeckhauser (2016). The "CAPS" prediction system and stock market returns. *Review of Finance 20*(4), 1363–1381.

Baker, M. & J. Wurgler (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance 61*(4), 1645–1680.

Baker, M. & J. Wurgler (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives 21*(2), 129–152.

Balduzzi, P., E. J. Elton, & T. C. Green (2001). Economic news and bond prices: Evidence from the US treasury market. *Journal of Financial and Quantitative Analysis 36*(04), 523–543.

Barber, B. M. & T. Odean (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies 21*(2), 785–818.

Ben-Davis, I., F. Franzoni, A. Landier, & R. Moussawi (2013). Do hedge funds manipulate stock prices? *The Journal of Finance 68*(6), 2383–2434.

Bernile, G., J. Hu, & Y. Tang (2016). Can information be locked up? Informed trading ahead of macro-news announcements. *Journal of Financial Economics 121*(3), 496–520.

Blankespoor, E., G. S. Miller, & H. D. White (2013). The role of dissemination in market liquidity: Evidence from firms' use of Twitter. *The Accounting Review 89*(1), 79–112.

Bogousslavsky, V. (2016). Infrequent rebalancing, return autocorrelation, and seasonality. *The Journal of Finance 71*(6), 2967–3006.

Böhme, R. & T. Holz (2006). The effect of stock spam on financial markets. *Working Paper, Technische Universität Dresden*.

Bollerslev, T., J. Li, & Y. Xue (2016). Volume, volatility and public news announcements. *Working paper. Duke University, Durham*.

Bollerslev, T., V. Todorov, & S. Z. Li (2013). Jump tails, extreme dependencies, and the distribution of stock returns. *Journal of Econometrics 172*, 307–324.

Bommel, J. V. (2003). Rumors. *The Journal of Finance 58*(4), 1499–1520.

Bonner, S. E., Z.-V. Palmrose, & S. M. Young (1998). Fraud type and auditor litigation: An analysis of SEC accounting and auditing enforcement releases. *Accounting Review 73*, 503–532.

Boudoukh, J., R. Feldman, S. Kogan, & M. Richardson (2013). Which news moves stock prices? A textual analysis. *Working Paper, NBER*.

Boudt, K. & M. Petitjean (2014). Intraday liquidity dynamics and news releases around price jumps: Evidence from the DJIA stocks. *Journal of Financial Markets 17*, 121–149.

Boulland, R., F. Degeorge, & E. Ginglinger (2017). News dissemination and investor attention. *Review of Finance 21*(2), 761–791.

Bradley, D., J. Clarke, S. Lee, & C. Ornthanalai (2014). Are analysts' recommendations informative? Intraday evidence on the impact of time stamp delays. *The Journal of Finance 69*(2), 645–673.

Brandes, U., D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, & D. Wagner (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering 20*(2), 172–188.

Brown, G. W. & M. T. Cliff (2005). Investor sentiment and asset valuation. *The Journal of Business 78*(2), 405–440.

Busse, J. A. & T. C. Green (2002). Market efficiency in real time. *Journal of Financial Economics 65*(3), 415–437.

Carhart, M. M., R. Kaniel, D. K. Musto, & A. V. Reed (2002). Leaning for the tape: Evidence of gaming behavior in equity mutual funds. *The Journal of Finance 57*(2), 661–693.

Cavallo, A. & R. Rigobon (2016). The billion prices project: Using online prices for measurement and research. *The Journal of Economic Perspectives 30*(2), 151–178.

Chen, H., P. De, Y. J. Hu, & B.-H. Hwang (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies 27*(5), 1367–1403.

Choi, H. & H. Varian (2012). Predicting the present with Google Trends. *Economic Record 88*(s1), 2–9.

Comerton-Forde, C. & T. J. Putniņš (2014). Stock price manipulation: Prevalence and determinants. *Review of Finance 18*(1), 23–66.

Cookson, J. A. & M. Niessner (2016). Why don't we agree? Evidence from a social network of investors. *Working Paper, Colorado University*.

Corrado, C. J. (1989). A nonparametric test for abnormal security-price performance in event studies. *Journal of Financial Economics 23*(2), 385–395.

Da, Z., J. Engelberg, & P. Gao (2011). In search of attention. *The Journal of Finance 66*(5), 1461–1499.

Da, Z., J. Engelberg, & P. Gao (2015). The sum of all FEARS: Investor sentiment and asset prices. *Review of Financial Studies 28*(1), 1–32.

Das, S. R. (2014). Text and context: Language analytics in finance. *Foundations and Trends in Finance 8*(3), 145–261.

Das, S. R. & M. Y. Chen (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science 53*(9), 1375–1388.

De Long, J. B., A. Shleifer, L. H. Summers, & R. J. Waldmann (1990). Noise trader risk in financial markets. *Journal of Political Economy 98*(4), 703–738.

DellaVigna, S. & J. M. Pollet (2009). Investor inattention and Friday earnings announcements. *The Journal of Finance 64*(2), 709–749.

Dewachter, H., D. Erdemlioglu, J.-Y. Gnabo, & C. Lecourt (2014). The intra-day impact of communication on euro-dollar volatility and jumps. *Journal of International Money and Finance 43*, 131–154.

Diesner, J., T. L. Frantz, & K. M. Carley (2005). Communication networks from the Enron email corpus. *Computational & Mathematical Organization Theory 11*(3), 201–228.

Dimpfl, T. & S. Jank (2016). Can Internet search queries help to predict stock market volatility? *European Financial Management 22*(2), 171–192.

Dougal, C., J. Engelberg, D. Garcia, & C. A. Parsons (2012). Journalists and the stock market. *Review of Financial Studies 25*(3), 639–679.

Drake, M. S., D. T. Roulstone, & J. R. Thornock (2012). Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research 50*(4), 1001–1040.

Driessen, J., T.-C. Lin, & O. Van Hemert (2013). How the 52-week high and low affect option-implied volatilities and stock return moments. *Review of Finance*, 369–401.

Einav, L. & J. Levin (2014). Economics in the age of big data. *Science 346*(6210), 1243089.

Engelberg, J. E., A. V. Reed, & M. C. Ringgenberg (2012). How are shorts informed? Short sellers, news, and information processing. *Journal of Financial Economics 105*(2), 260–278.

Eraker, B. & M. Ready (2015). Do investors overpay for stocks with lottery-like payoffs? An examination of the returns of OTC stocks. *Journal of Financial Economics 115*(3), 486–504.

Erdemlioglu, D., S. Laurent, & C. N. Neely (2015). Which continuous-time model is most appropriate for exchange rates? *Journal of Banking and Finance 61*, 256–268.

Fama, E. F. (1965). The behavior of stock-market prices. *Journal of Business*, 34–105.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics 49*(3), 283–306.

Faust, J., J. H. Rogers, S.-Y. B. Wang, & J. H. Wright (2007). The high-frequency response of exchange rates and interest rates to macroeconomic announcements. *Journal of Monetary Economics 54*(4), 1051–1068.

Foucault, T., J. Hombert, & I. Roşu (2016). News trading and speed. *The Journal of Finance 71*(1), 335–382.

Frieder, L. & J. Zittrain (2007). Spam works: Evidence from stock touts and corresponding market activity. *Hastings Communications and Entertainment Law Journal 30*, 479.

Gao, L., Y. Han, S. Z. Li, & G. Zhou (2017). Market intraday momentum. *Working Paper, Washington University in St. Louis*.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance 68*(3), 1267–1300.

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

Groß-Klußmann, A. & N. Hautsch (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance 18*(2), 321–340.

Grossman, S. J. & J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *The American Economic Review 70*(3), 393–408.

Hanke, M. & F. Hauser (2008). On the effects of stock spam e-mails. *Journal of Financial Markets 11*(1), 57–83.

Heston, S. L., R. A. Korajczyk, & R. Sadka (2010). Intraday patterns in the cross-section of stock returns. *The Journal of Finance 65*(4), 1369–1407.

Hirshleifer, D. & S. H. Teoh (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics 36*(1), 337–386.

Hoffmann, A. O. & H. Shefrin (2014). Technical analysis and individual investors. *Journal of Economic Behavior & Organization 107*, 487–511.

Huberman, G. & T. Regev (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *The Journal of Finance 56*(1), 387–396.

Ifrim, G., B. Shi, & I. Brigadir (2014). Event detection in Twitter using aggressive filtering and hierarchical tweet clustering. In *Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014*. ACM.

Jegadeesh, N. & D. Wu (2013). Word power: A new approach for content analysis. *Journal of Financial Economics 110*(3), 712–729.

Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economics 6*(2), 95–101.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall, Englewood Cliffs, NJ.

Kearney, C. & S. Liu (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis 33*(3), 171–185.

Kim, S.-H. & D. Kim (2014). Investor sentiment from Internet message postings and the predictability of stock returns. *Journal of Economic Behavior & Organization 107*, 708–729.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science 331*(6018), 719–721.

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

Kyle, A. S. & S. Viswanathan (2008). How to define illegal price manipulation. *The American Economic Review 98*(2), 274–279.

Lazer, D., R. Kennedy, G. King, & A. Vespignani (2014). The parable of Google Flu: Traps in big data analysis. *Science 343*(6176), 1203–1205.

Leung, H. & T. Ton (2015). The impact of Internet stock message boards on cross-sectional returns of small-capitalization stocks. *Journal of Banking & Finance 55*, 37–55.

Li, J. & J. Yu (2012). Investor attention, psychological anchors, and stock return predictability. *Journal of Financial Economics 104*(2), 401–419.

Lo, A. W. (2008). Efficient markets hypothesis. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

Loughran, T. & B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance 66*(1), 35–65.

Loughran, T. & B. McDonald (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *The Journal of Economic Perspectives 17*(1), 59–82.

Mathioudakis, M. & N. Koudas (2010). Twittermonitor: Trend detection over the Twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 1155–1158. ACM.

McLean, R. D. & J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Mei, J., G. Wu, & C. Zhou (2004). Behavior based manipulation: theory and prosecution evidence. *Working Paper, New York University*.

Moat, H. S., C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, & T. Preis (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports 3*.

Nardo, M., M. Petracco, & M. Naltsidis (2016). Walking down Wall Street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys 30*(2), 356–369.

Nelson, K. K., R. A. Price, & B. R. Rountree (2013). Are individual investors influenced by the optimism and credibility of stock spam recommendations? *Journal of Business Finance & Accounting 40*(9-10), 1155–1183.

Oliveira, N., P. Cortez, & N. Areal (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems 85*, 62 – 73.

Pang, B., L. Lee, & S. Vaithyanathan (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. Association for Computational Linguistics.

Petrovic, S., M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, & L. Shrimpton (2013). Can Twitter replace newswire for breaking news? In E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, & I. Soboroff (Eds.), *ICWSM*. The AAAI Press.

Pontiff, J. (1996). Costly arbitrage: Evidence from closed-end funds. *The Quarterly Journal of Economics 111*(4), 1135–1151.

Putniņš, T. J. (2012). Market manipulation: A survey. *Journal of Economic Surveys 26*(5), 952–967.

Ranaldo, A. (2008). Intraday market dynamics around public information arrivals. In J. Wiley & H. N. Sons (Eds.), *Stock Market Liquidity: Implications for Market Microstructure and Asset Pricing*.

Ranco, G., D. Aleksovski, G. Caldarelli, M. Grčar, & I. Mozetič (2015). The effects of Twitter sentiment on stock price returns. *PloS one 10*(9).

Riordan, R., A. Storkenmaier, M. Wagener, & S. S. Zhang (2013). Public information arrival: Price discovery and liquidity in electronic limit order markets. *Journal of Banking and Finance 37*(4), 1148–1159.

Roger, P. (2014). The 99% market sentiment index. *Finance 35*(3), 53–96.

Sabherwal, S., S. K. Sarkar, & Y. Zhang (2011). Do Internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance & Accounting 38*(9-10), 1209–1237.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *The Journal of Economic Perspectives 17*(1), 83–104.

Shirky, C. (2008). Here comes everybody: The power of organizing without organizations. *New York: Penguin Press*.

Shleifer, A. & R. W. Vishny (1997). The limits of arbitrage. *The Journal of Finance 52*(1), 35–55.

Smailović, J., M. Grčar, N. Lavrač, & M. Žnidaršič (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences 285*, 181–203.

Smales, L. A. (2014). Non-scheduled news arrival and high-frequency stock market dynamics: Evidence from the Australian securities exchange. *Research in International Business and Finance 32*, 122–138.

Solomon, D. H., E. Soltes, & D. Sosyura (2014). Winners in the spotlight: Media coverage of fund holdings as a driver of flows. *Journal of Financial Economics 113*(1), 53–72.

Sprenger, T. O., P. G. Sandner, A. Tumasjan, & I. M. Welpe (2014). News or noise? Using Twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting 41*(7-8), 791–830.

Sprenger, T. O., A. Tumasjan, P. G. Sandner, & I. M. Welpe (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management 20*(5), 926–957.

Sullivan, R., A. Timmermann, & H. White (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance 54*(5), 1647–1691.

Sun, L., M. Najand, & J. Shen (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance 73*, 147 – 164.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance 62*(3), 1139–1168.

Tetlock, P. C., M. Saar-Tsechansky, & S. Macskassy (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance 63*(3), 1437–1467.

Timmermann, A. & C. W. Granger (2004). Efficient market hypothesis and forecasting. *International Journal of forecasting 20*(1), 15–27.

Tumarkin, R. & R. F. Whitelaw (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal 57*(3), 41–51.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, & I. M. Welpe (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM 10*(1), 178–185.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives 28*(2), 3–27.

Wesolowski, A., N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, & C. O. Buckee (2012). Quantifying the impact of human mobility on malaria. *Science 338*(6104), 267–270.

Wongswan, J. (2009). The response of global equity indexes to us monetary policy announcements. *Journal of International Money and Finance 28*(2), 344–365.

Yuan, Y. (2015). Market-wide attention, trading, and stock returns. *Journal of Financial Economics 116*(3), 548–564.