UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

La collaboration des internautes peut-elle influencer le marché boursier ou est-ce ce dernier qui influence les internautes ?

Analyse de discussions et sentiments sur Reddit et Twitter

Auteur : Hugo Dendievel
Promoteur : François Fouss
Année académique 2021-2022
Master d'Ingénieur de gestion - Business Analytics

Résumé

La recherche d'une relation entre le marché boursier et les réseaux sociaux est un domaine vaste et animé par diverses recherches. Le but de ce mémoire est de s'intéresser à la relation entre le cours boursier d'une action et les messages postés sur Reddit et Twitter à propos de cette dernière. Plus précisément, nous nous intéressons au sentiment des messages, c'est-à-dire, l'attitude positive, négative ou neutre du message envers l'action.

Après l'extraction des messages depuis Reddit et Twitter, 13 actions ont été sélectionnées pour les analyses de ce mémoire. Les informations financières de ces dernières ont ensuite été extraites de la plateforme Yahoo ! Finance. Une analyse de sentiments a été appliquée aux messages extraits afin de les classifier selon leur sentiment.

Des analyses statistiques ont été appliquées à l'ensemble de données final composé du nombre de messages négatifs, positifs et neutres, du volume journalier et de la variation journalière de l'action. Ces analyses consistent en des tests de corrélation entre séries temporelles, des régressions linéaires ainsi que des tests de causalité de Granger. Ces deux dernières analyses ont permis d'observer des relations significatives entre le volume journalier et/ou la variation journalière d'une action et le nombre de messages positifs, négatifs et/ou neutres. Nous observons des résultats particulièrement prometteurs entre le sentiment des messages et le volume journalier.

Les résultats nous informent aussi d'une relation bilatérale : la collaboration des internautes semble influencer le marché boursier, mais ce dernier influence aussi les messages postés par les internautes.

Finalement, une comparaison entre actions a été réalisée, laissant entrevoir de moins bons résultats dans les analyses réalisées pour les actions ayant subi beaucoup de variation durant l'année observée et semblant avoir une couverture médiatique plus importante.

Remerciements

En préambule de ce mémoire, je souhaite adresser mes remerciements à toutes les personnes y ayant contribué.

Tout d'abord, je tiens à exprimer toute ma reconnaissance à mon promoteur, monsieur François Fouss. Je le remercie d'avoir accepté de m'encadrer, de m'avoir orienté, aidé et conseillé dans l'écriture de ce mémoire.

Je remercie également les personnes ayant participé à la relecture et à la correction de ce travail.

Finalement, j'aimerais remercier particulièrement ma famille et mes proches pour leur soutien tout au long de mon parcours universitaire.

Table des matières

Liste des figures

Liste des tableaux

Liste des annexes

Introduction

En 2021, l'affaire GameStop a ébranlé le marché boursier, remettant en question l'impact des petits investisseurs parmi les grands fonds d'investissement. En effet, une forte variation du cours boursier de l'action Gamestop a été observée suite à l'initiative d'internautes d'investir fortement dans cette dernière. (Lamy, 2021 ; Le Soir, 2021 ; Raveschot et Baudoux, 2021).

Cet évènement, bien que toutefois limité, met en exergue certaines questions liées à la place de l'investisseur au sein du marché boursier. Un petit investisseur peut-il impacter le cours d'une action ? Assurément pas seul, mais un groupe de petits investisseurs peut-il influencer les marchés financiers ? Ces communautés existent à travers différentes plateformes, que ce soit Twitter, Reddit ou autres, mais leur impact n'a que peu ou pas du tout été analysé.

Ces questions posent les bases des recherches et analyses réalisées au sein de ce mémoire et dirigent alors la rédaction de ce dernier. Plus concrètement, la question de recherche de ce mémoire consiste à déterminer si les internautes peuvent influencer le marché boursier et/ou si c'est ce dernier qui les influence.

Nous avons réalisé dans un premier temps une revue de la littérature nous permettant d'obtenir une vue d'ensemble de l'état de l'art concernant l'impact social sur le marché boursier.

Ensuite, afin de répondre à la question de recherche, nous suivons la méthodologie suivante : suite à l'extraction des messages, sélection des actions à analyser et extraction des informations financières de ces dernières, nous procédons à la préparation des données. Celle-ci est une étape essentielle avant l'une des principales analyses de ce mémoire : l'analyse de sentiments. Cette dernière nous permet en effet de déterminer si l'attitude d'un message envers une action est positive, négative ou neutre. C'est à partir de cette information et plus particulièrement du volume journalier de messages positifs, négatifs ou neutres que les analyses permettant de répondre aux hypothèses de ce mémoire seront appliquées.

Ces analyses, constituant la dernière partie de ce mémoire, seront menées à bien grâce à différents tests statistiques et permettent de déterminer si une relation statistiquement significative peut-être observée entre le nombre de messages positifs, négatifs et/ou neutres et deux variables propres au marché boursier : le volume journalier et la variation journalière d'une action.

Finalement, nous concluons en définissant les limites de ce mémoire ainsi que les recommandations managériales issues de cette analyse.

Chapitre 1

Fondements théoriques

Les questions présentées lors de l'introduction de ce mémoire dirigent l'écriture des fondements théoriques, ces derniers nous permettant de comprendre le fonctionnement des plateformes de microblogging, mais aussi l'impact que peut ou pourrait avoir chaque investisseur dans un environnement social. Cette partie théorique guidera les analyses et interprétations ultérieures et permettra à tout lecteur de comprendre le contexte de ce mémoire.

Dans cette revue de la littérature, nous nous intéressons tout d'abord aux plateformes qui seront les sources d'informations primaires pour la réalisation des analyses au sein de ce mémoire : Reddit et Twitter. Nous détaillons aussi les informations liées au marché boursier que nous pouvons retrouver sur chacune de ces plateformes. Nous nous intéressons ensuite à la finance comportementale et l'impact social sur le marché financier avant de présenter les études similaires à celles voulant être réalisées dans ce mémoire. Finalement, nous nous concentrons sur les techniques qui seront utilisées lors des analyses, telles que l'extraction de données et l'analyse de sentiments.

1.1 Reddit

Reddit est un réseau social fondé par Alexis Ohanian et Steve Huffman en 2005. Historiquement, le réseau social s'ornait du slogan "the front page of the internet", se présentant alors comme une source d'informations nouvelles et populaires sur internet (Anderson, 2015). Récemment, le slogan a été remplacé par "Dive Into Anything" (« Homepage - Reddit », s. d.).

Sur cette plateforme, les *Redditors*, dénomination pour les utilisateurs membres du réseau, postent des messages contenant du texte, des liens ou des médias. Ces messages peuvent recevoir des votes positifs ou négatifs des autres utilisateurs. Les utilisateurs peuvent également faire des commentaires sur la publication initiale, ainsi que

sur d'autres commentaires. Ces commentaires reçoivent également des votes positifs et négatifs. Chaque publication est assignée à un *subreddit*, un sous-groupe discutant d'une thématique particulière, ces derniers pouvant être créés et modérés par les utilisateurs. Les membres du réseau social peuvent alors s'abonner à certains subreddits, permettant ainsi de personnaliser leur expérience (Anderson, 2015).

En janvier 2021, Reddit possédait plus de 52 millions d'utilisateurs actifs chaque jour, avec plus de 100 000 communautés et 50 milliards de vues mensuelles (« Homepage - Reddit », s. d.).

Reddit n'est cependant pas une plateforme comme les autres. La plupart des publications soumises sur Reddit sont vues par très peu de personnes, tandis qu'une petite minorité peut atteindre un large public. La page d'accueil de Reddit sert de plaque tournante pour la communauté : les articles qui y apparaissent constituent la base des discussions entre les utilisateurs, et définissent l'identité du site. La page d'accueil présente les publications qui ont été approuvées par les membres de Reddit, et ces publications peuvent être prises comme exemple par les individus qui souhaitent soumettre du contenu que la communauté appréciera. Les utilisateurs ont pris conscience de cette particularité pour poster leurs messages (Mills, 2011).

De plus, une interaction existe entre les médias traditionnels et Reddit. Cette interaction entre anciens et nouveaux médias fait également de Reddit un lieu particulièrement intéressant où différents horizons peuvent converger. Les publications en première page peuvent servir de tremplin pour des campagnes d'action collective ou pour donner un coup de pouce à une entreprise qui n'a pas forcément de budget marketing (Mills, 2011).

Alors que le système de vote de Reddit et l'accent mis sur la première page peuvent donner lieu à une pensée collective centrée sur les mêmes idées, il semble également encourager un comportement pro social chez certains utilisateurs ; le site Web a l'habitude de soutenir de bonnes causes et de reconnaître les actes d'altruisme (Mills, 2011).

Les utilisateurs de Reddit font souvent des suggestions pour améliorer le site Web ; soit par une modification de son logiciel ou un changement dans le comportement collectif de ses utilisateurs. Ces modifications sont alors réalisées avec l'approbation des utilisateurs via le système de vote ; et ils sont rapidement accompagnés de centaines ou de milliers de commentaires dont les scores offrent un bon aperçu des sentiments de la communauté sur le sujet (Mills, 2011).

Finalement, Reddit se démarque des autres réseaux sociaux en proposant la création

de communautés autour de sujets spécifiques où chacun peut soumettre son opinion et proposer des idées qui peuvent donner suite à des discussions ou des votes. Ceux-ci impactent alors la visibilité de la publication au sein de la communauté et sur la page d'accueil de la plateforme (Anderson, 2015 ; Mills, 2011).

1.1.1 Les subreddits sur le marché boursier

Il convient maintenant de s'intéresser aux principaux subreddits liés aux marchés financiers et qui partagent des informations régulièrement en lien avec ces derniers, via des publications de leurs membres. Les messages de ces subreddits permettront, après avoir été extraits de la plateforme, d'analyser, s'il y a, l'impact sur le marché boursier.

r/wallstreetbets

WallStreetBets est un subreddit consacré à la négociation d'options à haut risque (Boylston et al., 2021).

> Selon Larousse, une **option** est un "Contrat par lequel un opérateur à la Bourse des valeurs acquiert le droit (mais non l'obligation) d'acheter (option d'achat) ou de vendre (option de vente) une certaine quantité de titres à un prix donné et jusqu'à une date fixée dans le contrat. (Il existe aussi des marchés d'options sur matières premières, devises et indices.)" (Éditions Larousse, s. d.-b).

Les membres de cette communauté d'internautes se rassemblent pour parler du marché boursier, partager des mèmes, et blaguer à propos de leurs investissements (Boylston et al., 2021).

> Selon Larousse, un **mème** est défini comme un "Concept (texte, image, vidéo) massivement repris, décliné et détourné sur Internet de manière souvent parodique, qui se répand très vite, créant ainsi le buzz" (Éditions Larousse, s. d.-a).

Les utilisateurs de WallStreetBets sont connus pour acheter des options bon marché qui approchent de l'expiration et qui sont sans valeur au moment de l'achat. Plus récemment, ils se sont fait remarquer avec l'action GameStop, un détaillant de produits de jeux vidéo, en voulant réaliser un *short squeeze* (Chohan, 2021).

r/investing, r/finance, r/stocks

Les subreddits *r/investing*, *r/finance* et *r/stocks* sont des subreddits financiers faisant circuler des informations objectives et sérieuses sur les méthodes d'investissement

(Boylston et al., 2021) et principalement axés sur l'investissement sur les marchés boursiers (« Financial news and views », s. d. ; « index - investing », s. d. ; « index - stocks », s. d.). Par opposition à r/wallstreetbets, qui aborde le monde financier avec humour et légèreté, par exemple au travers de paris ou de jeux d'argent, r/investing, r/finance et r/-stocks diffusent de réels conseils d'investissement. D'ailleurs, il arrive que les utilisateurs de r/wallstreetbets se moquent des personnes qui offrent des conseils d'investissement conservateurs ou trop sérieux en leur disant de retourner sur r/investing (Boylston et al., 2021).

1.1.2 L'affaire GameStop

GameStop est un détaillant proposant des jeux vidéo et des produits de divertissement dans plus de 4 000 magasins à travers 10 pays (« About GameStop | Gamestop Corp. » s. d.). Cette entreprise a été, au début de l'année 2021, fortement mentionnée dans la presse en lien avec une forte variation de sa valeur boursière (La Libre Eco avec AFP, 2021 ; Lamy, 2021 ; Raveschot et Baudoux, 2021) suite à un phénomène de *short squeeze*, expliqué ci-après. Il est cependant nécessaire de comprendre le concept de vente à découvert pour assimiler la notion de short squeeze.

Vente à découvert

La vente à découvert est un mécanisme impliquant généralement un gestionnaire de *fonds spéculatifs*, ou *hedge funds*, qui emprunte les actions d'une entreprise à un courtier (figure 1), pour ensuite les vendre sur le marché pour réaliser une plus-value. Ceci est fait dans l'espoir que le cours de l'action chutera suite à un évènement quelconque. Comme la figure 1 l'indique, quand la valeur des actions diminue, le gestionnaire de fonds spéculatifs peut racheter le montant des actions empruntées au courtier à une valeur inférieure et les restituer au courtier, réalisant ainsi un profit (Di Muzio, 2021). La vente à découvert représente alors l'achat, de ce que l'on appelle, des positions courtes (Chohan, 2021).

Les **fonds spéculatifs** ou **hedge funds** sont des pools d'investissement dirigés par des gestionnaires qui utilisent un large éventail de stratégies, dont souvent l'achat d'actifs avec de l'argent emprunté et la négociation d'actifs difficilement accessibles, afin de battre les rendements moyens des investissements pour leurs clients. Ils sont considérés comme des choix d'investissement alternatif risqués (The Investopedia Team, 2021).

La vente à découvert peut ainsi être imaginée en 3 étapes :

1. Le gestionnaire de fonds spéculatifs emprunte au courtier un certain nombre d'actions de la société X d'une certaine valeur, disons 100$. Le gestionnaire les vend immédiatement. Ce dernier s'engage aussi à remettre au courtier les actions empruntées à un moment précis et paiera des frais de transaction et des intérêts au courtier (figure 1) ;
2. Le gestionnaire attend et espère la baisse du cours de l'action. Par exemple, si cette dernière descend à 50$, il pourra racheter les actions deux fois moins chères. Cependant, si le prix augmente, il se retrouverait face à une possible perte d'argent, comme indiqué sur la figure 1 ;
3. Le gestionnaire de fonds spéculatifs rend les actions au courtier et perçoit un profit lié à la baisse du prix de l'action. Cela peut aussi être une perte si le cours de l'action n'a pas suivi les prédictions du gestionnaire.

(Di Muzio, 2021).

Figure 1 – Vente à découvert (Alliance Bernstein, 2013)

Short squeeze

Le *short squeeze* est un scénario dans lequel une augmentation initiale du prix d'une action incite les vendeurs à découvert à racheter leurs actions pour éviter une perte trop

élevée, ce qui fait de nouveau monter le prix (Xu et Zheng, 2016). En effet, un short squeeze pourrait entraîner une perte éventuellement infinie pour le vendeur à découvert si le prix augmente indéfiniment (Chohan, 2021).

Plus en détail, un short squeeze se produit lorsque le prix d'une action augmente très vite et promptement, comme présenté sur la figure 2, au-delà de ce que les analystes et les acteurs du marché boursier avaient prévu. Les short squeeze peuvent très fortement frapper les investisseurs qui vendent à découvert avec des actions empruntées, car ils pourraient finir par dépenser plus d'argent que prévu pour racheter et restituer les actions empruntées (Cliffe, s. d.).

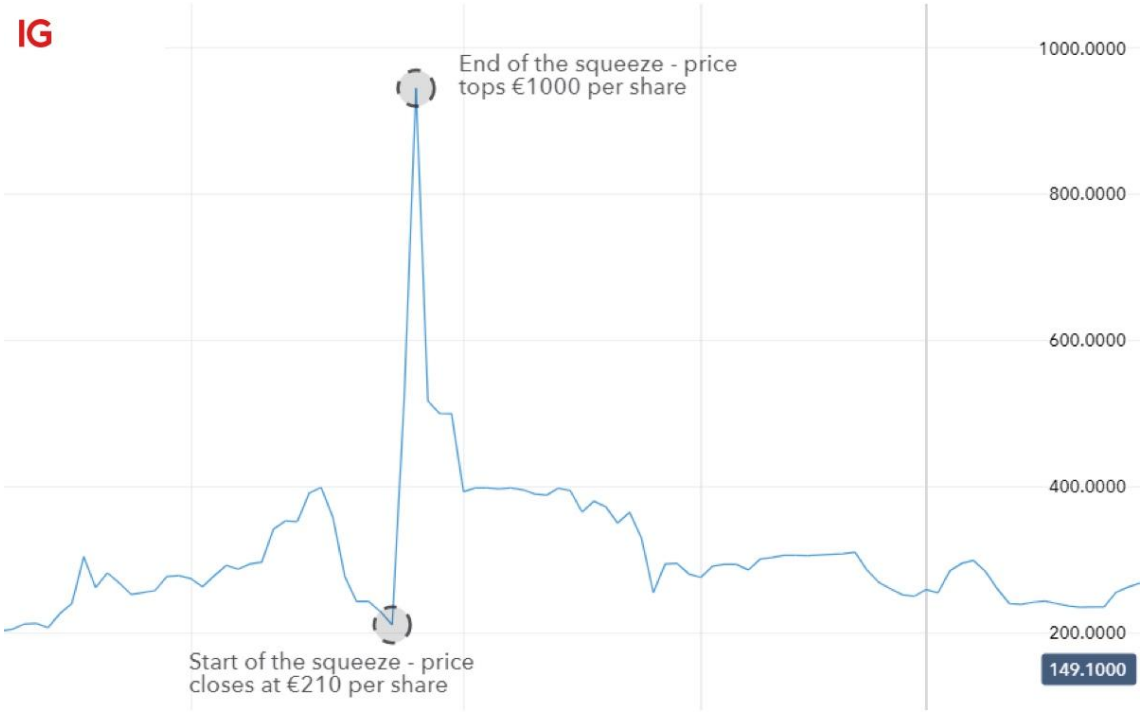

Figure 2 – Exemple de short squeeze (Cliffe, s. d.)

Déroulement

GameStop (GME sur le marché boursier) est une entreprise cotée en bourse qui voyait diminuer ses bénéfices, principalement en raison de l'augmentation des ventes en ligne et de la digitalisation (Vasileiou et al., 2021). Cette baisse de performance se traduisit par une baisse du prix de l'action, passant de 22,73$ le 3 janvier 2017 à 5,43$ le 6 janvier 2020 (« GME Interactive Stock Chart | GameStop Corp. Stock - Yahoo Finance », s. d.).

Avec l'arrivée de la pandémie du coronavirus, les potentiels acheteurs se sont tournés vers les achats en ligne suite aux mesures sanitaires, faisant de nouveau baisser le prix de l'action (Vasileiou et al., 2021), pour atteindre 2,80$ le 30 mars 2020 (« GME Interactive Stock Chart | GameStop Corp. Stock - Yahoo Finance », s. d.).

En effet, la crise du coronavirus a eu un impact négatif sur les activités et la performance financière de GameStop. L'entreprise s'attendait alors à une baisse de l'affluence dans leurs magasins ou à un impact sur la chaine d'approvisionnement (« Annual Reports | Financial Information | Investor Relations | Gamestop Corp. » s. d.). Sur base de ces perspectives négatives pour l'entreprise, les fonds spéculatifs ont vendu l'action à découvert. La vente à découvert n'est cependant pas une nouvelle stratégie. De nombreuses actions sont vendues à découvert et les positions courtes moyennes représentent environ 5% des actions en circulation (Vasileiou et al., 2021).

Ce qui rend la situation particulière dans le cas de GameStop, c'est que le pourcentage d'actions vendues à découvert par rapport au total des actions publiquement disponibles est resté proche de 100%. L'affaire GameStop est aussi le premier cas largement connu pour lequel un grand nombre de petits investisseurs, provenant de la communauté r/wallstreetbets sur Reddit, se sont opposés aux grands fonds spéculatifs (Vasileiou et al., 2021).

En effet, la participation des internautes de r/wallstreetbets peut s'analyser de la façon suivante :

— Tout d'abord, des fonds d'investissement américains ont tenté de gagner de l'argent, au travers de ventes à découvert, en pariant contre l'entreprise GameStop, prévoyant une baisse du cours de son action ;
— Leur stratégie n'a cependant pas fonctionné, car le prix de l'action a par la suite augmenté grâce aux investisseurs et à des résultats meilleurs qu'espérés. Un peu plus tard, en début d'année 2021, un analyste de Wall Street a assuré que l'action GameStop allait tout de même perdre de la valeur ;
— Les utilisateurs du subreddit r/wallstreetbets ont cependant décidé de lui donner tort en achetant en masse des actions GameStop.

(Lamy, 2021).

Les utilisateurs de r/wallstreetbets, présentés dans la section 1.1.1, ont alors mis en commun leurs énergies financières pour saboter les positions courtes détenues par de puissants hedge funds. De ce mouvement résulte une vente des positions courtes par les fonds d'investissement et une flambée du cours de l'action (Di Muzio, 2021). L'action GME a vu son cours en bourse passer de 18,80$ le 31 décembre 2020 à 483$ le 28 janvier 2021 (Vasileiou et al., 2021), impliquant une perte d'argent pour les fonds d'investissement ayant vendu à découvert l'action (Di Muzio, 2021).

1.2 Twitter

Twitter est un réseau social sur lequel les utilisateurs partagent leurs pensées, leurs actualités et leurs idées (Rigolin, 2018) en postant des messages de 280 caractères maximum (140 historiquement) (Gligorić et al., 2018). Les utilisateurs se suivent pour rester informés ou pour communiquer avec d'autres individus ou groupes (Rigolin, 2018). Ces relations sont cependant non obligatoirement réciproques entre les utilisateurs. Ces utilisateurs peuvent en effet suivre ou être suivis par quelqu'un sans que ce lien soit réciproque (Kwak et al., 2010).

Twitter a été lancé en 2006 par les cofondateurs Jack Dorsey, Noah Glass, Biz Stone et Evan Williams. L'idée originale était de créer une plateforme de communication permettant à un groupe d'amis de se suivre. Ce service d'information et de réseautage a depuis explosé en popularité (Rigolin, 2018). Au deuxième trimestre 2021, Twitter comptait 206 millions d'utilisateurs actifs quotidiennement dans le monde (Statista Research Department, 2021).

Les messages postés par les utilisateurs sont appelés *tweets* et sont à destination des *followers*, individus qui suivent le publicateur dudit tweet. Ces messages peuvent être ensuite *retweetés*, c'est-à-dire, partagés par un utilisateur à ses followers. Le contenu des messages est généralement accompagné de mentions d'utilisateurs en utilisant l'arobase @ suivie du nom d'utilisateur ou de hashtags # suivis de mots caractérisant ou accompagnant le thème du message (Kwak et al., 2010).

1.2.1 Les cashtags et le marché boursier

L'utilisation de *cashtags* sur Twitter, de manière analogue aux hashtags, permet de désigner un thème financier dans un tweet. Les cashtags sont des symboles boursiers précédés du signe dollar *$*. Par exemple, pour tweeter sur l'action Apple, le cashtag $\$AAPL$ est utilisé. Twitter étant l'une des sources les plus importantes d'informations en temps réel sur Internet, cela en fait une excellente plate-forme pour diffuser des informations rapidement à propos, entre autres, du marché boursier (Hentschel et Alonso, 2014). Ces symboles sont fréquemment utilisés par les utilisateurs et peuvent être source d'informations sur le marché en analysant, par exemple, le nombre de tweets contenant un cashtag spécifique (Oliveira et al., 2017).

Les tweets partageant des informations pertinentes peuvent être obtenus en recherchant sur le site Web de Twitter des mots-clés spécifiques (Yang et al., 2015), comme discuté précédemment.

Certaines communautés, dont les discussions sont centrées autour du marché boursier, peuvent aussi être observées sur Twitter. Ainsi, selon Yang et al., 2015, qui ont réalisé une étude s'intéressant à ce type de communauté sur Twitter, une régression du sentiment du marché boursier met en évidence l'influence significative de *nœuds critiques* sur les mouvements du marché financier. Grâce à la comparaison des performances, le sentiment exprimé par les noeuds principaux est plus significatif et impactant que ceux des autres membres de la communauté (Yang et al., 2015).
Ces communautés présentent donc bien un effet, bien que plus limité pour les noeuds non critiques.

Les **nœuds** dits **critiques** représentent un ensemble de nœuds d'un réseau dont la suppression entraîne une fragmentation maximale de ce dernier (Sarker et al., 2019).

Finalement, une forte corrélation a été observée entre l'humeur sociale et le mouvement des prix des actifs des marchés financiers, en se basant sur la communauté financière du réseau social. Le sentiment Twitter généré par les nœuds critiques de la communauté financière fournit un indicateur fiable pour prédire les mouvements des marchés financiers (Yang et al., 2015).

1.3 Finance comportementale

Depuis un certain nombre d'années, il a été démontré que le marché boursier est guidé par la psychologie des investisseurs (Baker et Wurgler, 2007 ; Nofer et Hinz, 2015). Les investisseurs sont en effet sujets à des erreurs ou au moins à des décisions émotionnelles. Certains effets ont pu être observés, comme l'effet de janvier, qui indique que les retours sont en moyenne plus élevés en janvier par rapport aux autres mois de l'année ; l'effet lundi, qui implique que les retours du lundi sont relativement faibles par rapport à ceux du vendredi précédent, ou encore, l'effet momentum, qui implique que les anciens gagnants ou perdants continuent à bien ou mal performer (Nofer et Hinz, 2015). Certains de ces effets semblent cependant être remis en question depuis quelques années maintenant (Jaffe et al., 1989 ; Patel, 2016 ; Wang et al., 1997).

Les chercheurs fournissent différentes explications à ces anomalies qui sont motivées par des sentiments et des émotions. Les chercheurs en finance comportementale se réfèrent à deux groupes d'investisseurs : les *rational arbitrageurs*, qui sont des investisseurs bien informés et qui ne sont pas enclins au sentiment, et les *noise traders*, qui s'appuient de manière irrationnelle sur le sentiment et d'autres informations non fondamentales (Shleifer et Summers, 1990). Les *noise traders* suivent les tendances et réagissent souvent de ma-

nière excessive ou insuffisante aux nouvelles (Nofer et Hinz, 2015). Les membres du subreddit r/wallstreetbets, présentés précédemment dans la section 1.1.1 pourraient être associés à ce type d'individus : les actions achetées par ces individus sont celles les plus discutées ou populaires au sein du subreddit et ces achats sont généralement réalisés sans analyse profonde du marché.

Les individus soutenant l'hypothèse du marché efficient expliquent que les *rational arbitrageurs* négocient contre les *noise traders*, ce qui permet de ramener les prix aux valeurs fondamentales après des chocs exogènes. Les premiers influencent donc les prix pendant une très courte période de temps avant que les deuxièmes ne prennent position jusqu'à ce que l'équilibre de marché soit atteint. Les chercheurs en finance comportementale ont cependant démontré que le pouvoir des *rational arbitrageurs* contre les *noise traders* reste limité (Nofer et Hinz, 2015).

D'autres chercheurs présentent un autre fonctionnement : de plus en plus de *noise traders* pourraient suivre d'autres *noise traders* lors de l'achat ou de la vente d'actions. On peut de nouveau penser que ce type de comportement pourrait se dérouler au sein du subreddit r/wallstreetbets. De cette manière, les *noise traders* achètent ou vendent en cas de hausse ou baisse des prix. Les *rational arbitrageurs* peuvent alors anticiper le comportement des *noise traders*. Il existe cependant d'autres facteurs qui limitent la capacité des investisseurs rationnels à négocier contre les investisseurs individuels mal informés. Étant donné que les investisseurs rationnels sont pour la plupart averses aux risques, le risque fondamental peut par exemple également empêcher les arbitragistes de négocier pendant un certain temps (Nofer et Hinz, 2015).

On observe ici que le sentiment peut donc influencer les cours des actions en cas d'arbitrage limité.

1.4 L'impact social sur le marché boursier

Afin de prendre une décision importante, les individus se basent généralement sur des informations issues de leur entourage, comme leurs proches ou leurs amis. Une de ces décisions peut être d'interagir avec le marché boursier. Cette participation au marché boursier par un grand nombre d'individus améliore l'efficacité de l'allocation des ressources, facilite le développement financier et provoque ainsi la croissance économique. Cette large participation pourrait cependant avoir un impact sur les rendements globaux (Liang et Guo, 2015).

L'interaction sociale pourrait donc servir de canal de diffusion pour les informations

liées au marché et influencer ainsi la participation des ménages au marché boursier. Cependant, les informations transmises par l'interaction sociale sont généralement biaisées : la prise de décision pourrait entraîner des croyances différentes entre les investisseurs, résultant en des fluctuations importantes des marchés financiers (Liang et Guo, 2015).

Il est donc important de s'intéresser à l'impact de l'interaction sociale sur la participation au marché boursier. Cet impact a été étudié à travers une étude empirique en Chine, réalisée par Liang P. et Guo, S., en se basant sur plus de 8000 ménages chinois.

Cet échantillon représentatif des ménages chinois a été utilisé pour étudier l'effet informationnel et l'effet multiplicateur des interactions sociales sur la participation au marché boursier en Chine. L'effet informationnel représente la communication de bouche à oreille qui permet aux investisseurs de facilement s'informer en parlant avec des pairs. L'effet multiplicateur, lui, représente l'influence par le comportement des individus de la communauté. Une personne est ainsi plus susceptible de participer au marché boursier si les individus qui l'entourent participent au marché boursier (Liang et Guo, 2015).

Les résultats de l'étude permettent de constater que l'accès à Internet et les interactions sociales augmentent la participation au marché boursier, mais ces deux canaux d'information se substituent l'un à l'autre. Cela suggère que les outils de communication modernes pourraient évincer l'effet informationnel des interactions sociales. L'effet marginal des interactions sociales, quant à lui, diminue si le ménage a accès à Internet. Il en ressort aussi que l'effet marginal des interactions sociales sur la participation au marché boursier est plus important dans les communautés à forte participation, soutenant ainsi l'effet multiplicateur social (Liang et Guo, 2015).

1.5 La recherche sur l'impact du microblogging sur le marché boursier

Avant de s'intéresser à l'impact du microblogging sur le marché boursier, il est important de définir ce qu'est le microblogging. Ce dernier est défini comme le fait de poster des messages courts sur une plateforme à destination des internautes (Cambridge Dictionary, s. d.). Twitter et Reddit font partie de ces plateformes de microblogging (Anderson, 2015 ; Kwak et al., 2010). Les papiers scientifiques recherchant un lien entre le microblogging et le marché boursier sont principalement basés sur l'analyse de messages issus de Twitter. En effet, peu d'études s'intéressent à l'analyse de messages issus de Reddit, et encore moins, de messages liés au marché boursier. L'analyse de l'état de l'art se fera donc en se basant principalement sur des études à propos de Twitter.

1.5.1 Données

Les données utilisées au sein de ces différentes études dépendent des méthodes déployées, mais les données primaires sont toujours des messages extraits de Twitter ou Reddit grâce à leurs *API* respectifs (Foufi et al., 2019 ; Nisar et Yeung, 2018 ; Nofer et Hinz, 2015 ; Oliveira et al., 2017 ; Ranco et al., 2015 ; Talamás, 2021 ; Yang et al., 2015).

Une **API** est un ensemble de règles et de protocoles permettant de créer et d'intégrer des logiciels. L'API permet alors à un produit ou un service de communiquer avec d'autres produits ou services en faisant office d'interface pour échanger des informations (« What is an API ? », 2017). Twitter et Reddit proposent tous les deux des API permettant de communiquer avec la plateforme afin d'extraire un certain nombre d'informations, dont les messages postés par les utilisateurs ainsi que des caractéristiques associées à ces derniers (date et heure du message, popularité...) (« reddit.com : documentation sur l'API », s. d. ; « Twitter API Documentation », s. d.).

Des données issues du marché boursier, telles que les prix d'ouverture et de clôture, le volume de transactions ou encore le rendement de certaines actions ou indices, sont aussi extraites et utilisées au cours des analyses (Nisar et Yeung, 2018 ; Nofer et Hinz, 2015 ; Oliveira et al., 2017 ; Ranco et al., 2015 ; Talamás, 2021 ; Yang et al., 2015).

- Le **prix d'ouverture** représente la gamme de prix auxquels les premières offres sont faites ou les premières transactions sont conclues sur un marché boursier (« Opening price Definition », s. d.).

- Le **prix de clôture** représente le prix de la dernière transaction d'une action effectuée au cours d'une séance de bourse d'une journée sur le marché boursier (« Closing price Definition », s. d.).

- Le **volume de transactions** représente le nombre d'actions négociées chaque jour (« Trading volume Definition », s. d.).

- Le **rendement** représente la variation de la valeur d'un portefeuille au cours d'une période d'évaluation (« Return Definition », s. d.).

1.5.2 Analyses

Deux types de méthodes se retrouvent au sein des analyses dans les différents papiers scientifiques. Certains se concentrent majoritairement sur des calculs statistiques tandis que d'autres se penchent principalement sur des méthodes d'apprentissage automatique, mais ces deux types d'analyses se retrouvent généralement appliquées conjointement. (Nisar et Yeung, 2018 ; Nofer et Hinz, 2015 ; Oliveira et al., 2017 ; Ranco et al., 2015 ; Talamás, 2021 ; Yang et al., 2015)

Méthodes statistiques

Différentes méthodes statistiques sont utilisées afin de faire ressortir des liens entre des messages à propos du marché boursier et les mouvements sur ce dernier. Des analyses de séries chronologiques sont notamment réalisées (Talamás, 2021).

Selon Talamás, 2021, une *méthode des moindres carrés* est appliquée sur des données de panel afin d'observer une relation entre des messages issus de Reddit et les mouvements de trois actions : AMC, GME et NOK. Les résultats ne sont cependant pas significatifs. Les analyses ultérieures s'intéressent à la méthode *ARIMA* qui est appliquée à l'action GME, mais les résultats ne sont pas concluants. Le modèle est, entre autres, incapable de prédire de grandes variations du cours de l'action, même en relâchant la confiance des prédictions (Talamás, 2021).

Une **ARIMA**, ou moyenne mobile intégrée autorégressive, est un modèle d'analyse statistique utilisant les données de séries chronologiques afin de comprendre l'ensemble de données ou prédire les tendances futures. Un modèle statistique est dit autorégressif s'il prédit des valeurs futures sur base de valeurs passées.

Ce modèle statistique est une forme de régression dont le but est d'évaluer la force d'une variable dépendante par rapport à d'autres variables indépendantes. L'un des objectifs du modèle peut être de prédire les mouvements futurs d'actions en examinant les différences entre les valeurs de la série.

Le terme ARIMA représente :

- **Autorégression (AR)**, qui fait référence à un modèle qui montre une variable qui régresse sur ses propres valeurs décalées ou antérieures ;
- **Intégré (I)**, qui représente la différenciation des observations pour permettre à la série chronologique de devenir stationnaire, autrement dit, les valeurs des données sont remplacées par la différence entre les valeurs des données et les

valeurs précédentes ;
— **Moyenne mobile (MA)**, qui incorpore la dépendance entre une observation et une erreur résiduelle d'un modèle de moyenne mobile appliqué aux observations retardées.

(Hayes, 2021a).

Une *régression linéaire* est ensuite réalisée, ainsi qu'une *régression polynomiale*. Les R^2 mesurés ne permettent cependant pas de conclure (Talamás, 2021).

La **régression linéaire**, ou régression des moindres carrés ordinaires, est une méthode d'analyse statistique qui estime la relation entre une ou plusieurs variables indépendantes et une variable dépendante. La relation, linéaire, est estimée en minimisant la somme des carrés de la différence entre les valeurs observées et prédites, grâce au modèle, de la variable dépendante. Cette relation est alors représentée par une droite (Poston, s. d.).

La **régression polynomiale**, quant à elle, est une régression s'intéressant à une relation non linéaire entre les variables dépendantes et indépendantes. La variable dépendante est ici liée à la variable indépendante ayant un nième degré. Il n'est pas nécessaire que la relation entre les variables dépendantes et indépendantes soit linéaire, nous obtenons ainsi une courbe qui représente la relation entre les données, à la place d'une droite (Pedamkar, 2019).

Le $\mathbf{R^2}$ est une mesure statistique qui représente la proportion de la variance d'une variable dépendante qui est expliquée par une ou plusieurs variables indépendantes dans un modèle de régression. Ainsi, si le R^2 d'un modèle est de 0.8, 80% de la variation observée peut être expliquée par le modèle (Fernando, 2021).

D'autres réalisent une procédure du filtre de Kalman pour créer un indicateur de sentiment quotidien unique à partir de Twitter. Le filtre de Kalman permet la production d'un indicateur de sentiment quotidien en combinant diverses mesures de sentiment de fréquences diverses. L'indicateur qui en résulte semble plus représentatif du sentiment général des investisseurs. Cette méthode permet d'informer sur la prédiction des rendements de certains portefeuilles et indices. Cependant, les indicateurs du filtre de Kalman sont moins efficaces pour la prévision du volume des transactions et de la volatilité (Oliveira et al., 2017).

Un **filtre de Kalman** est un algorithme qui prend des données d'entrée de plusieurs sources et estime des variables inconnues, malgré un niveau de bruit potentiellement élevé. Le filtre de Kalman a l'avantage de pouvoir prédire des valeurs inconnues avec plus de précision que si les prédictions individuelles étaient faites à l'aide de méthodes de mesure singulières (« Kalman Filter », 2019).

Le coefficient de *corrélation de Pearson* se retrouve également au sein de certaines analyses, notamment pour mesurer la dépendance linéaire entre la polarité des sentiments issus de tweets à propos d'actions et le rendement de ces dernières. Un test de causalité de Granger est aussi réalisé pour vérifier si des variables issues de Twitter, comme le nombre de tweets par jour, ou encore le nombre de tweets négatifs, neutres ou positifs par jour peuvent aider à la prédiction des rendements (Ranco et al., 2015).

"Le **coefficient de Pearson** est un indice reflétant une relation linéaire entre deux variables continues." (« Corrélation de Pearson », s. d.).

Le **test de causalité de Granger** est un test d'hypothèse statistique permettant de déterminer si une série chronologique est utile pour en prévoir une autre (Wei, 2016).

Il en ressort que la variable de polarité des sentiments n'est pas utile pour prédire le rendement des prix. Cependant, suite au test de Granger, on observe que la quantité d'attention sur Twitter est utile pour prédire la volatilité des prix (Ranco et al., 2015).

Méthodes d'apprentissage automatique

Des méthodes d'apprentissage automatique sont appliquées au sein de différentes publications scientifiques dans le but de rechercher des relations entre des messages à propos d'actions et les mouvements de ces dernières. La principale méthode appliquée et utilisée comme base d'autres analyses est l'analyse de sentiments. Cette dernière sera discutée en détail par la suite (section 1.7), mais il convient de s'intéresser à l'application de cette technique au sein des différents papiers scientifiques.

Ainsi, selon Oliveira et al., 2017, une analyse de sentiments a été réalisée à partir des données de Twitter sur le marché boursier. Afin de créer les indicateurs de sentiment des investisseurs, les scores de sentiment produits par l'analyse de sentiments ont été utilisés. Cette dernière se base sur un lexique récent adapté aux conversations de microblogging sur le marché boursier et accessible librement. Ce lexique contient environ 7000 unigrammes,

13000 bigrammes et les scores de sentiment respectifs pour les contextes positifs et négatifs (Oliveira et al., 2017).

Dans le cadre du **Natural Language Processing**, c'est-à-dire, une branche du machine learning qui consiste en la compréhension, la manipulation et la génération du langage naturel par les machines, et dont fait partie l'analyse de sentiments (Lina, 2020), le concept de **n-gramme** est important.

Le n-gramme représente une séquence de n éléments générés à partir d'un échantillon de texte. Ainsi, si l'on considère une phrase "Either my way or no way", il est possible de générer les n-grammes suivants (Fardeen, 2021) :

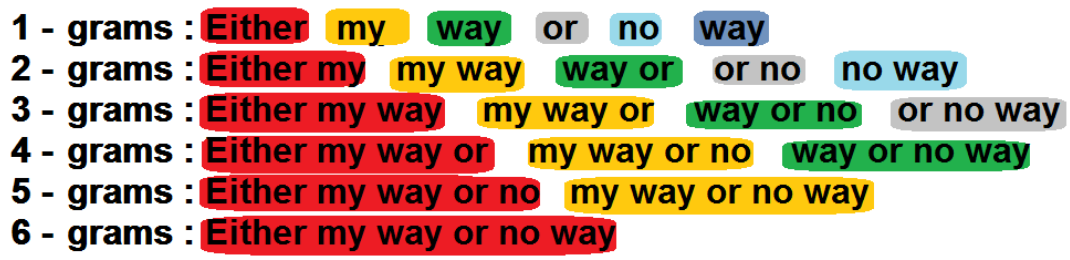

Figure 3 – N-grammes de la phrase "Either my way or no way" (Fardeen, 2021)

De cette manière, un score négatif indique un mot faisant penser à un contexte baissier sur le marché boursier et un score positif indique un contexte haussier. À partir de ce lexique, une analyse de sentiments a été réalisée sur chaque tweet et le score de sentiment de chaque tweet correspond à la somme du score de tous les éléments du lexique présents dans le message (Oliveira et al., 2017).

Ranco et al., 2015 s'intéressent aussi à l'analyse de sentiments. Ainsi, 100 000 tweets ont été annotés manuellement par des experts afin d'obtenir une base d'entrainement solide. De plus, environ 6 000 tweets ont été annotés deux fois par des experts différents afin de s'assurer de la qualité d'annotation. Les différentes étapes appliquées lors de l'analyse de sentiments sont les suivantes :

1. Un échantillon de tweets est annoté manuellement avec le sentiment ;
2. L'ensemble étiqueté est utilisé pour former et régler un classificateur ;
3. Le classificateur est évalué par validation croisée ;
4. Le classificateur est appliqué à l'ensemble des tweets collectés.

Le sentiment des tweets est ensuite approximé avec une échelle ordinale de trois valeurs : négatif, neutre ou positif (Ranco et al., 2015). Cette analyse de sentiments, bien

que déjà complexe, représente un input utilisé ensuite pour s'intéresser à la relation entre ces messages et le cours boursier des actions analysées. Des calculs statistiques, comme discutés précédemment, sont entre autres réalisés sur base des résultats, c'est-à-dire, l'annotation automatique des tweets comme négatifs, positifs ou neutres et qui sont liés à une action particulière (Nisar et Yeung, 2018 ; Ranco et al., 2015 ; Yang et al., 2015).

Oliveira et al., 2017 s'intéressent à des techniques d'apprentissage automatique appliquées aux résultats de l'analyse de sentiments. Un réseau de neurones, un SVM et une forêt aléatoire sont entre autres utilisés.

Les **réseaux de neurones** représentent un sous-ensemble de l'apprentissage automatique. Leur nom et leur structure sont inspirés du cerveau humain, imitant la façon dont les neurones biologiques fonctionnent. Les réseaux de neurones sont constitués de 3 types de couches de neurones (figure 4) : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie.

Chaque noeud, ou neurone artificiel, se connecte à un autre et possède un poids et un seuil. Si la sortie d'un nœud est supérieure à la valeur du seuil, le noeud s'active en envoyant des données à la couche suivante du réseau. Les réseaux de neurones s'appuient alors sur des données d'entraînement pour apprendre et améliorer leur précision au fur et à mesure qu'ils s'entrainent.

(« What are Neural Networks ? », 2021).

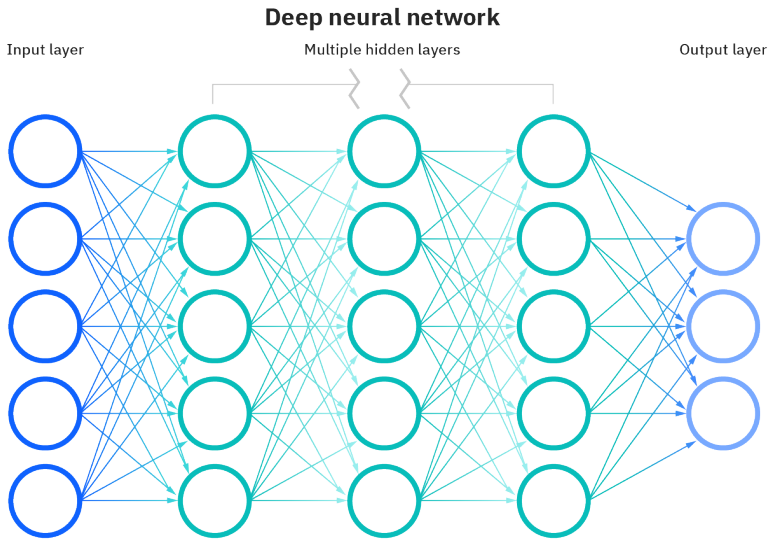

Figure 4 – Représentation d'un réseau de neurones (« What are Neural Networks ? », 2021)

Une **forêt aléatoire** est avant tout composée d'arbres de décision. L'arbre de décision est une structure hiérarchique qui est construite en utilisant les caractéristiques d'un ensemble de données. Chaque nœud de l'arbre de décision est divisé selon une mesure associée à un sous-ensemble des caractéristiques (figure 5) (Caie et al., 2021).

La forêt aléatoire, elle, est une collection d'arbres de décision, chacun associé à un ensemble d'échantillons générés à partir de l'ensemble de données d'origine. Les nœuds sont eux divisés en fonction de l'entropie ou de l'indice de Gini (Caie et al., 2021). L'algorithme effectue alors un apprentissage en parallèle sur plusieurs arbres de décision construits de manière aléatoire et entraînés sur des sous-ensembles de données différents (« Random Forest », s. d.).

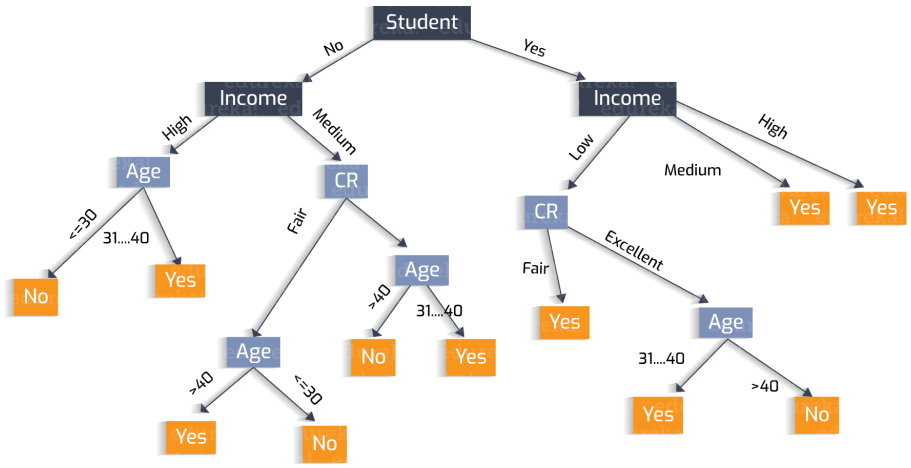

Figure 5 – Représentation d'un arbre de décision (Sharma, 2021)

Des trois méthodes abordées précédemment, le SVM semble proposer les résultats les plus intéressants avec des p-value inférieures à 5% pour certaines analyses. Des résultats significatifs sont donc observés en utilisant une de ces méthodes d'apprentissage automatique (Oliveira et al., 2017).

Finalement, selon Talamás, 2021, parallèlement aux méthodes statistiques présentées précédemment, un *SVR* est aussi appliqué afin d'observer une relation entre des messages issus de Reddit et les mouvements de trois actions : AMC, GME et NOK. Bien que les résultats ne soient pas significatifs (Talamás, 2021), il convient d'expliquer le SVR, lui-même basé sur le principe du SVM.

Un **SVR**, ou Support Vector Regression, est un algorithme de régression qui prend en charge les régressions linéaires et non linéaires. Cette méthode fonctionne sur le principe du Support Vector Machine, ou SVM (Pedamkar, 2020).

Avant de définir en détail un SVM, il est nécessaire de définir certains termes :

- Le **kernel** représente une fonction qui sera utilisée pour représenter des données de dimension inférieure dans une dimension supérieure ;
- L'**hyperplan** représente la ligne de séparation entre les classes de données dans le cadre d'un SVM (figure 6). Dans le SVR, il représente plutôt la ligne aidant à prédire les différentes valeurs ;
- Les **frontières** sont les deux lignes autres que l'hyperplan qui créent une marge autour de ce dernier (figure 6). Les vecteurs de support peuvent se positionner sur les frontières ou à l'extérieur de celles-ci ;
- Les **vecteurs de support** sont les points de données les plus proches des frontières (figure 6).

(Bhattacharyya, 2020).

Habituellement, un algorithme d'apprentissage tente d'apprendre les caractéristiques les plus communes d'une classe, le SVM fonctionne dans l'autre sens. Il trouve les exemples les plus similaires entre les classes. Si on considère des pommes et des citrons, un algorithme traditionnel chercherait les points communs entre toutes les pommes qui sont vertes et arrondies et entre tous les citrons qui sont jaunes et ont une forme elliptique. En revanche, le SVM va rechercher les pommes qui sont très similaires aux citrons, par exemple les pommes qui sont jaunes et ont une forme elliptique. Il s'agira alors d'un vecteur de support (figure 6). L'autre vecteur de support sera un citron similaire à une pomme (Zoltan, 2021).

L'objectif est alors de séparer au mieux les échantillons de classes différentes par un hyperplan en maximisant la marge autour de ce dernier (figure 6). La marge étant alors caractérisée par des frontières. Les données ne sont cependant pas toujours linéairement séparables, c'est donc là qu'intervient le kernel, permettant de les représenter dans une dimension supérieure afin de trouver une façon de séparer ces données (Versloot, 2019).

Le SVR, lui, diffère du SVM, car ce dernier est un classificateur utilisé pour prédire des catégories discrètes, tandis que le SVR est un régresseur utilisé pour prédire des variables ordonnées continues. Dans le cas du SVR, on essaye de déterminer la meilleure droite ou courbe représentant nos données (figure 7). Dans la régression

simple, l'idée est de minimiser le taux d'erreur tandis que dans le SVR, l'idée est d'ajuster l'erreur à l'intérieur d'un certain seuil ε (Pedamkar, 2020). Autrement dit, le but est de trouver une droite ou courbe qui maximise le nombre de points au sein de la marge (figure 7).

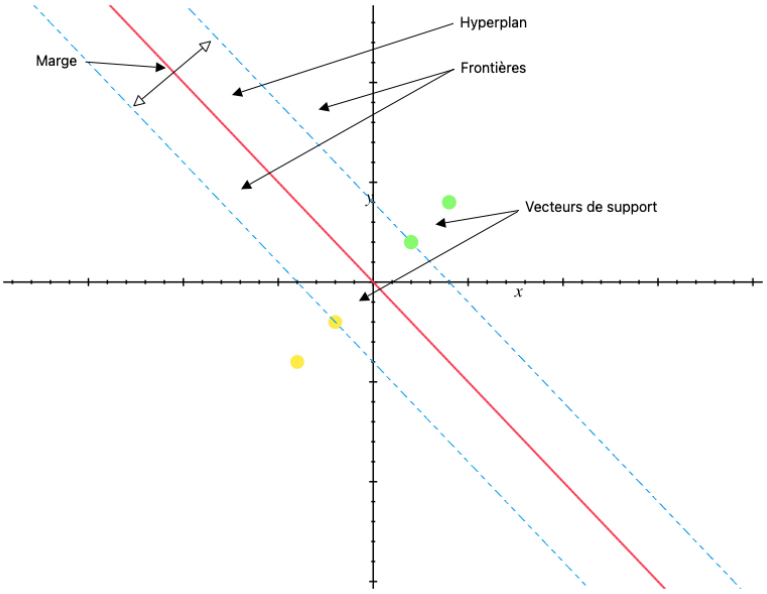

Figure 6 – Représentation d'un SVM

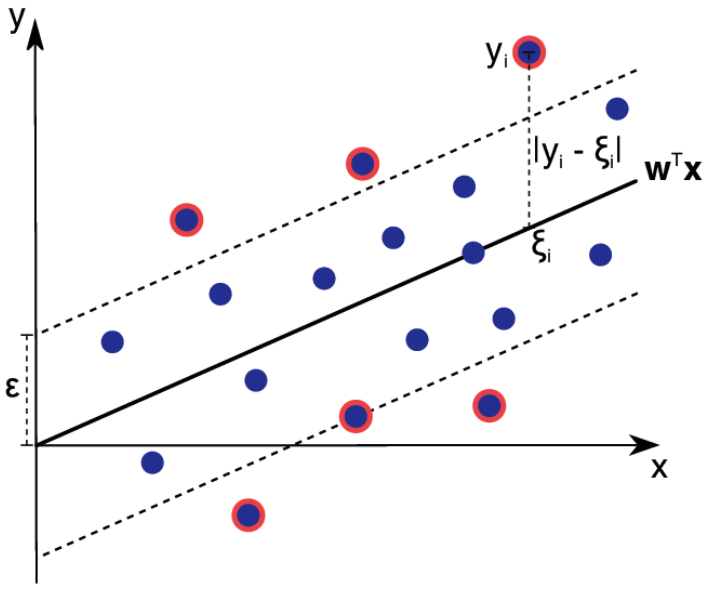

Figure 7 – Représentation d'un SVR (Rosenbaum et al., 2013)

1.5.3 Résultats

D'un point de vue global, les méthodes statistiques n'apportent pas de réponse directe au problème. Ces méthodes ne semblent en effet pas apporter de résultats satisfaisants dans la recherche d'un lien significatif entre des messages sur le marché boursier issus de Reddit et les variations de différentes actions (Talamás, 2021).

Certaines informations utiles peuvent cependant être retenues de ces analyses. Ainsi, selon Ranco et al., 2015, la quantité d'attention sur Twitter semble être utile pour prédire la volatilité des prix. De même, selon Yang et al., 2015, une régression s'intéressant au sentiment du marché permet de mettre en évidence l'influence significative des nœuds critiques sur les mouvements du marché financier.

Les méthodes se basant sur l'apprentissage automatique semblent alors plus prometteuses, bien que ne proposant pas de résultats se démarquant clairement des autres études. Selon Oliveira et al., 2017, le filtre de Kalman combiné à des analyses utilisant des techniques d'apprentissage automatique semble en effet offrir des perspectives positives quand à la recherche d'un lien entre des messages issus de plateformes de microblogging et le marché boursier.

1.6 Méthodes d'extraction de données

Les méthodes d'extraction de données consistent en des outils permettant d'extraire des informations pertinentes de plateformes diverses. Dans le cadre de ce mémoire, trois sources d'information principales seront utilisées : Reddit, Twitter et Yahoo ! Finance. Les deux premières seront utilisées afin de rassembler des messages liés au marché boursier tandis que la dernière sera source d'information directe du marché financier.

Les deux plateformes de microblogging proposent un API pour extraire les données (« api - reddit.com », s. d. ; « Getting access to the Twitter API », s. d.). Yahoo ! Finance ne propose cependant plus d'API, mais il est tout de même possible d'extraire des données historiques grâce à un package Python (Aroussi, 2019).

1.6.1 Reddit

L'accès à l'API de Reddit est ouvert et ne nécessite pas de démarche supplémentaire que la création d'un simple compte sur la plateforme (« api - reddit.com », s. d.).

Après obtention des informations nécessaires pour accéder à son API, le package PRAW, pour "The Python Reddit API Wrapper", basé sur le langage de programma-

tion Python (« PRAW 7.4.0 documentation », s. d.), permet un accès simple à l'API de Reddit. Ce package vise à être facile à utiliser et suit en interne toutes les règles de l'API de Reddit (« PRAW », 2021).

Il est ensuite possible d'extraire facilement les publications d'un subreddit, bien que comportant certaines limitations (« Subreddit — PRAW 7.4.1.dev0 documentation », s. d.). Cette procédure sera expliquée en détail lors de l'extraction des données.

Le code de base, présenté dans le code 1.1, consiste à importer le package pour ensuite définir les identifiants de connexion et pour finalement afficher les publications issues d'un subreddit (« Quick Start — PRAW 7.4.0 documentation », s. d.).

Code 1.1 – PRAW - Quick Start

```
import praw

reddit = praw.Reddit(
    client_id="my client id",
    client_secret="my client secret",
    user_agent="my user agent",
    username="my username",
    password="my password",
    )

for submission in reddit.subreddit("learnpython").hot(limit=10):
  print(submission.title)
```

(« Quick Start — PRAW 7.4.0 documentation », s. d.).

1.6.2 Twitter

Pour accéder à l'API de Twitter, la démarche est un peu plus complexe. Il est en effet nécessaire de demander un compte développeur et d'ensuite faire approuver l'utilisation de l'API en détaillant l'usage des données qui seront extraites du site (« Getting access to the Twitter API », s. d.).

À partir du moment où nous possédons l'accès à l'API, il est possible de travailler de deux manières : soit en accédant directement et manuellement à l'API de Twitter en programmant avec Python (Edward, 2021) ; ou en utilisant des packages basés sur Python, qui permettent d'extraire les données de la plateforme. Un de ces packages s'appelle *Tweepy* (« Tweepy », s. d.). La méthode employée dépendra des avantages et inconvénients de chaque méthode et sera détaillée lors de l'extraction des données.

Le code de base pour utiliser le package *Tweepy* est présenté dans le code 1.2. Ce code consiste à importer le package pour ensuite définir les informations d'accès à l'API de Twitter afin de finalement afficher le fil d'actualité de l'utilisateur contenant des tweets (« Getting started — tweepy 3.10.0 documentation », s. d.).

Code 1.2 – Tweepy - Getting started

```
import tweepy

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

public_tweets = api.home_timeline()
for tweet in public_tweets:
    print(tweet.text)
```

(« Getting started — tweepy 3.10.0 documentation », s. d.).

1.6.3 Yahoo ! Finance

Comme expliqué au début de cette section, Yahoo ! Finance a mis hors service son API permettant d'accéder facilement aux données historiques des marchés financiers. Heureusement, des méthodes alternatives fiables permettent à l'heure actuelle de continuer à profiter des informations précieuses de la plateforme. Ainsi, le package Python *yfinance* permet d'extraire facilement des informations financières (Aroussi, 2019).

Le code 1.3 présente le package, très simple d'utilisation, permettant de récupérer un grand nombre d'informations qui pourront nous être utiles au sein des futures analyses.

Code 1.3 – yfinance - Quick Start

```
import yfinance as yf

gme = yf.Ticker("GME")

# get stock info
gme.info
# get historical market data
hist = gme.history(period="max")
# show actions (dividends, splits)
gme.actions
# show dividends
gme.dividends
```

(Aroussi, 2022).

1.7 Analyse de sentiments

L'analyse de sentiments est un domaine très actif dans le traitement du langage. L'objectif est de définir des outils automatiques qui sont capables d'extraire des informations subjectives de textes, comme des avis ou des sentiments, dans le but de créer des connaissances qui peuvent être utilisées par un système d'aide à la décision ou un décideur (Pozzi et al., 2017).

L'analyse de sentiments présente encore plus d'intérêt avec les réseaux sociaux. L'interconnexion entre les utilisateurs actifs crée un espace de discussion permettant de relier un grand nombre d'individus autour de sujets et objectifs communs. Ces plateformes permettent alors d'exprimer et partager des émotions et opinions à travers le réseau. Ces données, récoltées puis analysées, permettent de comprendre et d'expliquer de nombreux phénomènes ainsi que de les prédire (Pozzi et al., 2017)

Comme abordé précédemment (section 1.5.2), l'analyse de sentiments peut être utilisée afin de traiter et d'analyser les messages issus des réseaux sociaux. Cette technique, pouvant être appliquée et utilisée de façons différentes en fonction de l'outil utilisé, nécessite néanmoins une tâche commune à toutes les analyses : le prétraitement. C'est ce dernier que nous allons aborder à travers les divers papiers scientifiques qui nous ont permis de rédiger la section 1.5.

Ranco et al., 2015 ont, lors du prétraitement des tweets, supprimé les URL, car elles ne représentent pas un contenu particulièrement pertinent. Les cashtags (voir section 1.2.1) et les mentions d'utilisateurs ont aussi été supprimés pour obtenir des tweets indépendants d'une action spécifique ou d'utilisateurs. Finalement, les répétitions de lettres ont aussi été traitées (par exemple : « coooool » devient « cool »). Oliveira et al., 2017 ont eux remplacé tous les cashtags, nombres, mentions et URL par un terme neutre.

Après cette étape de prétraitement, des étapes de transformation sont appliquées aux messages :

— Une *tokenization* est appliquée ;
— Les *stop words* sont aussi généralement retirés ;
— Un processus de *stemming* ou de *lemmatization* est appliqué.

(Oliveira et al., 2017 ; Yang et al., 2015).

La **tokenization** consiste à découper en morceaux, appelés tokens, une séquence de caractères selon une unité définie. Ces tokens sont généralement appelés termes ou mots, bien qu'un token peut représenter quelque chose de plus spécifique. À titre d'exemple, une phrase peut-être découpée en plusieurs tokens qui représenteront les différents mots de la phrase (« Tokenization », 2009).

Les **stop words**, ou mots vides en français, représentent l'ensemble de mots couramment utilisés dans une langue. Les stop words sont généralement supprimés lors de l'exploration de texte et le traitement du langage naturel pour éliminer les mots couramment utilisés qui transportent alors très peu d'informations (Ganesan, 2019).

Le but du **stemming** et de la **lemmatization** est de réduire les formes flexionnelles et dérivées d'un mot à une forme de base commune. Les deux méthodes fonctionnent cependant différemment. Le stemming fait référence à un processus qui coupe les extrémités des mots dans l'espoir de réaliser la transformation correctement la plupart du temps. La lemmatization fait, elle, référence à l'utilisation d'un vocabulaire et d'une analyse morphologique des mots visant à uniquement supprimer les terminaisons flexionnelles et à renvoyer la forme de base d'un mot (« Stemming and lemmatization », 2009). Ainsi, les mots "develop", "developed", "developing" et "development" seront remplacés par "develop" lors d'une procédure de stemming en retirant la fin de ces mots. Des mots comme "drive", "drives", "drove" ou "driven" seront remplacés plus correctement par "drive" lors d'une lemmatization alors qu'ils n'auraient pas tous été remplacés par la même racine lors d'un stemming.

Finalement, l'analyse de sentiments résout un problème de classification pour lequel les classes sont généralement définies comme positives et négatives. Le modèle va alors utiliser des caractéristiques qui sont les données textuelles transformées à l'aide d'un vectorizer. Ce dernier transforme les documents, ou ici les messages, en vecteurs afin de pouvoir ensuite appliquer un modèle de classification (Zhu, 2021).

Chapitre 2

Hypothèses

Suite à cette revue de littérature, et comme expliqué en introduction de ce mémoire, nous souhaitons tester diverses hypothèses à la fois sur l'impact des messages postés sur le marché boursier, et sur l'impact du marché boursier sur les messages postés.

Ainsi, deux hypothèses génériques peuvent être formulées :

— Le nombre de messages [sentiment] postés sur [plateforme] à propos d'une action [impacte/n'impacte pas] la [variable] du cours boursier de cette dernière.

— La [variable] du cours boursier d'une action [impacte/n'impacte pas] le nombre de messages [sentiment] postés sur [plateforme] à propos de cette dernière.

Ces deux types d'hypothèses seront déclinées selon le sentiment des messages, la plateforme, le type d'impact sur le cours boursier ainsi que la variable étudiée.

Comme expliqué lors de la revue de littérature, le sentiment des messages peut être positif, négatif ou neutre. Nos hypothèses considèrent un impact lorsque les messages sont positifs ou négatifs, mais aucun impact s'ils sont neutres.

Concernant la plateforme, les hypothèses seront déclinées pour Reddit et Twitter.

Finalement, après lecture des papiers scientifiques présentés dans les fondements théoriques de ce mémoire et après analyse des variables financières à notre disposition, deux variables seront utilisées : la variation journalière et le volume journalier.

Les hypothèses qui seront testées dans ce mémoire sont présentées dans les pages suivantes.

2.1 Impact des messages postés sur le cours boursier

2.1.1 Reddit

Impact sur la variation journalière du cours boursier

H1a : Le nombre de messages **positifs** postés sur Reddit à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.

H1b : Le nombre de messages **négatifs** postés sur Reddit à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.

H1c : Le nombre de messages **neutres** postés sur Reddit à propos d'une action **n'impacte pas** la **variation journalière** du cours boursier de cette dernière.

Impact sur le volume journalier de l'action

H2a : Le nombre de messages **positifs** postés sur Reddit à propos d'une action **impacte** le **volume journalier** de cette dernière.

H2b : Le nombre de messages **négatifs** postés sur Reddit à propos d'une action **impacte** le **volume journalier** de cette dernière.

H2c : Le nombre de messages **neutres** postés sur Reddit à propos d'une action **n'impacte pas** le **volume journalier** de cette dernière.

2.1.2 Twitter

Impact sur la variation journalière du cours boursier

H3a : Le nombre de messages **positifs** postés sur Twitter à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.

H3b : Le nombre de messages **négatifs** postés sur Twitter à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.

H3c : Le nombre de messages **neutres** postés sur Twitter à propos d'une action **n'impacte pas** la **variation journalière** du cours boursier de cette dernière.

Impact sur le volume journalier de l'action

H4a : Le nombre de messages **positifs** postés sur Twitter à propos d'une action **impacte** le **volume journalier** du cours boursier de cette dernière.

H4b : Le nombre de messages **négatifs** postés sur Twitter à propos d'une action **impacte** le **volume journalier** du cours boursier de cette dernière.

H4c : Le nombre de messages **neutres** postés sur Twitter à propos d'une action **n'impacte pas** le **volume journalier** du cours boursier de cette dernière.

2.2 Impact du cours boursier sur les messages postés

2.2.1 Reddit

Impact de la variation journalière du cours boursier

H5a : La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **positifs** postés sur Reddit à propos de cette dernière.

H5b : La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **négatifs** postés sur Reddit à propos de cette dernière.

H5c : La **variation journalière** du cours boursier d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Reddit à propos de cette dernière.

Impact du volume journalier de l'action

H6a : Le **volume journalier** d'une action **impacte** le nombre de messages **positifs** postés sur Reddit à propos de cette dernière.

H6b : Le **volume journalier** d'une action **impacte** le nombre de messages **négatifs** postés sur Reddit à propos de cette dernière.

H6c : Le **volume journalier** d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Reddit à propos de cette dernière.

2.2.2 Twitter

Impact de la variation journalière du cours boursier

H7a : La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **positifs** postés sur Twitter à propos de cette dernière.

H7b : La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **négatifs** postés sur Twitter à propos de cette dernière.

H7c : La **variation journalière** du cours boursier d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Twitter à propos de cette dernière.

Impact du volume journalier de l'action

H8a : Le **volume journalier** d'une action **impacte** le nombre de messages **positifs** postés sur Twitter à propos de cette dernière.

H8b : Le **volume journalier** d'une action **impacte** le nombre de messages **négatifs** postés sur Twitter à propos de cette dernière.

H8c : Le **volume journalier** d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Twitter à propos de cette dernière.

Chapitre 3

Méthodologie

La recherche sur une éventuelle relation entre des messages issus de plateformes de microblogging comme Reddit et Twitter et le marché boursier peut suivre différentes méthodologies. Après avoir parcouru différents papiers scientifiques, et ayant sélectionné les techniques semblant présenter les résultats les plus encourageants, nous proposons ici une méthodologie liant à la fois *Machine Learning* et statistiques afin d'aborder la question de recherche.

La méthodologie reposera sur 5 grands axes : l'extraction de données, le nettoyage des données, l'analyse exploratoire des données, l'analyse de sentiments et finalement, les analyses statistiques.

La première étape consiste à extraire les messages des deux plateformes, Reddit et Twitter, durant une période de temps déterminée à partir de la date de commencement des analyses de ce mémoire et qui représente environ un an, entre le 1er octobre 2020 et le 24 octobre 2021. La première extraction concernera Reddit, ce qui permettra ensuite de déterminer les actions que nous analyserons. Nous pourrons ensuite extraire les tweets relatifs à ces actions. Elle consistera aussi à extraire les informations financières des actions sélectionnées.

Nous aborderons ensuite le nettoyage des données, partie dans laquelle uniquement les informations essentielles seront conservées.

La troisième étape consistera à découvrir les données en notre possession à l'aide d'une analyse exploratoire, qui permettra d'avoir un premier aperçu des données récupérées.

Après ces 3 premières étapes se concentrant principalement sur la manipulation de données, l'analyse de sentiments permettra d'associer un caractère positif, négatif ou neutre aux messages extraits de Reddit et Twitter.

Finalement, les analyses finales permettront de déceler, s'il existe, un lien entre le sentiment des messages et le cours boursier des actions sélectionnées.

3.1 Extraction des données

Dans cette partie, nous aborderons la première étape de la méthodologie appliquée dans ce mémoire : l'extraction des données. Cette section sera séparée en 3 sous-sections, l'une abordant les techniques appliquées pour Reddit, la deuxième se concentrant sur Twitter et la dernière à propos de l'extraction des informations financières sur Yahoo ! Finance.

L'extraction de données représente une activité chronophage en fonction de la quantité de données souhaitée. Quelques semaines auront été nécessaires afin de constituer un ensemble de données conséquent.

3.1.1 Reddit

Étant donné la particularité de Reddit qui est, pour rappel, de proposer des espaces de discussion, appelés *subreddits*, dans lesquels les internautes peuvent poster des messages et qui peuvent être commentés par d'autres personnes, il est nécessaire de prendre en compte non seulement les messages, mais également leurs commentaires.

Les messages ont été extraits des 4 subreddits financiers principaux de la plateforme, présentés dans la section 1.1.1 : *r/wallstreetbets*, *r/investing*, *r/finance* et *r/stocks*. L'extraction a été réalisée sur tous les messages postés pendant une durée d'un an dans ces 4 subreddits, comme expliqué en introduction de ce chapitre.

Lors d'une première tentative d'extraction de messages issus de Reddit, uniquement les messages avaient été considérés, réduisant la quantité de données et pouvant impacter la qualité des analyses suivantes. Un code python avait été réalisé à cet effet (voir annexe A.1). Une seconde extraction a donc été réalisée afin d'extraire les messages et les commentaires, offrant ainsi une plus grande quantité de données.

Cette seconde extraction de données pour Reddit a été réalisée grâce un outil open source, disponible sur GitHub (« pistocop/subreddit-comments-dl », s. d.). Étant donné les limitations associées à l'API de Reddit, principalement en termes de nombre de requêtes autorisées par minute (« API · reddit-archive/reddit Wiki », s. d.), cet outil se base sur l'API de Pushshift.

Pushshift est une plateforme de collecte, d'analyse et d'archivage de données qui collecte des données Reddit et les met à la disposition des chercheurs. L'ensemble de données Reddit de Pushshift est mis à jour en temps réel. L'accès aux données est simplifié, permettant de réduire le temps passé dans les phases de collecte, de nettoyage et de stockage des données (Baumgartner et al., 2020).

Les messages et commentaires récoltés sont enregistrés dans des fichiers CSV, permettant ensuite de les réutiliser dans la suite des analyses. L'outil nécessite néanmoins un accès à un compte Reddit comme le nécessite l'API officiel de Reddit. La commande présentée dans le code 3.1 , appelant un code Python, permet d'extraire les données de la plateforme avec l'outil présenté plus tôt.

Code 3.1 – subreddit-comments-dl - Extraction des données

```
python src/subreddit_downloader.py <subreddit>
        --batch-size 512
        --laps 1000
        --reddit-id <reddit-id>
        --reddit-secret <reddit-secret>
        --reddit-username <reddit-username>
        --utc-after <unixtimestamp>
```

(« pistocop/subreddit-comments-dl », s. d.).

Dans cette commande, plusieurs paramètres sont demandés : $< subreddit >$ représente le subreddit sur lequel nous voulons extraire les données, $< reddit - id >$, $< reddit - secret >$ et $< reddit - username >$, des informations propres à un compte Reddit et $< unixtimestamp >$ représente une date à partir de laquelle les données sont extraites, exprimée en temps Unix, le nombre de secondes qui se sont écoulées depuis le 1er janvier 1970 (« Unix Time Stamp - Epoch Converter », s. d.).

Finalement, la commande du code 3.2 permet de construire et structurer les données extraites précédemment.

Code 3.2 – subreddit-comments-dl - Construction de l'ensemble de données

```
python src/dataset_builder.py
```

(« pistocop/subreddit-comments-dl », s. d.).

À l'issue de cette extraction, 1 014 613 messages et 8 419 339 commentaires ont été extraits (avant nettoyage), représentant environ 24 Go de données.

Sélection des actions

Après extraction des messages et commentaires issus de Reddit, il est nécessaire de déterminer les actions les plus discutées sur la plateforme afin de réaliser une sélection. Un code a alors été réalisé afin d'extraire les symboles boursiers des actions dans les messages et de les comptabiliser (voir annexe A.2). Ce code retire une sélection de mots non nécessaires à la recherche des actions dans les messages et cherche ensuite les symboles à partir d'une liste contenant la majorité des symboles boursiers existants.

Afin d'éviter un biais dans les résultats de ce mémoire, principalement dû à l'engouement autour de l'affaire GameStop (voir section 1.1.2), les actions GameStop (**GME**) et AMC Entertainment Holdings (**AMC**) ont été délibérément retirées de la liste des actions les plus discutées.

À l'issue de cette sélection, 13 actions/fonds ont été déterminés :

1. Apple (**AAPL**) : Apple conçoit, fabrique et commercialise des smartphones, ordinateurs, tablettes et accessoires dans le monde entier (« Apple Inc. (AAPL) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
2. Advanced Micro Devices (**AMD**) : AMD est une entreprise mondiale de semiconducteurs (« Advanced Micro Devices, Inc. (AMD) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
3. BlackBerry (**BB**) : BlackBerry fournit des logiciels et services de sécurité intelligents aux entreprises et aux gouvernements du monde entier (« BlackBerry Limited (BB) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
4. Clover Health (**CLOV**) : Clover Health est un assureur de soins de santé aux États-Unis (« Clover Health Investments, Corp. (CLOV) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
5. Cenntro Electric Group (**CENN**) (anciennement Naked Brand (**NAKD**)) : Cenntro Electric Group est une entreprise leader dans le domaine de la technologie des véhicules électriques (Freehold, 2022).
6. NIO (**NIO**) : NIO conçoit, développe, fabrique et vend des véhicules électriques intelligents en Chine (« NIO Inc. (NIO) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
7. Nokia (**NOK**) : Nokia fournit des solutions de réseaux mobiles et fixes dans le monde entier (« Nokia Corporation (NOK) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
8. Palantir Technologies (**PLTR**) : Palantir Technologies construit et déploie des plateformes logicielles pour la cellule de renseignement aux États-Unis afin de faciliter

les enquêtes et les opérations de lutte contre le terrorisme (« Palantir Technologies Inc. (PLTR) Stock Price, News, Quote & History - Yahoo Finance », s. d.).

9. Rocket Companies (**RKT**) : Rocket Companies exerce des activités dans les domaines de l'immobilier, des prêts hypothécaires et du commerce électronique aux États-Unis et au Canada (« Rocket Companies, Inc. (RKT) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
10. Sundial Growers (**SNDL**) : Sundial Growers se consacre à la production et à la commercialisation de produits à base de cannabis destinés au marché canadien (« Sundial Growers Inc. (SNDL) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
11. Virgin Galactic (**SPCE**) : Virgin Galactic est une entreprise aérospatiale qui développe des vols spatiaux habités pour des particuliers et chercheurs aux États-Unis (« Virgin Galactic Holdings, Inc. (SPCE) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
12. SPDR S&P 500 Trust ETF (**SPY**) : ce fonds est un des plus grands fonds négociés en bourse, regroupant différentes actions (« SPDR S&P 500 ETF Trust (SPY) Stock Price, News, Quote & History - Yahoo Finance », s. d.).
13. Tesla (**TSLA**) : Tesla conçoit, développe, fabrique, loue et vend des véhicules électriques et des systèmes de production et de stockage d'énergie (« Tesla, Inc. (TSLA) Stock Price, News, Quote & History - Yahoo Finance », s. d.).

Ces 13 actifs assez hétéroclites, qui seront nommés par leur symbole (entre parenthèses, en gras) dans la suite de ce mémoire, sont les plus discutés sur Reddit pour la période sélectionnée. Afin de réaliser des analyses similaires sur les deux plateformes, les messages à propos de ces mêmes actifs seront extraits de Twitter.

3.1.2 Twitter

Pour l'extraction de données issues de Twitter, de la même façon que pour Reddit, l'API officielle n'a pas été utilisée, car trop limitée (« Rate limits », s. d.). La librairie Python *Twint* (« TWINT - Twitter Intelligence Tool », 2022) a été utilisée afin d'extraire les tweets abordant les actions sélectionnées.

Les tweets publiés sur la période du 01/10/2020 au 24/10/2021 ont été extraits, mais dans ce cas, seulement les messages contenant le *cashtag* (voir section 1.2.1) associé aux actions sélectionnées ont été extraits. Cette particularité a permis d'accélérer le processus d'extraction et de réaliser un premier nettoyage des données.

Contrairement à l'utilisation d'un outil préexistant comme pour Reddit, un code Python a été réalisé afin d'assurer l'extraction des tweets (voir annexe A.3). Cet algorithme extrait jour par jour les tweets associés à une action particulière et les enregistre dans un fichier CSV.

À l'issue de l'extraction, un fichier CSV a donc été généré pour chaque jour de la période considérée, et ce par action. Un autre code a donc été nécessaire afin de former l'ensemble de données final (voir annexe A.4). Ce code concatène l'ensemble des CSV pour une action au sein d'un seul CSV par action.

Le nombre total de tweets extraits se somme à un peu plus de 6 millions pour 13 actifs (6 462 684 exactement). Cela représente un peu moins de 4 Go de données.

3.1.3 Yahoo ! Finance

Le dernier lot de données nécessaire aux analyses de ce mémoire concerne les informations financières. Nous entendons par informations financières différentes métriques propres aux marchés financiers telles que :

— Le cours à l'ouverture ;
— Le cours à la fermeture ;
— Le volume journalier.

Ces informations financières sont décrites dans la section 1.5.1. Ces données sont extraites pour chaque action, et ce pour chaque jour entre le 01/10/2020 et le 24/10/2021. L'ensemble de données final comprend donc un fichier CSV pour chaque action, dans lequel chaque ligne correspond aux informations financières d'une journée.

Afin d'extraire ces données, la librairie Python *yfinance* (Aroussi, 2022) est utilisée. Un code a été écrit afin d'extraire les données financières des 13 actifs sélectionnés (voir annexe A.5). Le tableau 1 présente un exemple de la première ligne des données extraites pour l'action Apple (AAPL).

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
2020-10-01	116.71	116.79	114.92	115.87	116120400	0.0	0

Table 1 – Données financières provenant de Yahoo ! Finance pour l'action AAPL

En plus du cours à l'ouverture (Open), à la fermeture (Close) et du volume journalier (Volume), d'autres informations sont présentes comme le prix journalier le plus élevé

(High), le prix journalier le plus faible (Low), ou encore d'autres informations liées aux dividendes (Dividends) ou à la division d'actions (Stock Splits). Seules la date et les informations présentées en début de cette section seront conservées pour la suite des analyses.

3.2 Préparation des données

La préparation des données est une étape importante permettant d'assurer un ensemble de données propre à l'analyse et particulièrement à l'analyse de sentiments.

3.2.1 Reddit

Après extraction des messages et commentaires, il est nécessaire de déterminer les messages et commentaires associés ne concernant que les actions sélectionnées, présentées dans la section 3.1.1.

Deux codes ont été écrits à cette fin. Le premier permet de classer les commentaires et messages par action ainsi que de filtrer la date entre le 01/10/2020 et le 24/10/2021. Le second s'occupe de concaténer dans un seul fichier les messages et commentaires pour chaque action. Ces codes sont présentés respectivement dans les annexes A.6 et A.7. Pour chaque action, différents mots clés ont été sélectionnés afin de réaliser la recherche et les éventuels doublons ont été retirés avant l'enregistrement du nouveau fichier CSV. Ces mots clés sont présents dans ce même code.

3.2.2 Twitter

Pour Twitter, étant donné qu'uniquement les tweets correspondant aux actions sélectionnées ont été extraits, la phase de nettoyage ne consiste qu'en une vérification des dates, comme pour Reddit, avec le code présenté dans l'annexe A.8.

3.2.3 Yahoo ! Finance

Pour les données venant de Yahoo ! Finance, de nouvelles métriques ont été calculées à partir des données d'origine. Pour rappel, trois informations nous intéressent particulièrement : le cours à l'ouverture (Open), le cours à la fermeture (Close) et le volume journalier (Volume). Les données nettoyées pour l'action Apple (AAPL) sont présentées dans le tableau 2.

Date	Open	Close	Volume
2020-10-01	116.71	115.87	116120400

Table 2 – Données financières nettoyées pour l'action AAPL

À partir de ces 3 données, une nouvelle variable a pu être calculée : la variation journalière (Daily variation), qui correspond à la différence entre le cours à la fermeture et le cours à l'ouverture. Cette nouvelle métrique a été calculée et insérée dans les fichiers CSV correspondants à chaque action à l'aide d'un code présenté dans l'annexe A.9. Les nouvelles données calculées sont représentées dans le tableau 3.

Date	Open	Close	Volume	Daily variation
2020-10-01	116.71	115.87	116120400	-0.84

Table 3 – Données financières complètes pour l'action AAPL

Cette nouvelle variable nous permettra de rechercher un lien entre la variation du cours boursier d'une action et le sentiment des messages. Cette relation sera expliquée en détail dans les sections à ce propos.

3.3 Analyse exploratoire des données

Afin d'avoir une vision préliminaire des données en notre possession, nous réalisons une analyse exploratoire pour décrire, pour chaque action, les données financières ainsi que les messages associés sur la période de temps considérée.

Différentes métriques à propos des informations financières seront aussi présentées : la variation journalière moyenne (DAVar) et le volume journalier moyen (DAVol) sur l'année ainsi que les cours à l'ouverture du 01/10/2020 (Open20) et du 01/10/2021 (Open21) permettront d'avoir un aperçu général de l'évolution du cours de l'action sur l'année.

Pour chacune des actions, une carte de chaleur de la répartition annuelle des messages a été réalisée. Des diagrammes en bâtonnets ont aussi été réalisés afin d'observer les jours où le plus de messages ont été postés à propos des différentes actions. Ces graphiques sont disponibles dans l'annexe B. Le code utilisé pour réaliser cette analyse est présent dans l'annexe A.10.

3.3.1 Répartition annuelle des messages

Nous observons plusieurs tendances grâce aux cartes de chaleurs des différentes actions. Premièrement, pour Reddit (annexe B.1.1), nous remarquons une répartition assez sporadique tout au long de l'année, quelle que soit l'action. Cela peut probablement être expliqué par le nombre inférieur de messages extraits sur la plateforme.

Ensuite, pour Twitter (annexe B.2.1), la répartition est beaucoup plus constante pour une partie des actions, pouvant peut-être déjà indiquer Twitter comme étant une plateforme plus intéressante pour obtenir un flux de données abondant et plus constant. De plus, on observe ici que les messages sont principalement postés du lundi au vendredi, jours pendant lesquels le marché boursier est ouvert, alors que ce n'est pas forcément le cas pour Reddit.

3.3.2 Jours présentant le plus de messages

Les diagrammes en barre réalisés pour Reddit (annexe B.1.2) et Twitter (annexe B.2.2) permettent d'observer plus en détail les pics de messages observés sur les cartes de chaleurs. De manière générale, si le flux de données est plus constant, comme on peut l'observer sur Twitter, peu de différence est observée entre les 10 jours pour lesquels le plus de messages ont été postés. Pour Reddit par contre, nous pouvons observer de grosses différences, avec des jours particuliers où le nombre de messages est très élevé en comparaison au reste de l'année.

3.3.3 Informations financières

Les métriques à propos des informations financières décrites en début de ce chapitre ont été calculées pour chaque action. Le tableau 4 présente un récapitulatif de ces dernières.

Plusieurs observations peuvent-être faites. Premièrement, le nombre de messages extrait pour chaque action est majoritairement plus élevé sur Twitter, confirmant de nouveau la plus grande abondance d'information sur Twitter. Ensuite, la variation journalière moyenne sur l'année semble être assez limitée pour toutes les actions. Finalement, les cours à l'ouverture au début et à la fin de l'année permettent de calculer la variation annuelle des actifs. On observe que la majorité a bien performé avec un cours à l'ouverture plus élevé après un an pour 11 des 13 actions.

Tickers	Nombre de messages		Informations financières				
	Twitter	Reddit	DAVar %	DAVol	Open20	Open21	Var. annuelle
AAPL	550816	47161	0.01%	95605480	116.71	141.69	21.40%
AMD	352751	25817	-0.02%	48491140	83.05	102.59	23.52%
BB	162935	255346	-0.10%	28320050	4.59	9.78	113.07%
CLOV	219540	2361	-0.54%	26628590	12.77	7.51	-41.19%
CENN	169142	10757	0.25%	7402826	1.65	10.64	544.84%
NIO	412888	104864	0.15%	96047540	21.68	36.63	68.95%
NOK	171738	112630	-0.08%	45228740	3.91	5.5	40.66%
PLTR	279811	189563	0.02%	60293920	9.68	24.2	150.00%
RKT	69986	8269	-0.18%	11747010	19.08	16.04	-15.93%
SNDL	259577	19164	-0.92%	256486700	0.25	0.68	172.00%
SPCE	136264	14535	0.08%	21839380	19.25	25.06	30.18%
SPY	843290	19837	0.02%	72933920	331.83	429.47	29.42%
TSLA	2833946	184831	0.07%	31099770	440.76	778.4	76.60%

TABLE 4 – Analyse exploratoire - Résumé

3.4 Analyse de sentiments

Après avoir exploré les données en notre possession, une analyse de sentiments va permettre de caractériser les messages récupérés des deux plateformes. Ainsi, un modèle préentrainé sur 58 millions de tweets (CARDIFF NLP, s. d.) permettra de caractériser les messages selon trois catégories : positif, négatif ou neutre. Le nombre de messages positifs, négatifs ou neutres sera ensuite utilisé pour rechercher si une relation existe entre les messages postés sur Reddit et Twitter et le marché boursier.

Le modèle utilisé, appelé *Twitter-roBERTa-base for Sentiment Analysis* et appuyé par une publication scientifique (BARBIERI et al., 2020), se base sur un modèle de *Transformers* bien connu : le modèle *BERT*.

> Un modèle de Transformers est un réseau neuronal qui apprend le contexte et le sens en suivant les relations dans des données séquentielles comme les mots dans une phrase (MERRITT, 2022).

Le modèle BERT, signifiant *Bidirectional Encoder Representations from Transformers*, est conçu pour pré-entraîner des représentations d'un texte non étiqueté en conditionnant conjointement le contexte à gauche et à droite. BERT peut-être utilisé pour un large éventail de tâches, telles que répondre à des questions ou réaliser de l'inférence linguistique (Devlin et al., 2019).

Afin d'appliquer cette analyse de sentiments, diverses étapes ont été réalisées. Un nettoyage des messages a été réalisé et sera décrit dans la section 3.4.1. L'analyse de sentiments a ensuite été appliquée aux messages afin de déterminer le sentiment de ces derniers. Ceci sera décrit dans la section 3.4.2. Finalement, à partir des données récoltées sur Yahoo ! Finance et le sentiment des messages, l'ensemble de données final a pu être constitué afin de l'utiliser lors des analyses subséquentes. Cette dernière étape constituera la section 3.4.3. Le code appliquant l'analyse de sentiments est disponible dans l'annexe A.11.

3.4.1 Nettoyage des données

Pour le nettoyage des données, différentes méthodes ont été testées pour définir le nettoyage optimal à réaliser pour l'analyse de sentiments.

Ainsi, quatre méthodes ont été créées. La première pour retirer les noms d'utilisateurs ainsi que les URL des messages. La seconde pour retirer les caractères spéciaux, la troisième pour retirer les tickers, c'est-à-dire, les symboles boursiers des actions et finalement, la dernière pour retirer les stop words (définis dans la section 1.7). Les quatre méthodes ont été appliquées seules et simultanément pour déterminer la combinaison optimale de nettoyage.

Après utilisation et évaluation de l'impact du nettoyage lors de l'analyse de sentiments, seule la première méthode a été conservée et appliquée. Les noms d'utilisateurs et les URL ont donc été retirés des messages à destination de l'analyse de sentiments.

Ce résultat peut facilement être expliqué : des informations contextuelles sont perdues lors d'un nettoyage trop intensif des données, produisant de moins bons résultats pour le modèle BERT (Bricken, 2021a, 2021b).

3.4.2 Application de l'analyse de sentiments

Après nettoyage des messages, l'analyse de sentiments a été appliquée sur ces derniers. Le modèle utilisé permet de travailler avec différents sentiments comme la colère, la joie,

la tristesse ou encore l'optimisme (« TweetEval », 2022), mais nous avons préféré travailler avec des classes plus générales. C'est pourquoi l'analyse détermine si les messages sont positifs, négatifs ou neutres.

L'analyse de sentiments a donc été appliquée sur l'ensemble des actions pour Reddit et Twitter. Ce processus était l'un des plus longs de cette analyse avec environ 10h de temps de calcul par action.

3.4.3 Préparation des données aux analyses finales

Après l'analyse de sentiments, un ensemble de données final a été constitué afin de réaliser les dernières analyses. À cette fin, les résultats de l'analyse de sentiments ainsi que les informations financières présentées précédemment ont été rassemblés en un fichier CSV final pour chaque action, et ce pour chaque plateforme.

Le tableau 5 présente la première ligne des données pour l'action AAPL sur Twitter.

Date	neutral	positive	negative	Volume	Daily variation
2020-10-01	922	511	79	116120400	-0.84

Table 5 – AAPL - Ensemble de données final sur Twitter

L'ensemble de données final contient, pour chaque jour où le marché boursier était ouvert entre le 01/10/2020 et le 24/10/2021, le nombre de messages négatifs (negative), positifs (positive) et neutres (neutral), le volume (Volume) ainsi que la variation journalière (Daily variation).

Les sections suivantes aborderont la manière dont les analyses finales utilisent les données collectées, nettoyées et générées durant tout le cheminement précédent.

3.5 Analyses finales

Dans cette section, nous nous intéressons aux relations entre le nombre de messages positifs, négatifs et neutres et la variation journalière ou le volume journalier des actions en utilisant des méthodes statistiques.

Nos données sont apparentées à des séries temporelles : elles contiennent l'évolution du sentiment des messages ainsi que de la variation journalière et du volume journalier entre le 01/10/2020 et le 24/10/2021.

Une **série temporelle** est une séquence d'observations qui se produisent successivement durant une certaine période de temps (Hayes, 2021b).

Nous nous intéresserons d'abord à des tests de corrélation entre séries temporelles pour ensuite nous intéresser à la régression linéaire, pour enfin terminer avec le test de causalité de Granger. Les résultats seront présentés dans le chapitre 4.

Après avoir étudié la possibilité de réaliser des analyses basées sur des méthodes d'apprentissage automatique et après des premiers essais sur nos données, nous avons décidé de ne pas suivre cette direction dû à un manque de données. En effet, malgré la quantité assez importante de messages extraits, l'ensemble de données final ne contient que 268 lignes représentant les jours où le marché boursier est ouvert durant une année. Cette quantité de données semble alors assez limitée pour entrainer un modèle permettant de rechercher une relation entre ces variables. Une extraction sur plusieurs années pourrait générer un ensemble de données assez conséquent, mais l'extraction de données étant chronophage, cela dépasse les limites de ce mémoire.

Les analyses ayant été considérées et appliquées afin d'avoir un premier aperçu sont alors les suivantes : un SVR, une forêt aléatoire ainsi qu'un réseau de neurones, tous les 3 présentés dans la section 1.5.2. Ces algorithmes ne présentaient pas des résultats intéressants dû au manque de données présenté ci-dessus.

3.5.1 Tests de corrélation

À partir des données en notre possession, nous réalisons des tests de corrélation entre nos séries temporelles. Cela nous permettra d'obtenir un premier aperçu de la relation entre nos différentes variables.

Nous utiliserons la corrélation de Pearson, introduite dans la section 1.5.2. Soit Neg le nombre de messages négatifs, Pos le nombre de messages positifs, Neu le nombre de messages neutres, Var la variation journalière et Vol le volume journalier, les relations entre les variables mentionnées en vert dans le tableau 6 seront testées.

	Neg	Pos	Neu	Var	Vol
Neg	1				
Pos		1			
Neu			1		
Var				1	
Vol					1

TABLE 6 – Matrice de corrélation entre les variables de l'ensemble de données

Ces premiers résultats nous permettront de diriger les analyses suivantes quant aux relations semblant présenter le plus d'intérêt ainsi que d'aider à l'interprétation des résultats. Le code utilisé pour réaliser ces analyses est disponible dans l'annexe A.12.

3.5.2 Régression linéaire

Après les tests de corrélation, une autre méthode cherchant une relation linéaire entre les données sera appliquée : la régression linéaire (présentée dans la section 1.5.2).

Nous considérons dans un premier temps le nombre de messages négatifs, positifs et neutres comme variables explicatives et la variation journalière et le volume journalier comme variables à expliquer. Dans un second temps, l'inverse sera considéré pour tester la relation opposée.

Pour le premier cas, nous avons élaboré deux équations contenant l'ensemble de nos variables liées aux messages. En effet, intégrer l'ensemble des variables explicatives dans nos équations plutôt qu'une seule variable indépendante permet de limiter un biais d'omission.

> Un **biais d'omission** se produit lorsqu'une ou plusieurs variables pertinentes ont été omises d'un modèle statistique. Le modèle attribue alors l'effet des variables manquantes à celles présentes au sein du modèle (« Biais de variable omise », s. d.).

Soit Neg le nombre de messages négatifs, Pos le nombre de messages positifs, Neu le nombre de messages neutres, Var la variation journalière et Vol le volume journalier, voici les différentes équations de régression qui devraient être évaluées :

$$Var = a + bPos + cNeg + dNeu$$

$$Vol = a + bPos + cNeg + dNeu$$

Cependant, en utilisant ces équations, une forte multicolinéarité est présente pour certaines des actions, même après centrage des données par rapport à la moyenne. Il est donc nécessaire d'utiliser d'autres équations ne contenant pas toutes les variables explicatives présentes ci-dessus. Les équations utilisées pour chacune des actions seront décrites dans la section 4.2.

La **multicolinéarité** est l'apparition de fortes corrélations entre deux ou plusieurs variables indépendantes dans un modèle de régression multiple (Hayes, 2022).

Pour la relation inverse, c'est-à-dire, considérer le nombre de messages positifs, négatifs et neutres comme variables à expliquer et la variation journalière où le volume comme variables explicatives, plusieurs équations doivent être testées. En effet, dans ce cas-ci, il est nécessaire de créer une équation pour chaque variable à expliquer afin de connaître l'effet sur cette dernière. Nous considérerons aussi la somme de ces variables à expliquer. Voici ces équations :

$$Pos + Neg + Neu = a + bVar + cVol \qquad Pos = a + bVar + cVol$$

$$Pos + Neg = a + bVar + cVol \qquad Neg = a + bVar + cVol$$

$$Pos + Neu = a + bVar + cVol \qquad Neu = a + bVar + cVol$$

$$Neg + Neu = a + bVar + cVol$$

De manière similaire, les équations contenant à la fois la variation journalière et le volume journalier semblent présenter de la multicolinéarité. Les équations finales présentées dans la section 4.2 ne contiendront donc pas les deux variables en même temps.

Afin d'évaluer la régression linéaire, le niveau de significativité ainsi que le R^2, présenté dans la section 1.5.2, seront analysés.

Le **niveau de significativité** est le seuil déterminant si le résultat d'une analyse peut être considéré comme statistiquement significatif. Le niveau de significativité est généralement fixé à 5%. Ce dernier représente la probabilité de rejeter l'hypothèse nulle lorsqu'elle est vraie (« Niveau de signification », s. d.).

Le code utilisé pour réaliser ces analyses est disponible dans l'annexe A.13.

3.5.3 Test de causalité de Granger

Après avoir appliqué des tests de corrélation et une régression linéaire à nos données, intéressons-nous au test de causalité de Granger. Ce dernier est défini dans la section 1.5.2, mais une définition plus avancée permet de mieux comprendre ce test :

> Une causalité au sens de Granger est observée de X1 sur X2, lorsque les valeurs passées de X1 contiennent des informations qui aident à prédire X2 au-delà des informations contenues dans les valeurs passées de X2 seules (Seth, 2007).

Afin d'appliquer le test de causalité de Granger, une série d'étapes doivent être suivies :

1. Tester le niveau d'intégration des séries temporelles, autrement dit, tester si nos séries temporelles sont stationnaires ou non.
2. Si elles ne le sont pas, remplacer nos séries temporelles par la différence entre la série temporelle et elle-même décalée de n observations avec $n >= 1$ jusqu'à ce que la série soit stationnaire. n représente alors le niveau d'intégration.
3. Appliquer un modèle VAR à nos séries temporelles pour déterminer le délai (lag) optimal entre nos séries, c'est-à-dire, le décalage optimal à appliquer pour comparer les deux séries temporelles.
4. Vérifier qu'il n'y ait pas d'autocorrélation dans les résidus.
5. Si des séries temporelles ont le même ordre d'intégration, vérifier s'ils elles sont cointégrées. En effet, si deux séries temporelles sont cointégrées, il doit y avoir une causalité Granger entre elles.
6. Calculer le test de causalité de Granger en prenant en considération le lag calculé à l'étape 3.

(Adhikary, 2020).

> Une **série temporelle stationnaire** possède des propriétés qui ne dépendent pas du moment où la série est observée (Kwiatkowski et al., 1992).

> Le **modèle vectoriel autorégressif** (VAR) est un modèle de série temporelle mettant en relation les observations actuelles d'une variable avec les observations passées de cette même variable et les observations passées des autres variables du système (« Introduction to the Fundamentals of Vector Autoregressive Models - Aptech », 2021).

L'**autocorrélation** se produit dans une série temporelle lorsqu'une variable et une version décalée d'elle-même sont corrélées (Banton, 2021).

Deux ou plusieurs séries temporelles non stationnaires sont **cointégrées** si elles sont corrélées à long terme (Corporate Finance Institute, s. d.).

Ces étapes seront appliquées entre la variation journalière ou le volume journalier et une des trois séries temporelles liées aux sentiments. Plusieurs tests seront donc réalisés pour tester la relation dans les deux sens. Le code utilisé pour réaliser ces analyses est disponible dans l'annexe A.14.

Chapitre 4

Résultats et interprétation

4.1 Tests de corrélation

Comme expliqué dans la section 3.5.1, un test de corrélation de Pearson a été appliqué à nos variables et ce pour chaque action et chaque plateforme.

En utilisant la même notation que présentée dans la section 3.5.1, les tableaux 7 et 8 nous permettent d'obtenir une vue d'ensemble des résultats.

	neg_pos	neg_neu	neg_Var	neg_Vol	pos_neu	pos_Var	pos_Vol	neu_Var	neu_Vol	Var_Vol
AAPL	0,92	0,96	-0,11	0,41	0,94	-0,11	0,36	-0,11	0,43	-0,19
AMD	0,86	0,96	-0,05	0,23	0,95	0,06	0,13	0,00	0,23	0,18
BB	0,96	0,98	-0,33	0,46	1,00	-0,23	0,54	-0,25	0,52	-0,10
CLOV	0,93	0,97	-0,28	0,20	0,92	-0,20	0,23	-0,18	0,20	0,06
NAKD	0,98	0,99	-0,12	0,70	0,97	-0,02	0,75	-0,22	0,67	0,28
NIO	0,97	0,97	-0,05	0,71	0,98	-0,02	0,75	-0,03	0,73	0,06
NOK	0,98	0,99	0,21	0,93	1,00	0,31	0,96	0,27	0,95	0,38
PLTR	0,96	0,99	-0,01	0,37	0,98	0,06	0,39	0,02	0,38	0,19
RKT	0,90	0,93	0,02	0,21	0,97	0,05	0,37	-0,03	0,31	0,48
SNDL	0,90	0,93	-0,54	0,70	0,98	-0,33	0,82	-0,41	0,79	-0,15
SPCE	0,84	0,88	-0,01	0,28	0,95	0,12	0,28	0,07	0,29	0,05
SPY	0,88	0,94	0,07	0,08	0,92	0,09	0,01	0,07	0,06	-0,33
TSLA	0,97	0,99	0,09	0,59	0,99	0,10	0,58	0,09	0,60	0,09

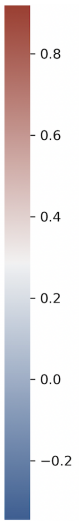

TABLE 7 – Scores de corrélation - Reddit

Nous pouvons directement en retirer certaines observations. Tout d'abord, les scores de corrélation entre le nombre de messages positifs, négatifs et neutres sont assez élevés, comme présentés dans les colonnes neg_pos, neg_neu et pos_neu. Ainsi, nous ne pourrons inclure ces variables ensemble lors des régressions linéaires, car elles sont trop corrélées.

	neg_pos	neg_neu	neg_Var	neg_Vol	pos_neu	pos_Var	pos_Vol	neu_Var	neu_Vol	Var_Vol
AAPL	0,68	0,79	-0,35	0,59	0,92	0,13	0,48	0,01	0,54	-0,17
AMD	0,58	0,62	-0,04	0,13	0,90	0,18	0,48	0,15	0,58	0,14
BB	0,84	0,89	-0,37	0,38	0,99	-0,18	0,71	-0,20	0,67	-0,10
CLOV	0,84	0,90	-0,18	0,67	0,98	0,10	0,86	0,00	0,84	0,04
NAKD	0,89	0,91	-0,36	0,57	1,00	-0,10	0,78	-0,13	0,76	0,22
NIO	0,67	0,81	-0,13	0,75	0,95	0,20	0,78	0,09	0,79	0,07
NOK	0,93	0,94	-0,12	0,66	1,00	0,05	0,84	0,09	0,85	0,36
PLTR	0,74	0,84	0,05	0,80	0,94	0,39	0,82	0,22	0,84	0,15
RKT	0,94	0,97	0,28	0,96	0,98	0,53	0,97	0,41	0,98	0,45
SNDL	0,81	0,90	-0,21	0,75	0,98	0,02	0,86	-0,06	0,86	-0,15
SPCE	0,70	0,87	-0,17	0,72	0,92	0,20	0,88	0,02	0,90	0,05
SPY	0,73	0,91	-0,37	0,80	0,86	-0,01	0,70	-0,23	0,80	-0,27
TSLA	0,54	0,87	-0,09	0,04	0,80	0,08	0,11	0,02	0,13	0,10

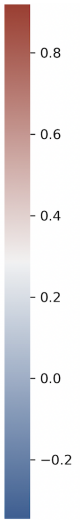

TABLE 8 – Scores de corrélation - Twitter

Deuxièmement, nous pouvons déjà observer des scores de corrélation élevés entre la variation journalière ou le volume journalier et le nombre de messages positifs, négatifs ou neutres. Cela peut indiquer une relation, qui sera analysée dans les sections prochaines. Finalement, les scores de corrélation semblent être plus élevés pour le volume journalier que pour la variation journalière. Ceci peut indiquer que les résultats des régressions incluant le volume journalier et le nombre de messages positifs, négatifs ou neutres présenteront probablement des résultats plus probants en comparaison à la variation journalière.

4.2 Régression linéaire

Intéressons-nous maintenant à l'application de régressions linéaires sur nos données. Pour chaque action, les équations présentées dans la section 3.5.2, ou des variations de ces dernières ont été appliquées. Celles présentant les meilleurs résultats ont été conservées. Les variations peuvent soit être une suppression de variable pour contrer de la multicolinéarité ou l'application de la fonction logarithme népérien sur nos variables suite à l'analyse de l'allure de la relation entre nos variables prises paire par paire.

Les tableaux 9, 10 et 11, 12 présentent les meilleurs résultats des régressions linéaires pour chaque action et chaque plateforme, et ce pour chaque sens de relation considéré. Tous les coefficients des équations sélectionnées présentent un niveau de significativité égal ou inférieur à 5%. Le niveau de significativité inférieur ou égal à 5% nous informe qu'il y a 95% de chance ou plus que les coefficients de nos variables explicatives soient différents de 0 et que ces variables aient, donc, un effet sur la variable à expliquer. Certaines lignes sont grisées, car aucun résultat satisfaisant n'a pu être trouvé dans certains cas.

	Equation	R_squared	Coef b	Coef c
AAPL	$Var = a + bPos + cNeg$	0,36	0,01	-0,02
AMD	$Vol = a + bNeu + cNeg$	0,43	50118,69	-69800,43
BB	$Vol = a + bNeu + cNeg$	0,69	113587,04	-163559,77
CLOV	$Vol = a + bPos + cNeg$	0,75	194864,52	-54526,69
NAKD	$Vol = a + bPos + cNeg$	0,68	39617,82	-24800,23
NIO	$Vol = a + bPos + cNeg$	0,70	156291,03	393801,36
NOK	$Vol = a + bNeu + cNeg$	0,87	90633,50	-123897,23
PLTR	$Vol = a + bPos + cNeg$	0,75	117062,35	361311,24
RKT	$Vol = a + bPos + cNeg$	0,97	103732,92	209841,59
SNDL	$Vol = a + bPos + cNeg$	0,76	427399,61	205846,97
SPCE	$Vol = a + bNeu + cNeg$	0,83	90745,94	-133920,72
SPY	$Vol = a + bPos + cNeg$	0,67	43469,57	121200,68
TSLA				

TABLE 9 – Régressions linéaires du sentiment des messages sur les informations financières - Twitter

	Equation	R_squared	Coef b
AAPL	$Neg = a + bVol$	0,35	0,0000013
AMD	$Neu = a + bVol$	0,34	0,0000097
BB	$Pos + Neg = a + bVol$	0,51	0,0000040
CLOV	$Pos + Neg = a + bVol$	0,74	0,0000043
NAKD	$Pos + Neg = a + bVol$	0,61	0,0000258
NIO	$Neu + Neg = a + bVol$	0,65	0,0000054
NOK	$Pos + Neg = a + bVol$	0,70	0,0000065
PLTR	$Neu + Neg = a + bVol$	0,73	0,0000082
RKT	$Neu + Neg = a + bVol$	0,97	0,0000154
SNDL	$Pos + Neg = a + bVol$	0,75	0,0000021
SPCE	$Neu + Neg = a + bVol$	0,79	0,0000126
SPY	$Neu + Neg = a + bVol$	0,66	0,0000178
TSLA			

TABLE 10 – Régressions linéaires des informations financières sur le sentiment des messages - Twitter

	Equation	R_squared	Coef b	Coef c
AAPL				
AMD				
BB	$Vol = a + bPos + cNeg$	0,35	131385,30	-73394,32
CLOV				
NAKD	$Vol = a + bPos + cNeg$	0,95	438618,60	-70893,49
NIO	$Vol = a + blog(Pos) + cNeu$	0,54	9748602,17	149819,57
NOK	$Vol = a + blog(Pos) + cNeu$	0,92	6238716,74	53320,48
PLTR				
RKT				
SNDL	$Vol = a + bPos + cNeg$	0,76	3769813,77	-533558,70
SPCE				
SPY				
TSLA	$Vol = a + bPos + cNeu$	0,37	-107758,171	55767,633

TABLE 11 – Régressions linéaires du sentiment des messages sur les informations financières - Reddit

	Equation	R_squared	Coef b
AAPL			
AMD			
BB	$Pos = a + bVol$	0,29	0,0000059
CLOV			
NAKD	$Pos + Neg = a + bVol$	0,83	0,0000047
NIO	$Pos + Neg = a + bVol$	0,55	0,0000009
NOK	$Pos + Neg = a + bVol$	0,94	0,0000048
PLTR			
RKT			
SNDL	$Pos + Neg = a + bVol$	0,75	0,0000002
SPCE			
SPY			
TSLA	$Neu + Neg = a + bVol$	0,35	0,0000157

TABLE 12 – Régressions linéaires des informations financières sur le sentiment des messages - Reddit

Prenons l'action RKT dans le tableau 9, l'interprétation des résultats se fait de la manière suivante : toutes choses étant égales par ailleurs, pour chaque message positif supplémentaire, le volume journalier de l'action RKT augmente d'environ 103732 transactions. De la même manière, toutes choses étant égales par ailleurs, pour chaque message

négatif supplémentaire, le volume journalier de l'action RKT augmente d'environ 209842 transactions. Le R^2 de 97% nous informe que 97% de la variance de la variable dépendante est expliquée par les variables indépendantes de notre modèle.

À partir de ces résultats, différentes observations peuvent-être faites. Tout d'abord, la variation journalière ne semble pas expliquer ou être expliquée par le sentiment des messages. Cela infirme alors les hypothèses H1a à H1c, H3a à H3c, H5a à H5c et H7a à H7c. En effet, la majorité des régressions réalisées à l'aide de la variation journalière présentaient des résultats significatifs, mais avec un R^2 très proche de 0. Cependant, le volume journalier semble lui être une variable intéressante pour rechercher une relation avec le sentiment des messages. La plupart des régressions présentant des résultats prometteurs sont en effet basées sur cette variable.

Il est aussi important de noter que les résultats sont bien meilleurs avec les données de Twitter que de Reddit. En effet, pour beaucoup d'actions sur Reddit, aucun résultat satisfaisant n'a pu être observé. Pour Twitter, la majorité des R^2 se situe au-dessus de 60%. Pour Reddit, seulement 3 des 13 actions présentent des résultats similaires. Cela pourrait s'expliquer par une quantité de données moins importante (voir section 3.3) et une moins bonne répartition annuelle des messages (voir annexe B). Nous pouvons aussi observer que si les données ne permettent pas d'obtenir des résultats satisfaisants, cela est le cas, quel que soit le sens de relation considéré.

Les hypothèses H2c, H4c, H6c, H8c concernant l'impact inexistant des messages neutres ne peuvent pas être confirmées. En effet, la variable Neu se retrouve dans les équations présentant les meilleurs résultats avec des coefficients significatifs et différents de 0.

Les hypothèses H2a à H2b, H4a à H4b, H6a à H6b et H8a à H8b semblent être confirmées, en partie, pour la majorité des actions. Étant donné que les R^2 oscillent entre 60% et 90%, d'autres variables non présentes dans ces modèles rentrent aussi en jeu dans l'explication du marché boursier. Il n'en reste pas moins que le sentiment des messages et plus particulièrement leur volume semble être un bon proxy du volume journalier. Il n'est par contre pas certain que ces résultats soient généralisables à n'importe quelle action, mais ces derniers vont dans le sens d'une possible relation entre le volume journalier du marché boursier et le sentiment des messages postés par les internautes.

Finalement, les régressions pour certaines actions semblent mieux performer que d'autres. Cela peut être lié aux scores de corrélation : nous observons des coefficients plus élevés entre les variables indépendantes et dépendantes pour les actions présentant de meilleurs résultats.

4.3 Test de causalité de Granger

La dernière analyse de cette section concerne le test de causalité de Granger. Les étapes présentées dans la section 3.5.3 ont été appliquées pour tester la relation entre chaque série temporelle. Comme pour la régression linéaire, les relations dans les deux sens ont été testées. Finalement, des tests de causalité de Granger avec des sommes de séries temporelles liées aux sentiments des messages ont aussi été réalisés.

Afin d'analyser les résultats, nous procédons de la suite : nous calculons pour chaque test de Granger le nombre d'actions pour lesquelles nous pouvons observer une causalité au sens de Granger, c'est-à-dire, pour lesquelles le test de Granger présente un niveau de significativité inférieur à 5%. Nous sélectionnons ensuite les tests de causalité présentant plus de la moitié d'actions significatives pour Reddit et Twitter (tableaux 13 et 14).

Relation	Number of stocks with significative results
$Var \sim Neg$	10
$Var \sim Pos + Neg + Neu$	10
$Var \sim Pos + Neg$	10
$Var \sim Pos + Neu$	10
$Var \sim Neg + Neu$	9
$Var \sim Neu$	9
$Var \sim Pos$	8
$Neg \sim Var$	8
$Neu \sim Var$	8
$Pos + Neg + Neu \sim Var$	8
$Pos + Neg \sim Var$	8
$Pos + Neu \sim Var$	8
$Neg + Neu \sim Var$	8
$Pos \sim Vol$	8
$Neg \sim Vol$	8
$Pos + Neg + Neu \sim Vol$	8
$Pos + Neg \sim Vol$	8
$Pos + Neu \sim Vol$	8
$Neu \sim Vol$	7
$Neg + Neu \sim Vol$	7

TABLE 13 – Tests de causalité de Granger présentant plus de la moitié d'actions significatives - Twitter

Relation	Number of stocks with significative results
$Pos + Neu \sim Vol$	8
$Neg + Neu \sim Vol$	7
$Pos + Neg + Neu \sim Vol$	7
$Neu \sim Vol$	7

TABLE 14 – Tests de causalité de Granger présentant plus de la moitié d'actions significatives - Reddit

Nous pouvons tout d'abord observer, comme c'était le cas pour la régression linéaire, que les résultats sont beaucoup plus intéressants avec les données issues de Twitter. Ensuite et contrairement aux régressions linéaires, ce sont ici les relations avec la variation journalière qui semblent offrir les meilleurs résultats pour Twitter. Pour Reddit, c'est le volume journalier qui génère les meilleurs résultats. Le volume journalier offre néanmoins des résultats satisfaisants pour Twitter aussi.

Cette contradiction dans les résultats pourrait trouver une partie de sa réponse dans l'analyse effectuée. En effet, le test de causalité de Granger vérifie si une série temporelle est utile pour en prévoir une autre, mais n'informe pas de la force de la relation comme peut le faire une régression linéaire via ses coefficients. Nous ne pouvons pas non plus conclure de la proportion de la variance expliquée par le modèle avec le R^2 comme avec une régression linéaire. Des résultats significatifs avec la variation journalière ont été observés avec la régression linéaire, mais ces résultats ne présentaient pas un R^2 satisfaisant.

Nous pouvons cependant observer qu'il existe une causalité au sens de Granger entre la variation journalière et le nombre de messages positifs, négatifs et neutres, semblant aller dans le sens des équations H1a, H1b, H3a, H3b, H5a, H5b, H7a et H7b. Cependant, nous ne pouvons pas confirmer les hypothèses H1c, H3c, H5c, H7c concernant l'impact inexistant des messages neutres. En effet, la variable Neu se retrouve dans les tests présentant les meilleurs résultats.

Nous observons aussi une causalité au sens de Granger entre le volume journalier et le sentiment des messages pour un peu plus de la moitié des actions. Ces relations ne sont cependant observées que dans un sens : le volume journalier permet de prédire le sentiment des messages. Ces observations vont alors dans le sens des hypothèses H6a, H6b, H8a et H8b.

4.4 Récapitulatif des hypothèses testées

Après revue des résultats de nos différentes analyses, nous présentons dans cette section le tableau 15 reprenant les différentes hypothèses vérifiées ou infirmées par la régression linéaire (LR) ou le test de causalité de Granger (GR).

Number	Hypotheses	Verified by LR	Verified by GR
H1a	Le nombre de messages **positifs** postés sur Reddit à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.	✗	✓
H1b	Le nombre de messages **négatifs** postés sur Reddit à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.	✗	✓
H1c	Le nombre de messages **neutres** postés sur Reddit à propos d'une action **n'impacte pas** la **variation journalière** du cours boursier de cette dernière.	✗	✗
H2a	Le nombre de messages **positifs** postés sur Reddit à propos d'une action **impacte** le **volume journalier** de cette dernière.	✓	✗
H2b	Le nombre de messages **négatifs** postés sur Reddit à propos d'une action **impacte** le **volume journalier** de cette dernière.	✓	✗
H2c	Le nombre de messages **neutres** postés sur Reddit à propos d'une action **n'impacte pas** le **volume journalier** de cette dernière.	✗	✗
H3a	Le nombre de messages **positifs** postés sur Twitter à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.	✗	✓
H3b	Le nombre de messages **négatifs** postés sur Twitter à propos d'une action **impacte** la **variation journalière** du cours boursier de cette dernière.	✗	✓
H3c	Le nombre de messages **neutres** postés sur Twitter à propos d'une action **n'impacte pas** la **variation journalière** du cours boursier de cette dernière.	✗	✗
H4a	Le nombre de messages **positifs** postés sur Twitter à propos d'une action **impacte** le **volume journalier** du cours boursier de cette dernière.	✓	✗

H4b	Le nombre de messages **négatifs** postés sur Twitter à propos d'une action **impacte** le **volume journalier** du cours boursier de cette dernière.	✓	✗
H4c	Le nombre de messages **neutres** postés sur Twitter à propos d'une action **n'impacte pas** le **volume journalier** du cours boursier de cette dernière.	✗	✗
H5a	La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **positifs** postés sur Reddit à propos de cette dernière.	✗	✓
H5b	La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **négatifs** postés sur Reddit à propos de cette dernière.	✗	✓
H5c	La **variation journalière** du cours boursier d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Reddit à propos de cette dernière.	✗	✗
H6a	Le **volume journalier** d'une action **impacte** le nombre de messages **positifs** postés sur Reddit à propos de cette dernière.	✓	✓
H6b	Le **volume journalier** d'une action **impacte** le nombre de messages **négatifs** postés sur Reddit à propos de cette dernière.	✓	✓
H6c	Le **volume journalier** d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Reddit à propos de cette dernière.	✗	✗
H7a	La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **positifs** postés sur Twitter à propos de cette dernière.	✗	✓
H7b	La **variation journalière** du cours boursier d'une action **impacte** le nombre de messages **négatifs** postés sur Twitter à propos de cette dernière.	✗	✓
H7c	La **variation journalière** du cours boursier d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Twitter à propos de cette dernière.	✗	✗
H8a	Le **volume journalier** d'une action **impacte** le nombre de messages **positifs** postés sur Twitter à propos de cette dernière.	✓	✓

H8b	Le **volume journalier** d'une action **impacte** le nombre de messages **négatifs** postés sur Twitter à propos de cette dernière.	✓	✓
H8c	Le **volume journalier** d'une action **n'impacte pas** le nombre de messages **neutres** postés sur Twitter à propos de cette dernière.	✗	✗

TABLE 15 – Hypothèses confirmées ou infirmées - Résumé

Nous observons que les deux tests vont dans le même sens en confirmant que le volume journalier impacte le nombre de messages positifs ou négatifs, et ce à la fois pour Reddit et Twitter. Les deux analyses infirment aussi l'absence d'impact des messages neutres, quel que soit le sens de relation considéré.

4.5 Interprétation entre les actions

Afin d'analyser pourquoi certaines actions semblent mieux performer que d'autres lors des analyses, nous voulons analyser et comparer dans cette dernière partie les actions ayant le mieux performé avec celles présentant les moins bons résultats de manière générale.

À cette fin, et suite à l'analyse des résultats présentés précédemment, nous sélectionnons les actions RKT et TSLA sur Twitter, mais les résultats présentés ici peuvent être observés avec toutes les actions analysées au sein de ce mémoire. Étant donné la quantité de données plus importante sur Twitter, cette comparaison semble plus pertinente que d'utiliser des données issues de Reddit où la quantité de données varie beaucoup plus d'une action à l'autre.

Nous pouvons tout d'abord observer les figures 8 et 9 représentant les données pour chaque action au cours de l'année. Nous comprenons directement la différence de résultats en observant nos données : les variations des différentes séries temporelles surviennent au même moment pour RKT, ce qui n'est pas particulièrement le cas pour TSLA. Nous observons de manière générale beaucoup plus de mouvements dans les données de l'action TSLA.

Prenons deux autres actions : AAPL et NOK. La première n'a pas non plus particulièrement bien performé lors des analyses alors que la seconde présente des résultats plus satisfaisants. Les figures 10 et 11 présentent les données pour ces actions.

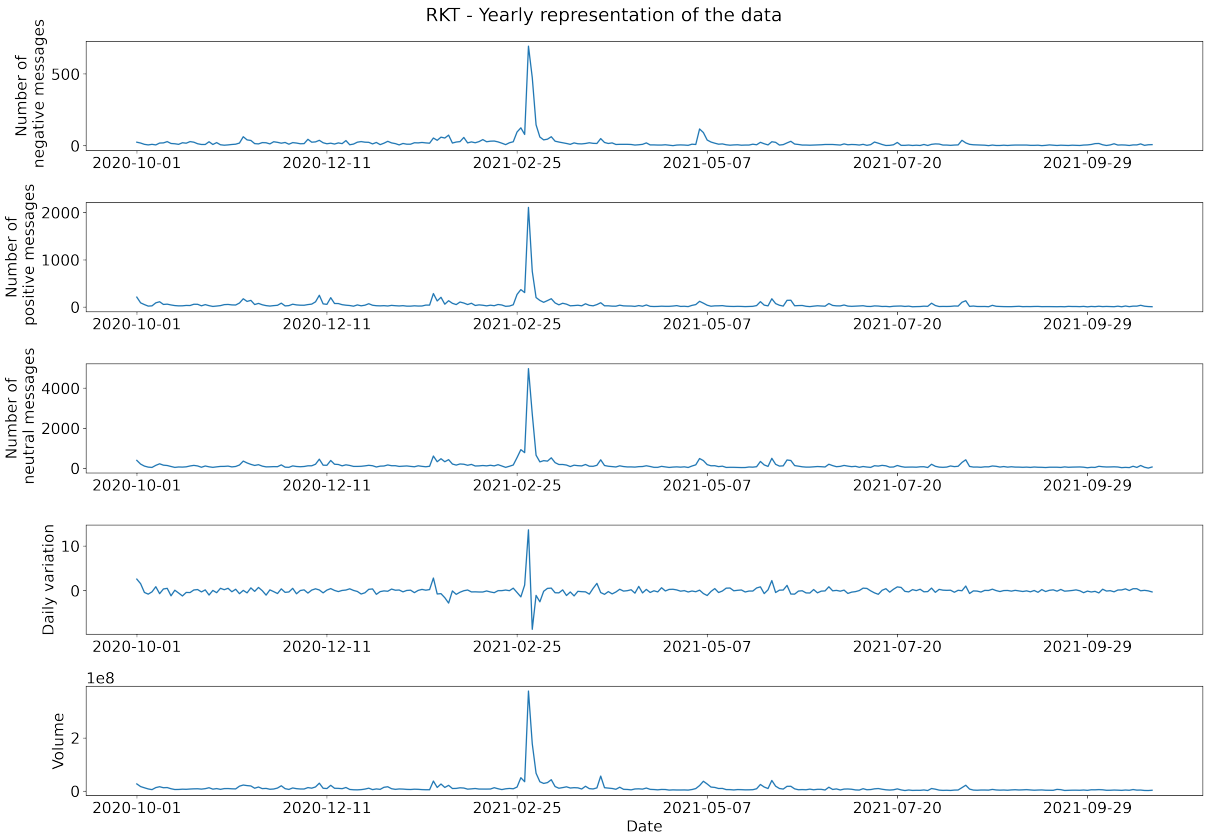

FIGURE 8 – RKT - Représentation annuelle des données

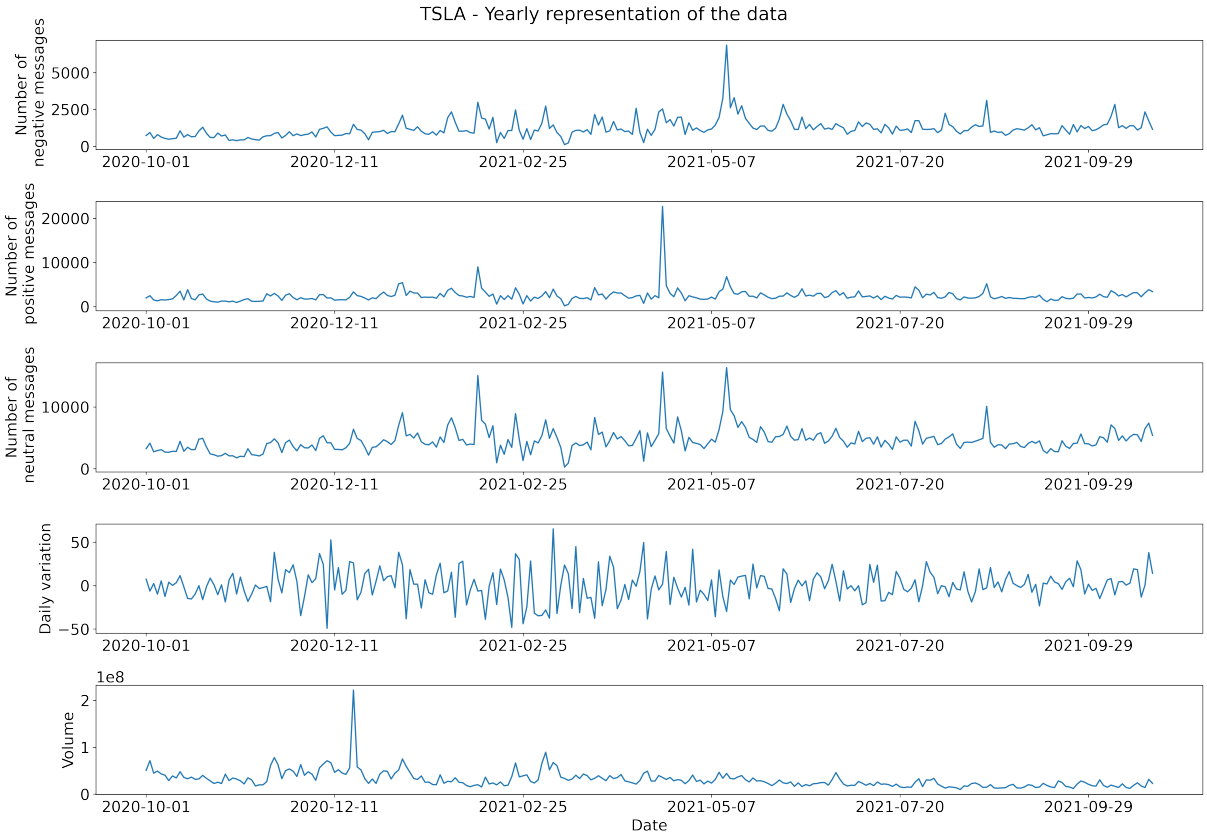

FIGURE 9 – TSLA - Représentation annuelle des données

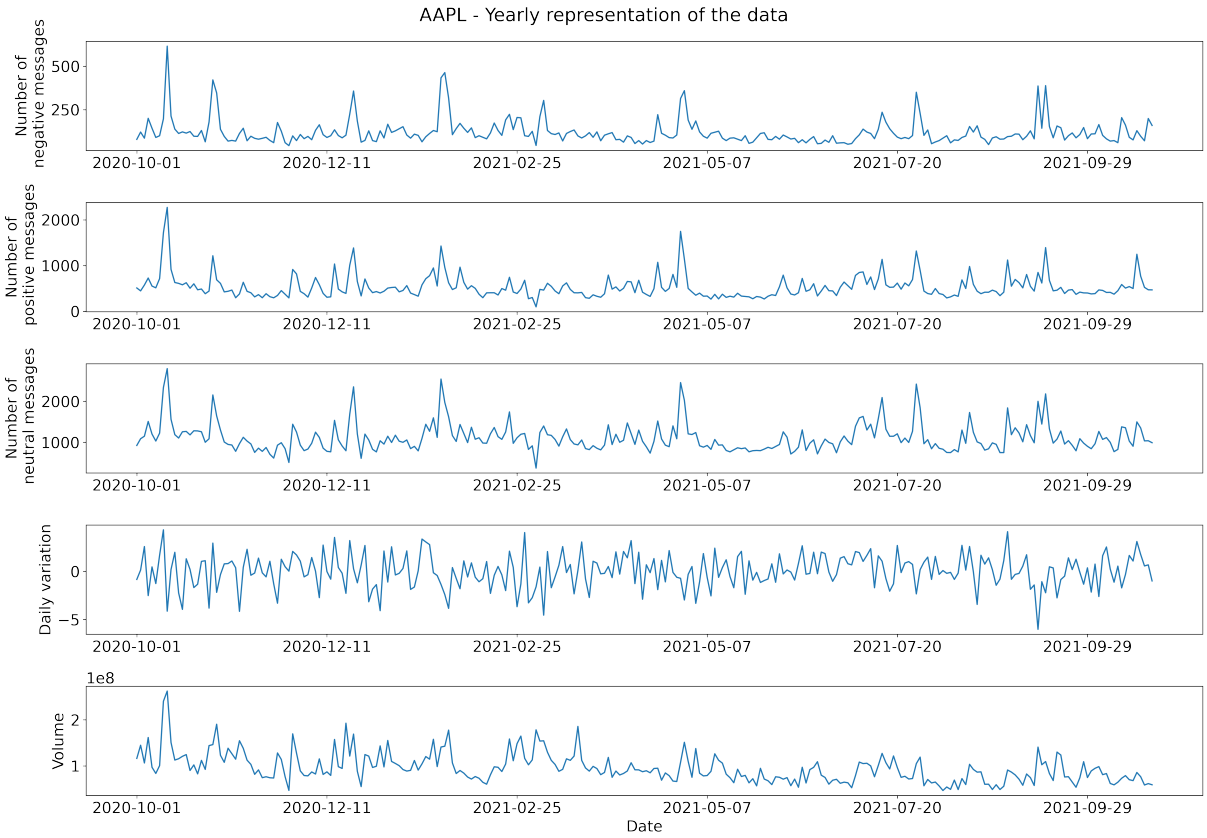

FIGURE 10 – AAPL - Représentation annuelle des données

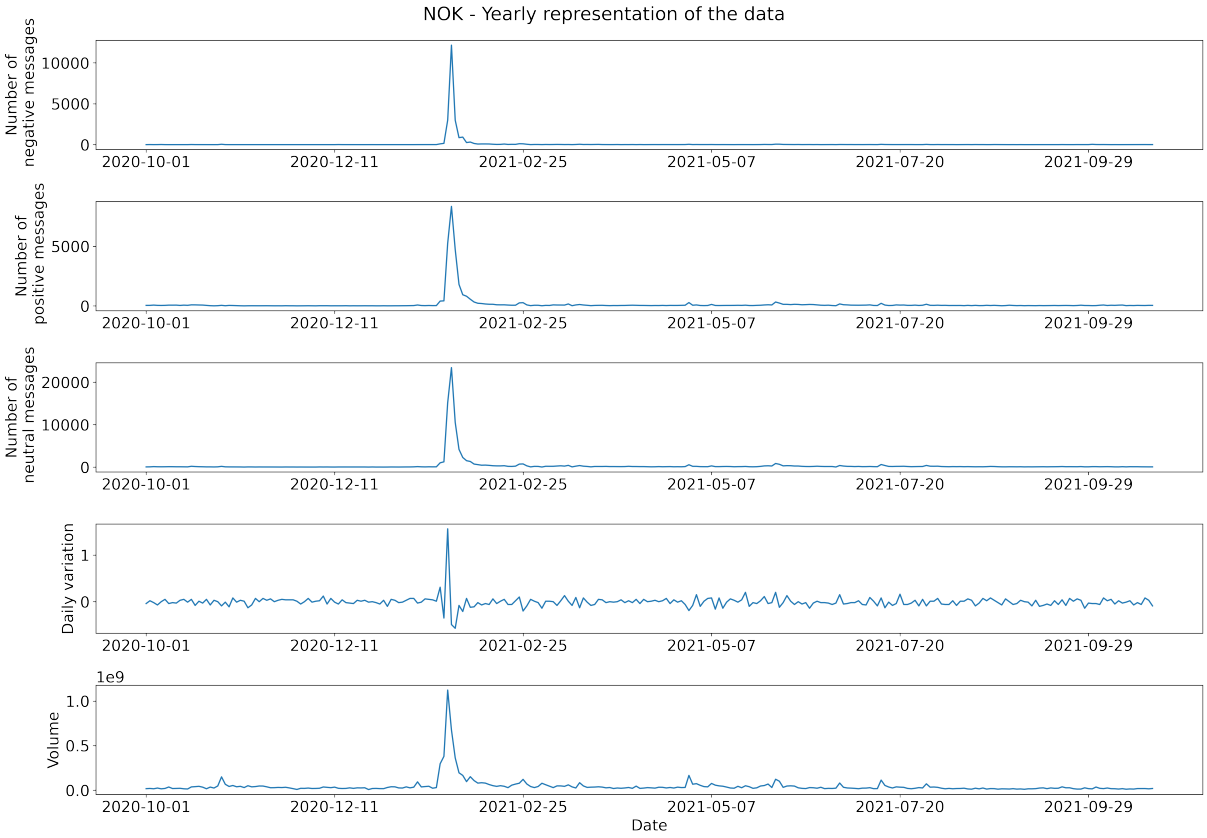

FIGURE 11 – NOK - Représentation annuelle des données

Comme précédemment, l'action AAPL est caractérisée par beaucoup de mouvements, avec néanmoins quelques schémas pouvant être repérés dans les séries temporelles, alors que l'action NOK est caractérisée par une forte variation à un moment particulier et des mouvements assez calmes le reste de l'année.

Nous proposons l'interprétation suivante à partir des informations en notre possession. AAPL et TSLA sont deux actions d'entreprises à grande ampleur médiatique contrairement aux actions RKT et NOK. Cette couverture médiatique, liée à l'actualité de l'entreprise, implique deux choses : d'une part, la réaction des internautes sur les réseaux sociaux va être plus forte par rapport à l'entreprise et à l'action, et, d'autre part, les investisseurs qui suivent cette actualité auront plus de raisons de vendre ou d'acheter l'action en fonction de celle-ci. Ainsi, les actions moins connues et moins médiatisées semblent être moins sujettes à de grosses variations tout au long de l'année contrairement aux actions plus médiatisées. La figure 12 représente l'interaction entre les différents éléments nous permettant d'appuyer cette interprétation.

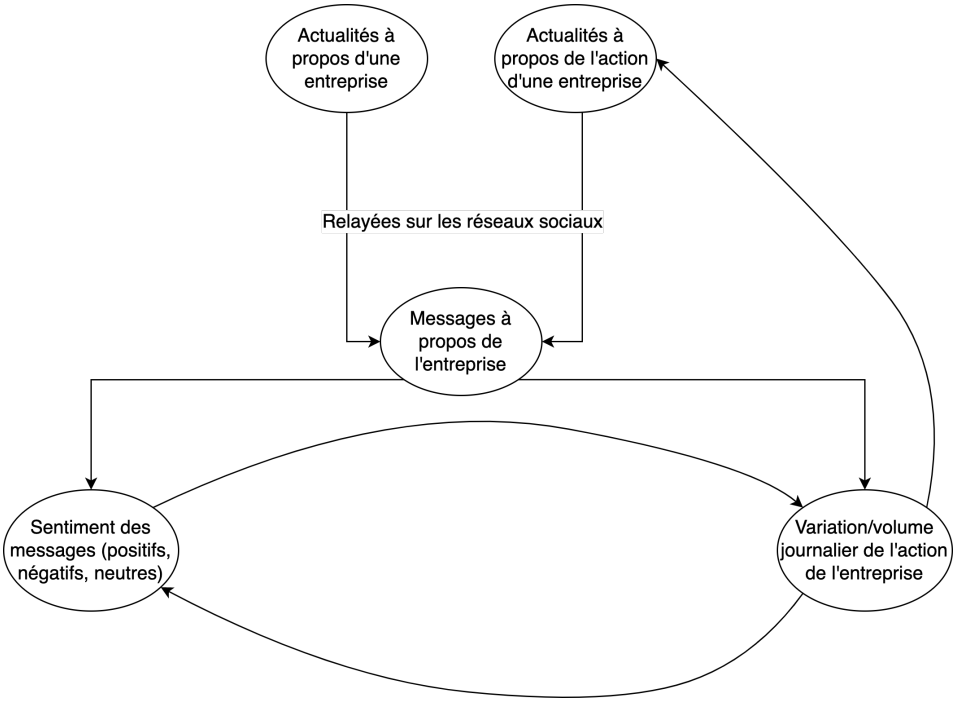

FIGURE 12 – Représentation graphique - Interprétation des résultats

Cette interprétation semble donner sens aux résultats observés entre les actions lors des analyses précédentes, mais reste néanmoins hypothétique. Une recherche approfondie de l'impact de la médiatisation d'une action sur la relation entre les réseaux sociaux et le marché boursier semble nécessaire pour pouvoir confirmer cette interprétation. Cette comparaison entre actions nous permet néanmoins d'observer visuellement une réelle différence entre les actions ayant bien performé et celles ayant moins bien performé.

Conclusion, limites et recommandations

Partis de la volonté d'analyser l'impact des internautes sur le marché boursier et de déterminer, s'il y a impact, le sens de ce dernier, c'est aussi une méthodologie à appliquer lors de recherches similaires que nous proposons dans ce mémoire. En effet, après avoir sélectionné nos deux plateformes, Reddit et Twitter, une revue de la littérature nous a permis de comprendre et analyser les recherches déjà effectuées dans ce domaine. La méthodologie à appliquer a ensuite dû être soigneusement choisie en adéquation avec les données avec lesquelles nous allions travailler.

La structure de la méthodologie proposée est alors la suivante : l'extraction, le nettoyage et l'exploration des données, suivis de l'analyse de sentiments nous ont permis de former un ensemble de données et de caractériser les messages selon leurs sentiments. Les résultats de cette dernière analyse ont alors pu être utilisés dans des analyses cherchant à déterminer si une relation existe entre le sentiment des messages et le marché boursier.

Trois analyses principales ont été réalisées : des tests de corrélation entre nos séries temporelles, des régressions linéaires multiples ainsi que des tests de causalité de Granger. Les régressions linéaires et tests de causalités de Granger présentent des résultats venant confirmer l'existence d'une relation entre le sentiment des messages et le marché boursier. Des relations statistiquement significatives ont en effet pu être observées entre le volume journalier (ou la variation journalière dans une moindre mesure) et le nombre de messages positifs, négatifs et/ou neutres.

Cette relation semble aussi se dérouler dans les deux sens : la collaboration des internautes semble pouvoir influencer le marché boursier, mais ce dernier influence aussi les messages postés par les internautes. De plus, Twitter semble être une source à privilégier pour l'analyse du marché boursier dû à une quantité de données disponible importante et un afflux de données continu.

Pour finir, afin d'analyser les différences de résultats entre actions, une comparaison a été réalisée. Nous avons pu ainsi observer que des actions présentant de fortes variations tout au long de l'année et qui semblent avoir une couverture médiatique plus importante

présentent des résultats moins encourageants lors des analyses.

Limites

Les résultats observés dans ce mémoire présentent cependant certaines limites. En effet, bien que des résultats prometteurs aient pu être constatés, ces derniers restent néanmoins dépendants de l'action étudiée : des variations dans les résultats sont présentes entre chaque action. Il est aussi intéressant de noter que la période considérée ne représente qu'une année, pouvant ainsi affecter les résultats présentés ici.

Le nombre d'actions considéré, 13, ne permet pas non plus de généraliser les résultats présentés ici. Une recherche à plus grande échelle, comptabilisant un grand nombre d'actions, serait nécessaire afin de pouvoir généraliser les observations.

Il est aussi important de noter que les analyses réalisées dans ce mémoire se concentrent avant tout sur des analyses statistiques et permettent d'observer si un lien ou une relation existe entre nos séries temporelles. Bien que les régressions linéaires permettent d'observer la force de l'impact, le but de ce mémoire n'était pas d'analyser de manière détaillée le sens de variation des variables lors de cet impact.

Afin d'aller plus loin, des algorithmes d'apprentissage automatique comme des forêts aléatoires ou réseaux de neurones n'ont pas pu être appliqués dans ce mémoire dû à la limite de données sur une année. Des analyses sur de plus grandes périodes pourraient apporter des résultats intéressants comme observé dans certaines publications scientifiques (Swathi et al., 2022).

Recommandations managériales

Nous avons pu voir qu'une relation existe entre les messages des internautes et le cours boursier d'actions, mais qu'en est-il d'un point de vue managérial ?

La connaissance de cette relation permet tout d'abord une meilleure compréhension du marché boursier. Bien que beaucoup d'autres facteurs rentrent en compte, une analyse approfondie des réseaux sociaux pourrait permettre une meilleure prévision de la performance d'un actif financier. Les résultats de ce mémoire semblent en effet encourageants et pourraient mener, conjointement à des analyses plus poussées, à un tableau de bord basé sur le sentiment des messages à propos d'une action.

Le but n'étant bien entendu pas de prédire le cours boursier d'une action afin de géné-

rer de l'argent, mais bien d'offrir un outil aux différentes parties prenantes s'intéressant à une action, que ce soit l'entreprise elle-même ou l'actionnaire, permettant de suivre le sentiment général à propos d'une action afin de mener à une meilleure gestion. Des solutions semblent déjà exister (« Home », s. d. ; « Market Sentiment Tracking Platform », 2020 ; « Social Media Sentiment Analysis for Stocks - SocialSentiment.io », s. d.), mais sont plutôt orientées vers l'investisseur voulant réaliser de meilleurs placements et ne semblent pas forcément être utilisées dans le monde de l'entreprise dans un but d'aide à la gestion.

Comme nous avons pu le voir durant la revue de littérature et ce mémoire, l'analyse de l'impact sociétal sur le marché boursier est un sujet en pleine effervescence et qui semble offrir de nouveaux moyens d'interpréter et comprendre le fonctionnement de ce dernier. Les analyses de ce mémoire semblent confirmer ceci, mais mettent aussi en avant la nécessité d'analyses plus complexes afin de comprendre pleinement ce phénomène.

Bibliographie

About GameStop | Gamestop Corp. (s. d.). Récupérée 31 août 2021, à partir de https://news.gamestop.com/about-gamestop

Adhikary, R. (2020). Testing for Granger Causality Using Python. Récupérée 25 avril 2022, à partir de https://rishi-a.github.io/2020/05/25/granger-causality.html

Advanced Micro Devices, Inc. (AMD) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/AMD/

Alliance Bernstein. (2013). How Does Short Selling Work ? Récupérée 2 septembre 2021, à partir de https://www.alliancebernstein.com/Microsites/ABI/US/Alternatives/portal/full_short_sell.html

Anderson, K. E. (2015). Ask me anything : what is Reddit ? *Library Hi Tech News*, *32*(5), 8-11. https://doi.org/10.1108/LHTN-03-2015-0018

Annual Reports | Financial Information | Investor Relations | Gamestop Corp. (s. d.). Récupérée 1 septembre 2021, à partir de https://news.gamestop.com/financial-information/annual-reports

api - reddit.com. (s. d.). Récupérée 6 septembre 2021, à partir de https://www.reddit.com/wiki/api

API · reddit-archive/reddit Wiki. (s. d.). Récupérée 5 février 2022, à partir de https://github.com/reddit-archive/reddit

Apple Inc. (AAPL) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/AAPL/

Aroussi, R. (2019). Reliably download historical market data from Yahoo ! Finance with Python. Récupérée 22 septembre 2021, à partir de https://aroussi.com/post/python-yahoo-finance

Aroussi, R. (2022). Download market data from Yahoo ! Finance's API. Récupérée 6 février 2022, à partir de https://github.com/ranaroussi/yfinance

Baker, M. & Wurgler, J. (2007). Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, *21*(2), 129-152. https://doi.org/10.1257/jep.21.2.129

Banton, C. (2021). Serial Correlation Definition. Récupérée 25 avril 2022, à partir de https://www.investopedia.com/terms/s/serial-correlation.asp

Barbieri, F., Camacho-Collados, J., Neves, L. & Espinosa-Anke, L. (2020). TweetEval : Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv :2010.12421*. Récupérée 10 avril 2022, à partir de http://arxiv.org/abs/2010.12421

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. (2020). The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 830-839. Récupérée 5 février 2022, à partir de https://ojs.aaai.org/index.php/ICWSM/article/view/7347

Bhattacharyya, I. (2020). Support Vector Regression Or SVR. Récupérée 23 septembre 2021, à partir de https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

Biais de variable omise. (s. d.). Récupérée 24 avril 2022, à partir de http://stringfixer.com/fr/Omitted_variable_bias

BlackBerry Limited (BB) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/BB/

Boylston, C., Palacios, B., Tassev, P. & Bruckman, A. (2021). WallStreetBets : Positions or Ban. *arXiv :2101.12110*. Récupérée 26 février 2021, à partir de http://arxiv.org/abs/2101.12110

Bricken, A. (2021a). Does BERT Need Clean Data ? Part 1 : Data Cleaning. Récupérée 23 avril 2022, à partir de https://towardsdatascience.com/part-1-data-cleaning-does-bert-need-clean-data-6a50c9c6e9fd

Bricken, A. (2021b). Does BERT Need Clean Data ? Part 2 : Classification. Récupérée 23 avril 2022, à partir de https://towardsdatascience.com/does-bert-need-clean-data-part-2-classification-d29adf9f745a

Caie, P. D., Dimitriou, N. & Arandjelović, O. (2021). Precision medicine in digital pathology via image analysis and machine learning. In S. Cohen (Éd.), *Artificial Intelligence and Deep Learning in Pathology* (p. 149-173). Elsevier. https://doi.org/10.1016/B978-0-323-67538-3.00008-7

Cambridge Dictionary. (s. d.). microblogging. Récupérée 30 août 2021, à partir de https://dictionary.cambridge.org/dictionary/english/microblogging

Cardiff NLP. (s. d.). cardiffnlp/twitter-roberta-base-sentiment · Hugging Face. Récupérée 10 avril 2022, à partir de https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Chohan, U. W. (2021). *Counter-Hegemonic Finance : The Gamestop Short Squeeze* (SSRN Scholarly Paper N° ID 3775127). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3775127

Cliffe, C. (s. d.). What is a short squeeze ? Récupérée 6 septembre 2021, à partir de https://www.ig.com/en/trading-strategies/what-is-a-short-squeeze-200701

Closing price Definition. (s. d.). Récupérée 6 septembre 2021, à partir de https://www.nasdaq.com/glossary/c/closing-price

Clover Health Investments, Corp. (CLOV) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/CLOV/

Corporate Finance Institute. (s. d.). Cointegration. Récupérée 25 avril 2022, à partir de https://corporatefinanceinstitute.com/resources/knowledge/other/cointegration/

Corrélation de Pearson. (s. d.). Récupérée 22 septembre 2021, à partir de http://www.biostat.ulg.ac.be/pages/Site_r/corr_pearson.html

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv :1810.04805*. Récupérée 10 avril 2022, à partir de http://arxiv.org/abs/1810.04805

Di Muzio, T. (2021). *GameStop Capitalism. Wall Street vs. The Reddit Rally (Part I)* (Working Paper). Toronto : The Bichler et Nitzan Archives. Récupérée 6 août 2021, à partir de https://www.econstor.eu/handle/10419/229951

Éditions Larousse. (s. d.-a). Définitions : mème - Dictionnaire de français Larousse. Récupérée 7 août 2021, à partir de https://www.larousse.fr/dictionnaires/francais/m%C3%A8me/10910896

Éditions Larousse. (s. d.-b). Définitions : option - Dictionnaire de français Larousse. Récupérée 7 août 2021, à partir de https://www.larousse.fr/dictionnaires/francais/option/56260

Edward, A. (2021). An Extensive Guide to collecting tweets from Twitter API v2 for academic research using Python 3. Récupérée 12 septembre 2021, à partir de https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a

Fardeen, A. (2021). Generating Unigram, Bigram, Trigram and Ngrams in NLTK. Récupérée 23 septembre 2021, à partir de https://machinelearningknowledge.ai/generating-unigram-bigram-trigram-and-ngrams-in-nltk/

Fernando, J. (2021). What Is R-Squared ? Récupérée 22 septembre 2021, à partir de https://www.investopedia.com/terms/r/r-squared.asp

Financial news and views. (s. d.). Récupérée 8 août 2021, à partir de https://www.reddit.com/r/finance/

Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C. & Song, M. (2019). Mining of Textual Health Information from Reddit : Analysis of Chronic Diseases With Extracted Entities and Their Relations. *Journal of Medical Internet Research*, *21*(6), e12876. https://doi.org/10.2196/12876

Freehold, N. J. (2022). Cenntro Electric Group Announces the Change of its Trading Symbol from “NAKD” to “CENN”. Récupérée 6 février 2022, à partir de https:

//ir.cenntroauto.com/news-releases/news-release-details/cenntro-electric-group-announces-change-its-trading-symbol-nakd

Ganesan, K. (2019). What are Stop Words ? Récupérée 25 septembre 2021, à partir de https://www.opinosis-analytics.com/knowledge-base/stop-words-explained/

Getting access to the Twitter API. (s. d.). Récupérée 12 septembre 2021, à partir de https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api

Getting started — tweepy 3.10.0 documentation. (s. d.). Récupérée 12 septembre 2021, à partir de https://docs.tweepy.org/en/stable/getting_started.html

Gligorić, K., Anderson, A. & West, R. (2018). How Constraints Affect Content : The Case of Twitter's Switch from 140 to 280 Characters. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). Récupérée 9 août 2021, à partir de https://ojs.aaai.org/index.php/ICWSM/article/view/15079

GME Interactive Stock Chart | GameStop Corp. Stock - Yahoo Finance. (s. d.). Récupérée 1 septembre 2021, à partir de https://finance.yahoo.com/chart/GME/

Hayes, A. (2021a). Autoregressive Integrated Moving Average (ARIMA). Récupérée 12 septembre 2021, à partir de https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

Hayes, A. (2021b). Understanding Time Series. Récupérée 23 avril 2022, à partir de https://www.investopedia.com/terms/t/timeseries.asp

Hayes, A. (2022). Multicollinearity. Récupérée 25 avril 2022, à partir de https://www.investopedia.com/terms/m/multicollinearity.asp

Hentschel, M. & Alonso, O. (2014). Follow the money : A study of cashtags on Twitter. *First Monday*. https://doi.org/10.5210/fm.v19i8.5385

Home. (s. d.). Récupérée 6 mai 2022, à partir de https://stocksnips.net/

Homepage - Reddit. (s. d.). Récupérée 8 août 2021, à partir de https://www.redditinc.com/

index - investing. (s. d.). Récupérée 8 août 2021, à partir de https://www.reddit.com/r/investing/wiki/index

index - stocks. (s. d.). Récupérée 8 août 2021, à partir de https://www.reddit.com/r/stocks/wiki/index

Introduction to the Fundamentals of Vector Autoregressive Models - Aptech. (2021). Récupérée 25 avril 2022, à partir de https://www.aptech.com/blog/introduction-to-the-fundamentals-of-vector-autoregressive-models/

Jaffe, J. F., Westerfield, R. & Ma, C. (1989). A twist on the Monday effect in stock prices : Evidence from the U.S. and foreign stock markets. *Journal of Banking & Finance*, *13*(4), 641-650. https://doi.org/10.1016/0378-4266(89)90035-6

Kalman Filter. (2019). Récupérée 13 septembre 2021, à partir de https://deepai.org/machine-learning-glossary-and-terms/kalman-filter

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is Twitter, a social network or a news media ? *Proceedings of the 19th international conference on World wide web*, 591-600. https://doi.org/10.1145/1772690.1772751

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root : How sure are we that economic time series have a unit root ? *Journal of Econometrics*, *54*(1-3), 159-178. https://doi.org/10.1016/0304-4076(92)90104-Y

La Libre Eco avec AFP. (2021). Comment l'affaire GameStop a "démocratisé le trading". Récupérée 31 août 2021, à partir de https://www.lalibre.be/economie/placements/2021/07/28/six-mois-plus-tard-laffaire-gamestop-a-democratise-le-trading-VDG4KO53S5FMDKTSF3ZXJ6ARZI/

Lamy, C. (2021). Affaire GameStop : les utilisateurs de Reddit bien décidés à en finir avec Wall Street. *Le Monde.fr*. Récupérée 31 août 2021, à partir de https://www.lemonde.fr/pixels/article/2021/01/29/affaire-gamestop-les-utilisateurs-de-reddit-bien-decides-a-en-finir-avec-wall-street_6068149_4408996.html

Le Soir. (2021). Gamestop : la bousculade de Wall Street par des «boursicoteurs» continue. Récupérée 1 septembre 2021, à partir de https://www.lesoir.be/352402/article/2021-02-01/gamestop-la-bousculade-de-wall-street-par-des-boursicoteurs-continue

Liang, P. & Guo, S. (2015). Social interaction, Internet access and stock market participation—An empirical study in China. *Journal of Comparative Economics*, *43*(4), 883-901. https://doi.org/10.1016/j.jce.2015.02.003

Lina, F. (2020). NLP- Natural Language Processing : Introduction. Récupérée 23 septembre 2021, à partir de https://datascientest.com/introduction-au-nlp-natural-language-processing

Market Sentiment Tracking Platform. (2020). Récupérée 6 mai 2022, à partir de https://www.stockgeist.ai/

Merritt, R. (2022). What Is a Transformer Model ? Récupérée 10 avril 2022, à partir de https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

Mills, R. (2011). Researching Social News – Is reddit.com a mouthpiece for the 'Hive Mind', or a Collective Intelligence approach to Information Overload ? Sheffield Hallam University. Récupérée 1 septembre 2021, à partir de https://eprints.lancs.ac.uk/id/eprint/61646/

NIO Inc. (NIO) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/NIO/

Nisar, T. M. & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements : A short-window event study. *The Journal of Finance and Data Science*, *4*(2), 101-119. https://doi.org/10.1016/j.jfds.2017.11.002

Niveau de signification. (s. d.). Récupérée 24 avril 2022, à partir de https://toolbox.eupati.eu/glossary/niveau-de-signification/?lang=fr

Nofer, M. & Hinz, O. (2015). Using Twitter to Predict the Stock Market. *Business & Information Systems Engineering*, *57*(4), 229-242. https://doi.org/10.1007/s12599-015-0390-4

Nokia Corporation (NOK) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/NOK/

Oliveira, N., Cortez, P. & Areal, N. (2017). The impact of microblogging data for stock market prediction : Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125-144. https://doi.org/10.1016/j.eswa.2016.12.036

Opening price Definition. (s. d.). Récupérée 6 septembre 2021, à partir de https://www.nasdaq.com/glossary/o/opening-price

Palantir Technologies Inc. (PLTR) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/PLTR/

Patel, J. B. (2016). The January Effect Anomaly Reexamined In Stock Returns. *Journal of Applied Business Research (JABR)*, *32*(1), 317-324. https://doi.org/10.19030/jabr.v32i1.9540

Pedamkar, P. (2019). Polynomial Regression | Uses and Features of Polynomial Regression. Récupérée 13 septembre 2021, à partir de https://www.educba.com/polynomial-regression/

Pedamkar, P. (2020). Support Vector Regression | Learn the Working and Advantages of SVR. Récupérée 13 septembre 2021, à partir de https://www.educba.com/support-vector-regression/

pistocop/subreddit-comments-dl : Download subreddit comments. (s. d.). Récupérée 5 février 2022, à partir de https://github.com/pistocop/subreddit-comments-dl

Poston, D. L. (s. d.). Ordinary Least Squares Regression | Encyclopedia.com. Récupérée 13 septembre 2021, à partir de https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/ordinary-least-squares-regression

Pozzi, F. A., Fersini, E., Messina, E. & Liu, B. (2017). Challenges of Sentiment Analysis in Social Networks : An Overview. In F. A. Pozzi, E. Fersini, E. Messina & B. Liu (Éd.), *Sentiment Analysis in Social Networks* (p. 1-11). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-804412-4.00001-2

PRAW 7.4.0 documentation. (s. d.). Récupérée 6 septembre 2021, à partir de https://praw.readthedocs.io/en/stable/

PRAW : The Python Reddit API Wrapper. (2021). Récupérée 6 septembre 2021, à partir de https://github.com/praw-dev/praw

Quick Start — PRAW 7.4.0 documentation. (s. d.). Récupérée 12 septembre 2021, à partir de https://praw.readthedocs.io/en/stable/getting_started/quick_start.html

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. & Mozetič, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. *PLOS ONE*, *10*(9), e0138441. https://doi.org/10.1371/journal.pone.0138441

Random Forest. (s. d.). Récupérée 24 septembre 2021, à partir de https://dataanalyticspost.com/Lexique/random-forest/

Rate limits. (s. d.). Récupérée 5 février 2022, à partir de https://developer.twitter.com/en/docs/twitter-api/rate-limits

Raveschot, B. & Baudoux, N. (2021). #1 GameStop : quand les petits investisseurs prennent le pouvoir. Récupérée 31 août 2021, à partir de https://www.lecho.be/podcast/tracker/1-gamestop-quand-les-petits-investisseurs-prennent-le-pouvoir/10300973.html

reddit.com : documentation sur l'API. (s. d.). Récupérée 30 août 2021, à partir de https://www.reddit.com/dev/api

Return Definition. (s. d.). Récupérée 6 septembre 2021, à partir de https://www.nasdaq.com/glossary/r/return

Rigolin, V. H. (2018). What is Twitter ? How Do I Get Started ? Why Should I Become a User ? *Journal of the American Society of Echocardiography*, *31*(3), A31-A32. https://doi.org/10.1016/j.echo.2018.01.005

Rocket Companies, Inc. (RKT) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/RKT/

Rosenbaum, L., Dörr, A., Bauer, M. R., Boeckler, F. M. & Zell, A. (2013). Inferring multi-target QSAR models with taxonomy-based multi-task learning. *Journal of Cheminformatics*, *5*(1), 33. https://doi.org/10.1186/1758-2946-5-33

Sarker, S., Veremyev, A., Boginski, V. & Singh, A. (2019). Critical Nodes in River Networks. *Scientific Reports*, *9*(1), 11178. https://doi.org/10.1038/s41598-019-47292-4

Seth, A. (2007). Granger causality. *Scholarpedia*, *2*(7), 1667. https://doi.org/10.4249/scholarpedia.1667

Sharma, N. (2021). Understanding the Mathematics behind Decision Trees. Récupérée 24 septembre 2021, à partir de https://heartbeat.comet.ml/understanding-the-mathematics-behind-decision-trees-22d86d55906

Shleifer, A. & Summers, L. H. (1990). The Noise Trader Approach to Finance. *Journal of Economic Perspectives*, *4*(2), 19-33. https://doi.org/10.1257/jep.4.2.19

Social Media Sentiment Analysis for Stocks - SocialSentiment.io. (s. d.). Récupérée 6 mai 2022, à partir de https://socialsentiment.io/

SPDR S&P 500 ETF Trust (SPY) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/SPY/

Statista Research Department. (2021). Twitter : most users by country. Récupérée 2 septembre 2021, à partir de https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Stemming and lemmatization. (2009). Récupérée 25 septembre 2021, à partir de https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Subreddit — PRAW 7.4.1.dev0 documentation. (s. d.). Récupérée 6 septembre 2021, à partir de https://praw.readthedocs.io/en/latest/code_overview/models/subreddit.html

Sundial Growers Inc. (SNDL) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/SNDL/

Swathi, T., Kasiviswanath, N. & Rao, A. A. (2022). An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03175-2

Talamás, J. (2021). *Social media Effects on the market : Reddit Data analysis on Stocks*. https://doi.org/10.13140/RG.2.2.24180.88960

Tesla, Inc. (TSLA) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/TSLA/

The Investopedia Team. (2021). What Is a Hedge Fund ? Récupérée 22 septembre 2021, à partir de https://www.investopedia.com/terms/h/hedgefund.asp

Tokenization. (2009). Récupérée 25 septembre 2021, à partir de https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

Trading volume Definition. (s. d.). Récupérée 6 septembre 2021, à partir de https://www.nasdaq.com/glossary/t/trading-volume

Tweepy. (s. d.). Récupérée 12 septembre 2021, à partir de https://www.tweepy.org/

TweetEval. (2022). Récupérée 23 avril 2022, à partir de https://github.com/cardiffnlp/tweeteval

TWINT - Twitter Intelligence Tool. (2022). Récupérée 5 février 2022, à partir de https://github.com/twintproject/twint

Twitter API Documentation. (s. d.). Récupérée 30 août 2021, à partir de https://developer.twitter.com/en/docs/twitter-api

Unix Time Stamp - Epoch Converter. (s. d.). Récupérée 5 février 2022, à partir de https://www.unixtimestamp.com/

Vasileiou, E., Bartzou, E. & Tzanakis, P. (2021). *Explaining Gamestop Short Squeeze using Intraday Data and Google Searches* (SSRN Scholarly Paper N° ID 3805630).

Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3805630

Versloot, C. (2019). Understanding SVM and SVR for Classification and Regression. Récupérée 23 septembre 2021, à partir de https://www.machinecurve.com/index.php/2019/09/20/intuitively-understanding-svm-and-svr/

Virgin Galactic Holdings, Inc. (SPCE) Stock Price, News, Quote & History - Yahoo Finance. (s. d.). Récupérée 6 février 2022, à partir de https://finance.yahoo.com/quote/SPCE/

Wang, K., Li, Y. & Erickson, J. (1997). A New Look at the Monday Effect. *The Journal of Finance*, *52*(5), 2171-2186. https://doi.org/10.1111/j.1540-6261.1997.tb02757.x

Wei, W. (2016). Vertical specialization and increasing productive employment : Comparing impacts of conventional trade and processing trade patterns on labor market in China. In W. Wei (Éd.), *Achieving Inclusive Growth in China Through Vertical Specialization* (p. 71-138). Chandos Publishing. https://doi.org/10.1016/B978-0-08-100627-6.00004-7

What are Neural Networks ? (2021). Récupérée 24 septembre 2021, à partir de https://www.ibm.com/cloud/learn/neural-networks

What is an API ? (2017). Récupérée 30 août 2021, à partir de https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces

Xu, W. & Zheng, Y. (2016). *The Short Squeeze : The 'Invisible' Cost of Short Sales* (SSRN Scholarly Paper N° ID 2783374). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.2783374

Yang, S. Y., Mo, S. Y. K. & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, *15*(10), 1637-1656. https://doi.org/10.1080/14697688.2015.1071078

Zhu, Z. (2021). A Step-by-Step Tutorial for Conducting Sentiment Analysis. Récupérée 25 septembre 2021, à partir de https://towardsdatascience.com/a-step-by-step-tutorial-for-conducting-sentiment-analysis-9d1a054818b6

Zoltan, C. (2021). SVM and Kernel SVM. Récupérée 13 septembre 2021, à partir de https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**