Creating an NLP-based financial indicator using Reddit messages

A proposal to forecast Reddit micro-investors impact on stocks (price and volatility), create an early warning financial indicator, test it on GameStop short squeeze and measure the interaction effect between social networks using Natural Language Processing.

Guillaume Karklins-Marchay

(under the supervision of Julien Fouquau)

February 2021

Glossary

Micro-investors: new investors using free trading platforms online such as *Robinhood* holding stocks for small investments. Their decisions are a new determinant element on the market.

Bombing: massive and coordinated action lead by a sub-group of investors / micro-investors to influence, destabilize or volunteering create loss for another investor or sub-group of investors.

Machine Learning: *"Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so"* [g].

NLP (Natural Language Processing) analysis: use of machine learning models, sentiment analysis and word-embedding systems to quantify and qualify the meaning and power of text-data (messages, subreddit, tweets…).

Social Network Porosity: idea that an information shared on a social network will be published on other social networks within a certain amount of time. Porosity is also a physical formula, that could be adapted to social network information.

"Guerrilla" effect: social networks popularity is permanently evolving. Now, Reddit is the centre of attention, but this could change in the future with for example a private conversation influencing a stock. This quick and unpredictable change of a group with no leader is apparent to the one of guerrillas and should be considered to create a robust indicator.

1. The problem

Reddit users from the forum "wallstreetbets" [e] have recently massively invested on different financial stocks with the objective to either earn money or sink funds that took short positions on certain companies [8]. The price of the most famous one, Gamestop, has increased by 1700% in one month, making the fund *Melvin Capital* notably loose several billion dollars in a few days. These micro-investors mostly used *Robinhood*, a free trading application on mobile phones. Ironically, the app roots a part of orders to *Citadel Securities,* a company making use of the information for algorithmic trading… and which provided co-jointly the $2 billion bailout to *Melvin Capital.*

This event has created a huge turmoil in the financial community, politics and in the media, as it is the first-time micro-investors coordinate efficiently their actions on a stock through social networks. Such an event was unpredictable, especially with traditional forecasting and hedging methods [d].

Some financial investors now fear that such micro-investors using trading applications could lead to large financial losses for certain actors on the market, increasing at the same time the general market volatility. The regulator has also been worried by such movements, as they could lead to unsustainable long-term positions with bombing on certain stocks, unsignificant quoted value and even market manipulation [b].

In the article *"GameStop: comment une horde d'investisseurs individuels fait trembler Wall Street" ["GameStop: how a horde of individual investors rock Wall Street"] from* the French newspaper Les Echos, Aurel BGC has declared, *"How can you invest your long-term savings if valuation no longer depends on a rational business valuation model, but on discussions on social networks?"* [a].

However, early 2020, *Matthias Schnaubelt & al.* [2] have successfully demonstrated that using social networks data (Tweets) can translate to statistically and economically significant excess returns on the S&P500 with Machine Learning algorithms (specifically Random Forests on short-time windows). Especially, this return is strongly linked to temporally clustered tweets, corresponding to high-event situations. In other words, social networks can be successfully used to take financial decisions [1][4].

The objective of this thesis is to test the data produced by discussions on Reddit using NLP. It can provide precious data on micro-investors' decisions, the volume implied, their positions and ultimately create useful financial indications. The question of interactions between social networks is also central in this process as the spread of information can strongly influence the outcome on real markets.

2. The proposal

This Master's thesis will try to answer the following questions:

1- Can we predict orders done by micro-investors on the market and their impact using "wallstreetbets" messages?
2- Was the volatility peak created by Reddit's users on Gamestop (and others) forecastable?
3- Can we create a financial indicator based on Reddit data useful for investors in the future? Is this indicator reliable on the Gamestop (& others) case?
4- Can this method be generalized to other subreddits, platforms, threads and social networks to avoid the "guerrilla effect"? What is the interaction effect between the different social networks?

All the work will be performed using Python and packages available to the large-public and the data produced by Reddit forum 'wallstreetbets" [f]. A pre-grouped dataset will be used and the Reddit API.

3. Main benefits and current use of social networks for finance

Such large losses show a general lack of observation from social networks from investment companies. Traditional forecasting methods relies on historical data; however, the Gamestop event demonstrates that social networks are and should now be a part of the valuation and forecasting process.

With the recent advancements in Machine Learning and Natural Language Processing (NLP) [3] it is now possible to use social networks as a forecasting tool. Even if papers using them are increasingly popular in the forecasting research, their use remains limited in real financial applications [5][6][7].

Following the Gamestop event, Thinknum Alternative Data quickly developed a Reddit mention tracker system and Nomura / Wolfe Research a Wolfe Retail Red Alert tool [c].

Those systems have been mediatized and are showing a rushing demand from investment companies, as micro-buyers are creating a fear among traditional hedge funds and investors.

However, such tools are still basic: for example, Thinknum Alternative Data counts in live the occurrences of companies on the subreddit, meaning there is an opportunity gap to develop more advanced NLP systems [c]. Also, such systems are neither open-source or free: the objective of this thesis is to develop a general pipeline and financial indicator that could be used by everyone.

Finally, such tools are specific to networks and it is likely that other subreddits, forums and platforms could be used for bombing other stocks. This is the “Guerilla”-effect: how do we track evolutions over several platforms and adapt the analysis pipeline?

This Master Thesis will try to solve all these problems, providing tools that can be used freely by investors to assess Reddit investors’ impact.

4. Objections and responses

Several questions regarding the technical aspects and limits of this approach could be raised. This proposal aims at answering to the followings:

How large volumes of text data will be handled? And the analysis of companies?
Python can handle large volume, especially certain packages proving support for large dataset operations. Companies stocks can be identified through their complete names or stock abbreviation and natural language processing allows extracting information in messages with accuracy.

Gamestop is the first (known) large impact of reddit on the market; therefore, it is complicated to create a financial indicator based on the current data.
Partly true. The objective is first to find patterns in Reddit data, see the link with the market and if the timing of appearance allows a certain forecasting, looking for several factors or statistics with strong early signal effects. In the Gamestop case, if we look at the complete story, traces of discussions on the stock were present late 2020 on Reddit.

Does Reddit influence the market or does the market influence discussions on Reddit?
This question can be responded with deep and detailed analysis of the timing difference between stocks and messages. Plus, social networks tend to create chain reactions: if the price of a stock is increased, Reddit’s users are likely to share their observations, leading to additional investments on the stock and increasing further the price with a delay. (Most Reddit users being unfamiliar or novice with exchanges -the forum is named “bet”-, associated with additional noise and volatility).

Emojis have been used a lot in the Gamestop case. Can we handle them?
Yes. Emojis have been used in Reddit messages: for example, snakes represent the hedge funds. This is a precious source of information for sentiment analysis.

By not gathering data on all social media, the model could miss important information.
Partly yes, the main impact being the delay between the publication on Reddit and its appearance on non-reddit media. There is a porosity between social media: for example, the GameStop bombing has been extremely shared on Twitter and financial information published on Twitter can also be shared on Reddit. The difference is the time of reactivity, as if the information is recovered directly at the source, it is made earlier than on other platforms. Similarly, some minor information published on Reddit are not published somewhere else.

Is there a limit of forecasting ability?
Yes. For example, we could think about a private Telegram group of micro-investors that could bomb a specific stock. Because the data is not publicly available, it is almost impossible to include information shared on a private group. Therefore, asymmetric information will still be persistent, but the impact should be restricted with potential porosity on forums and a restricted number of investors present on such groups.

Robinhood is already aware of micro-investors impact, as they are the gateway for micro-investors.
Yes. However, the collected data is used to obtain information before others for High-Frequency Trading and is provided at the buy/sell moment. Reddit is giving users the possibility to discuss, express many shades of sentiments in early discussions. Therefore, Reddit data allows early forecast while Robinhood data allows late forecast.

5. References

Academic

[1] Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, Werner Antweiler and Murray Z. Frank, The Journal of Finance, Vol. 59, No. 3 (Jun., 2004), pp. 1259-1294

[2] Separating the signal from the noise – Financial machine learning for Twitter, Matthias Schnaubelt & al., Journal of Economic Dynamics & Control, March 2020

[3] The Role of Big Data in Investing, Goldman Sachs Asset Management, GSAM PERSPECTIVES | BIG DATA EDITION, JULY 2016

[4] Exploiting social media with higher-order Factorization Machines: statistical arbitrage on high-frequency data of the S&P 500, Julian Knoll, Johannes Stübinger & Michael Grottke, Quantitative Finance, November 2018

[5] Information, Trading, and Volatility: Evidence from Firm-Specific News, Jacob Boudoukh, Ronen Feldman, Shimon Kogan, Matthew Richardson, The Review of Financial Studies, Volume 32, Issue 3, March 2019, Pages 992–1033

[6] PREDICTING RETURNS WITH TEXT DATA, Zheng Tracy Ke, Bryan T. Kelly and Dacheng Xiu, NATIONAL BUREAU OF ECONOMIC RESEARCH, August 2019

[7] News or Noise? Using Twitter to Identify and Understand Company-specific News Flow, TIMM O. SPRENGER, PHILIPP G. SANDNER, ANDRANIK TUMASJAN AND ISABELL M. WELPE, Journal of Business Finance & Accounting, October 2014

[8] Counter-Hegemonic Finance: The Gamestop Short Squeeze, SSRN, Usman W. Choha, 4 Feb 2021

Newspapers or online resources

[a] https://www.lesechos.fr/finance-marches/marches-financiers/gamestop-comment-une-horde-dinvestisseurs-individuels-fait-trembler-wall-street-1286065

[b] https://www.nicematin.com/amp/economie/5-questions-pour-tout-comprendre-a-laffaire-gamestop-qui-affole-la-bourse-de-wall-street-639560

[c] https://edition.cnn.com/2021/02/03/investing/wall-street-reddit-hedge-funds/index.html

[d] https://www.nbcnews.com/business/business-news/gamestop-reddit-explainer-what-s-happening-stock-market-n1255922

[e] https://en.wikipedia.org/wiki/GameStop_short_squeeze

[e] https://www.reddit.com/r/wallstreetbets/

[f] https://www.reddit.com/r/DataHoarder/comments/l7oxw9/creating_a_wallstreetbets_archive/

[g] https://en.wikipedia.org/wiki/Machine_learning