WallStreetBets: Evaluating the ability of public sentiment to influence financial markets

Amin Debabeche
EPFL
Lausanne, Switzerland
amin.debabeche@epfl.ch

Kamran Ali Nejad-Sattary
EPFL
Lausanne, Switzerland
kamran.nejad-sattary@epfl.ch

***Abstract*—It has long been seen that the valuation of an asset on the stock market is fluctuating following classical news related to the company itself. With the popularization of cheaper broker platforms and accessibility of the financial market to so called *tiny-investors*, financial related forums composed of many of these individuals arise. This research focuses in predicting stock returns of the S&P500 ETF with the help of sentiment analysis on a typical forum: *WallStreetBets*. This article is divided into 2 main sections. As a first step, data engineering is performed, and then followed by the modelling procedure. In this paper, a low accuracy results from predicting with such relationship. This may be a first step in proving that the index is not prone to fluctuation of such forum or the event to be so rare, meaning that it is not possible to forecast. Nevertheless, the methodology used in this article can be used to investigate such behaviour but on a intraday basis to find trends and predict the price of the stock in real-time.**

***Index Terms*—Prediction, Stock Price, Forum Influence, Sentiment Analysis**

I. Introduction

Technical indicators, heuristic or pattern-based signals produced by the price, volume or open interest of a security, remain one of the key stumbling blocks in asset trading. Although the number of publicly reported indicators has reached more than seventy, this still represents only a small portion of the totally possible data analysis achievable[1]. The lack of synthetically accurate time series models hinders access to potentially fruitful indicators to predict future price movements. Tackling the challenge of market projection with machine learning driven approaches is particularly timely as generative models in deep neural network for finance are coming of age[2]. These generative models enrich the toolbox of financial analysis by suggesting potentially promising indicators that lie outside of known scaffolds.

There are two salient challenges in predicting stock price movement and designing accurate tools. First, the market shall be efficient and the degree to which market prices reflect available relevant information would be perfect according to the Efficient Markets Hypothesis[3]. In this situation, an asset price incorporates all information, there are no under or over -valued securities available to be vanquished by the market. However the flaw in the Efficient Markets Hypothesis (EMH) is that not all investors perceive all available information in precisely the same manner. One investor evaluates a stock on the basis of its growth potential while another looks for undervalued opportunities, the latter will assess the stock differently to the stock's fair market value[4]. Second, markets are exposed to group manipulation abuse, consisting of deliberate attempts to interfere with the free and fair operation of an asset. As seen in section IV-C a blatant case of such exploitation is for a group of investors to coordinate and create a misleading appearances with respect to the price of GameStop (GME) commodity. The video game retailer has underwent a market squeeze, as individual investors who actively engage together on a Reddit Forum known as *WallStreetBets* (WSB) started to give significant attention to the stock with simultaneous buying the company's stock. This single short-squeeze incident built up into a potentially broader systemic risk, casting doubt on market stability and integrity[5].

Therefore, given the potential for large institutional or celebrity-type announcements to shock the markets, coupled with more recent online small-investor coordinated trading-efforts propelling one of the

largest short squeezes in history, the question of just how much investors of different scales may sway the markets, remains undetermined. This study sets out to examine this very phenomenon, utilising as case study subject the WallStreetBets subreddit. Is it possible, through large-scale coordinated efforts for small-volume investors, acting essentially as a distributed decision-making trading entity, to move the markets?

In this work, a particular focus on the question of predicting the Standard and Poor's 500, a stock market index tracking the performance of the 500 largest companies listed on the US stock exchanges. It can be seen that using an existing sentiment analysis on each individual comment using VADER Sentiment Analysis[6]. These Natural Language Process (NLP) frameworks show the highest ability to analyse the sentiment across social media comments and posts, being perfectly suited to the Reddit comments extracted. Some of the most famous indexes S&P500 is chosen to be sufficiently close to the actual WSB forum overall interests and the approximation of the daily forum resentment on market performance to be mainly galvanized by the top large companies listed on stock exchanges in the United States. A Temporal Convolutional Neural Network based model do not outperforms benchmark models, which consists of a simple historical probability, while being highly post independent and not requiring ultra-fine data processing. The model reaches 66% top-1 accuracy (50% top-2 accuracy) on a common benchmark data set. Importantly, it does not make use of any kind of handcrafted rules, regularly seen and needed in such dynamic forum. It can accurately predict subtle and selective change in the populating users of WSB subReddit, as long as the indexes are in a nearly steady states in terms of their composition. Nevertheless, one would postulate that the forum interest is in phase with indexes' composition. The score predicted by the model has an ROC - AUC of 0.55 in terms of multi-classifying whether returns are correctly predicted, for a Long Short-Term Memory Neural Network architecture.

II. Data

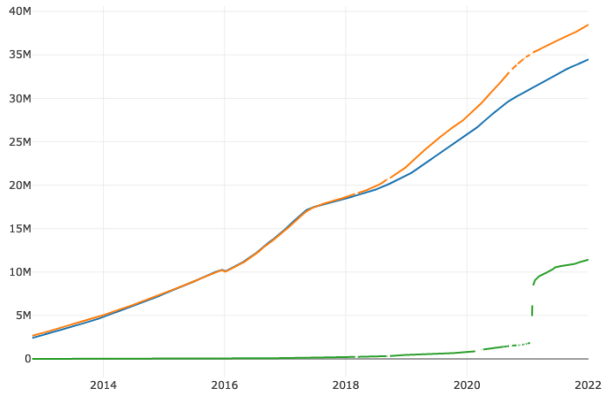
A. WallStreetBets Forum

WallStreetBets also known as WSB, is a Reddit forum, created in January 2012, on finance and the stock market, including topics such as trading stocks and options on the stock market. In January 2022, it had over 11.4 mio. subscribers generating roughly 7'155 daily comments and a total of more than 4.6 mio[7]. total comments. It has to be mentioned that Reddit itself is on the top ten most daily visited website. The subReddit is known for its aggressive trading strategies, which primarily revolve around highly speculative, leveraged options trading, to bet on stock that show popularity within the community. Mainly composed of young traders, with none or little investment and risk practices. With the venue of lowered commission brokers together with trading applications, it has contributed to the growth of such gambling trading strategy. The subReddit forum has gained increasingly exposure following the previously discussing GameStop market squeeze, with an increase of 8.9 mio. members

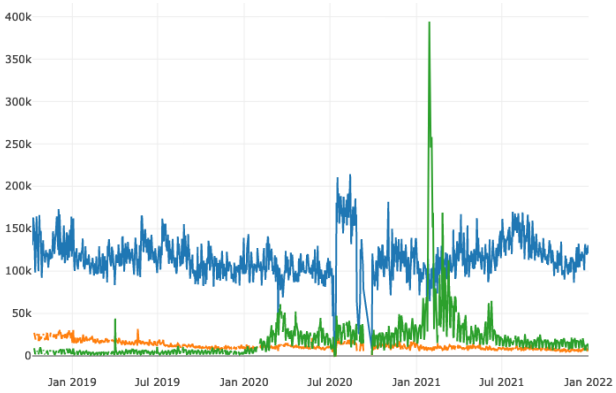
According to the amount of daily comments and subscribers, these portray a fair amount trust in the influence of the subReddit on market stocks. Reviewing the rules of the forum and thanks to extensive admins daily management the following subject or behaviors are quickly deleted from the forum:

- Market manipulation, posts containing misleading information and is made for the purpose of manipulating the market for a security is prohibited. Particularly since January 2021, GameStop short-squeeze;
- Focus on standard asset, posts on
  - non-reporting penny asset;
  - microcap, asset with less than .5 billion Market Cap;
  - Over-the-counter stocks;
  - Low volume options;
  - Cryptocurrencies;
  - "any other worthless securities".

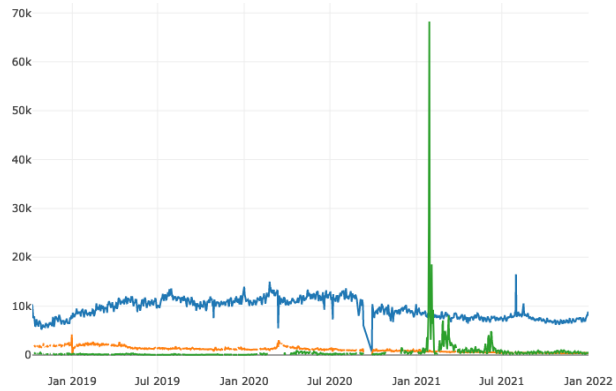
  are forbidden, due to their susceptibility to scams, pump dump or criminal schemes;
- Political discussion, any kind of advertisement, fundraising or begging;



(a)



(b)



(c)

Fig. 1. Ranked top 1-r/funny, top 2-r/AskReddit and top 4-wallstreetbets Reddit forums statistics. In blue: funny, orange: AskReddit and green: wallstreetbets. (a) total subscribers to the subReddit. (b) Daily comments rate. (c) Daily posts rate.[7]

- Users showing very poor knowledge or ill-informed opinion are quickly banned;
- Micro-investors posts are forbidden. For example, only posts showing trading of minimum $10,000 of options or $25,000 of shares are allowed. Also gain/loss posts must show either gains or losses of more than $2,500 for options or $10,000 for shares.

From this set of rules, there remains an underlying assumption that there is a significant chance for a momentum to be created from a particular post which creates an effect on an individual ticker and by extension an index. Note that online users visiting Reddit without being registered do not add up to the total number of the subReddit members. Again due to accessible market brokers and the large volume of *tiny* investors, there could be a relation between posts sentiments and the market movement.

Nonetheless, the question concerning the community diversity and if this diversity reflects the portion of low-investors section, is to be settled and is considered to be sufficiently close to reality for the sake of this research.

B. Raw Data

The raw data has been downloaded from Kaggle, a public repository of community published data and code. Two sets have been investigated during this present work, both were derived using Pushshift Multithread API Wrapper (PMAW) a minimalist wrapper for the Pushshift API which uses multithreading to retrieve Reddit comments and submissions. With the mining work of Kaggle users, the first data set, *wsb_comment˙raw.csv* (referred as Big data set), is a file of 11.91 GB containing 33,396,559 unclean unique comments up until February 16^{th}, 2021. The second data set, *wsbData.json* (referred as Small data set), a file of 1.12 GB containing approximately 2.5 mio. unclean unique comments up until December 22^{nd}, 2018.

Both data sets are similar in content type and the overall code was created using the Big data set, although the Small data set is used to train the model. One could train again the final model to take into account extreme event as the GameStop squeeze, since the Small data set ends before it happened.

In this particular project, the authors avoided data requests, by using an existing these existing database. It could be interesting to scrap the up-to-date data to gather more relevant data and then intensify the training step. This could ensure better wrangling, as both Big and Small data sets are bulk exports with no information on the data gathering or threads extraction information. An eye validation consisting of on cross-checking manually threads in Reddit and match some comments ID, authors and contents. One could question the confidence in the whole data set sanity.

From both data sets, the columns of primary interest in the research are:

- author;
- body;
- created_utc;
- link_id;
- score;
- subreddit;
- is_submitter.

C. Preprocessing

C.1 Data Cleaning

It has not been mentioned in the previous section, but originally there were more columns to the data sets with no purpose in the modelling. The decision was made to drop them and to work only the previously defined column of interest. The dataframe was also cleaned by removing deleted or updated posts to avoid outliers. As it remains Reddit comments, some atypical steps are performed, such as deleting any reference to external URLs or user tagging. The usual data preprocessing in Natural Language Processing, lemmatizing, stemming, removing stopwords, were undertaken along with textual tagging using a refactoring text library[8].
At this step, an automatic checking procedure is performed but is not discussed in the present paper.

Algorithm 1: Prepocessing

```
Data: Raw Data
Result: Cleaned comments
if comment has not been 'deleted' then
    drop unnecessary columns;
    transform time to timestamp object;
    if Text exist then
        lower-casing letters;
        tokenize;
        remove url;
        remove user tag (i.e @amin);
        for tokenized text do
            lemmatize;
            stem words;
            remove punctuation;
            remove stopwords;
            compute part-of-speech (POS)
             tag;
            detailed POS tag;
            syntactic dependency;
            word shape;
        end
        untokenize;
    else
        drop row;
    end
else
    drop comment;
end
```

C.2 Sentiment Analysis

Deep learning has been widely utilised and advanced in various field such as computer vision, audio-visual recognition or more interestingly in natural language processing, and its overwhelming success as a data processing approach towards speech recognition or sentiment analysis which make it a perfect starting point in this research[2]. Multiple attempts to implement algorithmic trading by analyzing historical data current price movements and extraction information from social media discussion simultaneously, have succeeded in showing deep learning algorithms can predict stock values more accurately. These modelled predictions could then be used for fast trading decisions or high frequency trading. Using platform users' emotions to develop a strategy with lack of emotions, decisions and predictions deep learning models deliver are more objective and data-driven[9;10;11].

It has been decided to use Generative Pre-trained Transformers (GPT) to perform the sentiment analysis on each individual comment. Indeed, the existing solution are well developed[12]. The first version of

TABLE I
BIG DATA SET SAMPLE

author	body	created_utc	link_id	score	subreddit	is_submitter
LazyMeal	We're retarded and claim to be often. If...	1585123910	t3_fom9g6	1	wallstreetbets	False
math_salts	Yes	1585123909	t3_fod66b	1	wallstreetbets	False
Legendary_Squirrel	markets been open for 13 min...	1585123905	t3_fod66b	1	wallstreetbets	False
WSBMORONICT...	Spy can fuck around all it wants...	1585123901	t3_fod66b	1	wallstreetbets	False
[*deleted*]	[removed]	1585123897	t3_fom0hg	1	wallstreetbets	False
madamlazonga	you lost me at "bulls fucked"	1585123896	t3_fod66b	1	wallstreetbets	False

the model only use VADER sentiment Analysis, but multiple trials together with FLAIR and TextBLOB sentiment analysis framework were investigated[13]. The three rule-based sentiment analysis tools have been designed to be specifically attuned to sentiments expressed in social media. The main feature that have to be discussed is the sentiment lexicon, a list of features derived from text which are generally classified detaining a subjective ton and their development are laborious processes. A compounded value using the three analysis score exhibited good results but further investigation are to be done.

Please refer to the work of Panchbhai and al. for further information and comparison between these generative pre-trained transformers[14]. The present authors briefly review only the most popular tool for sentiment analysis: *VADER* (Valence Aware Dictionary for Sentiment Reasoning)[6].

Previous sentiment analysis existed, nevertheless the inherent nature of social media content poses serious challenges for these previously developed model in practical applications. In opposition, VADER, is a simple rule-based model for general sentiment analysis. It has been created using a combination of quantitative and qualitative methods. The first qualitative feature is the construction, and empirically validated, gold-standard list of lexical features associated to their sentiment intensity measures. This list is attuned to sentiment in microblog-like contexts, such as Facebook post or Instagram photo comments. It is then combined with consideration for general rules embodying syntactical and grammatical conventions for expressing and emphasizing sentiment intensity. Its effectiveness compared to multiple typical state-of-practice benchmarks including SentiWordNet, and machine learning oriented techniques relying on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms, and so on. VADER outperforms any of these benchmarks in general contexts. It even shows an accuracy of 0.96 when assessing the sentiment of tweets, transcending individual human raters.

VADER manifests sufficiently accurate term frequency–inverse document frequency (TF-IDF), a numerical statistic intending to reflect the importance of a word to a collection. For a pre-trained model to be safe enough to be used, the frequency apparition has to be considered, more precisely in a model such including a weighting factor in user behaviour toward a ticker[15]. The framework return four different score; positive, neutral, negative and compound, this latter score is computed by summing the first three valence scores of each word and normalized to be between -1 (most extreme negative) and 1 (most extreme positive). Only the compound value is used in this research as it is the most useful metric for a single uni-dimensional measure of sentiment.

$$compound = \frac{x}{\sqrt{x^2 + \alpha}} \tag{1}$$

where x = sum of valence scores and alpha = normalization constant (default value is 15)

C.3 Market Data

One single market data prices provider is used; Yahoo Finance. The prices are pulled in the date range of the concerning data set. Yahoo Finance is considered as trustful data resource supplier.

The return used in the model is calculated as follow:

$$return = \frac{\text{Opening price}_{t+1} - \text{Opening price}_t}{\text{Opening price}_t} \tag{2}$$

TABLE II
INDEXES INFORMATION

Symbol	Name	Volume
GSPC	S&P 500	2.814B

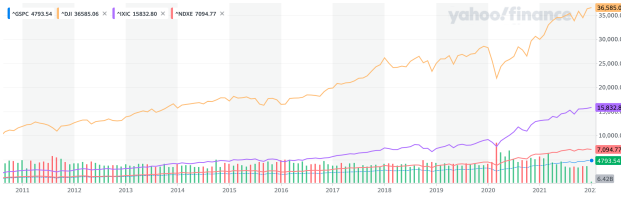


Fig. 2. Daily prices of indices from table II[16]. The volume shown is from SPY index.

D. Feature Engineering

In order to hopefully obtain results from the models to increase in accuracy, some feature engineering have been implemented on the data sets. This process of using expert domain knowledge to extract meaningful features from the data set is a standard in this field research. From domain experts and a review on quantitative risk management applied to financial market[17;18;19] to gain knowledge specific to the problem being solved, a few transformations and feature engineering are discussed. Note that not all extracted features are exploited, when either results indicate insufficient improvement or there is a lack in specific field knowledge.

D.1 Compound Score

Reddit has implemented a very simple tool: score. Every user comment receives a score that corresponds to the number of up-vote minus the number of down-vote, it received. This clearly shows the popularity of a comment or a post and can be used along with the sentiment analysis compound resulting of the sentiment analysis. The logic is that a comment with a high score reflects the sentiment of a group of individuals. On the other hand a comment with a highly negative score implies that the comment's content has not been approved by the community. The possible paths to use such tool are, firstly, to use only comments with a score higher than 30. Those are considered to be commonly accepted and to reflect the forum sentiment at some point in time. In the second place and more importantly, a composition of the compound with the score is formulated as follows, where the $daily\_score$ serves as a normalization factor, and corresponds to the absolute total score for a particular day. Note that implicitly the division of the comment score by the total daily score, represents the weight the community attributes to a particular post.

$$sentiment\_sc = \frac{comment\_sc^{var}}{daily\_sc} * sentiment \quad (3)$$

Where sc means score and var has been set by default to 1.1, but could be a parameter in the model as suggested in section .

D.2 Flair Threads

The comments can be filtered further by avoiding out of subject posts. In this particular forum there are posts directly created and managed by admins; *Daily Discussions*, and *Weekly Discussions*. From an experimental perspective, they are posts with content focused on analysis and investment decisions, usually the discussion are more controlled and most orderly behaviour this forum contains is on display. Notice that there may exist other threads with high quality investment conversations.

D.3 Comments correlation

Threads may albeit difficulty be separated from its parent post, because of the intrinsic syntax used in such corpus. Readers may have experienced it on diverse social media platform, but there is an implicit link between the parent posts and its child comments.

> Parent: Yeah, the lately abnormal returns of TSLA makes it an under-valued asset prone to correction, glad to earn some easy money.

> Child: indeed, it is certainly due to very poor acquisitions and awful CEO public management. I hate Elon.

Here in the child comment, one knows the post is implicitly talking about Tesla stock and its possibly under-valuation. However the subject and object are not explicitly discussed in the child comment. A human sentiment analysis in the parent would

give a positive value and the parent user would think the return to be positive also, while it is the opposite for the child user. The link between both comments are inversely proportional in sentiment. This is called the Oral Reaction Inversion (ORI) defined as the reaction of a person reacting with negation to assert its interlocutor.

To overcome the separation issue between parent-child comments, an easy instrument is to classify the parent in bins of positive and negative returns. Standard words as: buy, call, forward, long, up, grow, rise, green, hold, carry, bull are attributed to a market price increase, known as bullish. While on the other hand words such as: put, short, bear, sell, red are attributed to a market price decrease, known as bearish. This permits to create a new parameter that reflects more precisely the general discussion topic of an unique post. Concerning the ORI, one approximates for this behaviour to be present in small proportion enough to be negligible. From an empirical observation, user usually distinctly declare their intention and thinking, any type of irony or sarcasm like syntax are avoided in such threads. In addition to this, the title classification in bullish or bearish bins are considered to help in overcoming the paradox. This is also let to the model to reproduce close enough the general resentment. Because if many users are seeing Tesla CEO, for instance, as downgrading the value of the stock by its behaviour, it reflects the stock idiosyncratic risk, even if its behaviour can in reality induce an increase in the return.

D.4 Window Lag

Considering the fine filtering of the comments discussed above, there is now the necessity for intraday comments to be grouped by date time, as the model uses day intervals to predict return prices. To calculate a daily compound_score one could simply directly sum up sentiment scores for each post, although this implicitly ignores the scores, which is here considered the level of agreement of the community towards such post. To overcome this issue, the judiciously normalized compound scores discussed previously in (3) is performed are instead summed up as follows:

$$\text{daily\_SS} = \text{SS.sum()} \tag{4}$$

Where SS represents the sentiment score (positive, neutral, negative, or compound) discussed in (3), with parameter var set to 1.1, but this remains a parameter in the model as suggested in section .

This may seem simplistic, but the sum over every daily sentiment_score does include the dimension of the number of daily posts. The exponentiation allows for consideration that gives a sense of triggering to the parameter, i.e. the occurence of a rare (e.g. GME squeeze) event which should accompany an unusual number of daily comments.

For each time period t, windowing of n periods is then performed, such that for the prediction of the return from today's to tomorrow's opening price y_t, we have features X_{t-1} until X_{t_n} too. This allows the model to then capture the trend of previous returns for the SP500, as well as the previous days' sentiment scores. The number of day could also be a parameter in the model as suggested in section .

D.5 Fast Fourier Transform

Aberrant values, in particular measurement errors, present high effect on estimating the genetic parameters related to a selected study. To illustrate this, one can take as an example the comment of a user that has a compound value of one but has been in reality over estimated due to the benchmarks used in the pre-trained generative model. Such aberrant values cause observations response to selection pressure to differ considerably from an accurate prediction. One mechanism to transform and denoise a timeserie with meaningful data points to be used in the upcoming model, is to use a Fast Fourier Transform (FFT). This algorithm computes the Discrete Fourier Transform (DFT) of the whole sequence, and then following a dimensionality reduction, the Inverse Discrete Fourier Transform (IDFT) is computed. The reduction in the frequency domain consist of decomposing sequences of value into components of different frequencies. An FFT can cause long computation for large data set, such as a Big data set[20], but thanks to factorization using sparse matrix in more recent algorithms, it was possible to efficiently utilised it in research. Consider avoiding this feature engineering with enormously large dataset with low noise, the added contribution would be minimalistic.

Multiple transformations are thus available in the reduction, with different components sizes, the selection of interesting dimension has been investigated graphically and the Fourier-100 has shown the greatest engineering feature for this notable timeserie. This represent better swings and changes in the overall sentiment over time, specially in the case of predicting indices prices, as there are considered as stable asset.

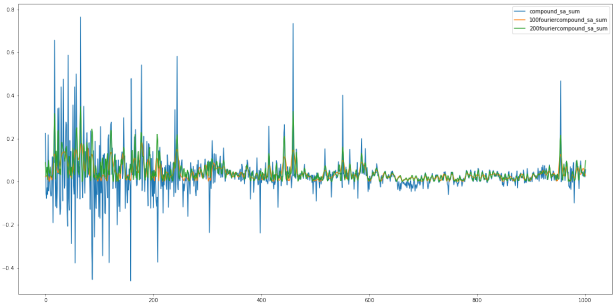


Fig. 3. Fourier transform of the sentiment analyse compound.

D.6 Return Normalisation

Despite its evident application in machine learning as part of most of the data preparation, the present authors are willing to mention the normalization as an important feature engineering step, particularly in speech emotion recognition[21;22]. This aids the training of the model weights, without distorting differences in the ranges of values or distribution. This normalisation can also by applied to the Fast Fourier Transform created, although as the Fourier Transform is only applied to the sentiment analysis scores which are already normalized as byproduct of the daily-weighted-scores, this is not necessary. Instead only the returns are normalized, utilising a minmax algorithm. This standardization step is even considered while building various deep neural networks. Batch normalisation units into deep feedforward neural networks have resulted in enhancement of modelling in several public datasets, for example MNSIT or CIFAR-100[23].

III. Model

A. Baseline

A first crucial step before diving straight into the modelling remains establishing a baseline to which we can compare our performance. While more complex baselines may be constructed, a simplistic baseline utilising simply the proportion of both classes, the daily positive and negative returns, was utilised. Indeed, should one be able to beat our model by simply randomizing across the classes with their historical probability, the model would serve no purpose. Overall in the dataset (post-windowing, splitting into training, validation and testing sets, and removing gaps between the different sets of the window size to avoid statistical biasing), there remains 314 days with negative returns, and 414 days with positive returns, which corresponds to a proportion of 43.13% and 56.87% overall. This class distribution does not however prevail across differing time periods.

The data is split into three separate datasets with a 0.7, 0.2 and 0.1 share respectively, which are named the training, the validation and the testing set. Note that due to the windowing operation, to avoid statistical biases, a gap is preserved of n periods between each of these sets, where n is the number of window channels the data contains. These three sets serve to train, and illustrate performance of the model on out-of-sample data - which is where the testing set draws it's use, while the validation set is used to select best-performing parameters which preserve model-generalizability and avoid overfitting to the training data. The training set obtains class proportions of 45.5% and 55.5%. Should one randomize (e.g. with a biased coin-flip) the two classes with such probabilities, and assuming this proportion of upward and downward returns was to be replicated in the future, then one would obtain an accuracy of 50% over large simulations. This is what we aim to beat.

For this purpose, two models are curated, each with differing defining characteristics; the Temporal Convolutional (TCN) Neural Network, and the Long Short-Term Memory (LSTM) Neural Network. While the latter is considered the more traditional approach towards time series modelling, according to[24] this way of thinking is perhaps antiquated, justifying the contrast of both differing methodologies in this study. Not only do TCN's present potential for performance increases, they also may bring computational efficiency and an easier model to train due to a less memory bandwith

intensive architecture.

For both these models, a standard grid search of a variety of hyperparameters is first devised, and randomized over. Indeed as suggested in[25], a solution to avoid the exponential explosion of hyperparameters is to randomize over their combinations, as there is a high chance with even few iterations to find a minimally distant solution to the optimal one. These hyperparameters naturally differ between both models, and their performance during training was both recorded visually via a confusion matrix plot as well as a Receiver Operating Characteristics figure, and numerically by repeatedly updating a python list saved as a pickle file.

B. The Temporal Convolutional Neural Network

B.1 Hyperparameter tuning

Given the advantage of a speedy computational training of the Temporal Convolutional Network, a creation of a wider grid of parameters to iterate over was made possible. By pairing Convolutional layers which iteratively summarize different windows of the data with maxpooling, an attempt at understanding the different trends of the data was made. Potential of model overfitting was tackled by introducing the possibility of dropout, and paired with batch normalization to stabilise the training procedure. Differing optimizers, such as the traditional Stochastic Gradient Descent and the Adam optimizer, along with different activation functions such as the Rectified Linear Unit and Hyperbolic Tangent functions were put to use, before considering different kernel initializer distributions which may pair better or worse with these functions, namely the He Normal and Glorot Uniform distributions. Lastly, the number of neurons, and the number of stacked convolutional layers were tampered with.

B.2 Results

Selecting the hyperparameters which maximise the area under the curve (AUC) of the Receiver Operating Characteristic on the validation set, leads to a model which solely predicts the majority class unfortunately. No model was able to provide a better generalizability and classify the validation set better in terms of the AUC score than simply stating that everyday would provide a positive return. If instead, one were to select the model which maximises the accuracy across both classes in terms of true positives, one obtains a validation accuracy for the negative and positive returns of 50% and 66% respectively, and a test accuracy of 50% and 59% respectively, as figure 4 and 5. While this is perhaps better than flipping a weighted coin, this does not significantly differ from the flipped coin, and given the small sample size, does not provide for much conviction in the final model.

There were however an anomaly case where test accuracies of 100% and 59% were recorded, as figure 7 portrays, however the accompanying validation accuracies were not higher than in figure 4. It would be overfitting to the test set to simply observe this to select which hyperparameters performed best.

During training, it was observable that the loss was particularly difficult to reduce, perhaps as separability between the classes was not simple for the model to achieve. Further, in some cases, the combination of different kernel initializers and activations, would perhaps lead to local minima which completely disregarded the negative returns class, in favour of only predicting positive returns.

Despite extensive attempts at modifying the many parameters included in the model, and expanding upon the feature engineering as the archived notebooks display, these results persisted. Perhaps after all, the sentiment on such a forum is not sufficiently impactful to sway the market, or the feature engineering simply lacks. Attempts of including logarithmic scales for the normalization of the magnitude of score scales, of sentiment analysis over the entirety of the WallStreetBets subReddit, rather than merely the highly moderated discussion threads, and on entirely different corpora including different time periods of data, were made too.

C. The Long Short-Term Memory Neural Network

C.1 Hyperparameter tuning

Unlike the Temporal Convolutional Neural Network, the Long Short-Term Memory network was particularly slow in its training. Epochs remained consistently with a Graphics Processing Unit, multiple times slower than with the Temporal Convolutional Network. Further, the stateful nature of the created model, would only be able to predict

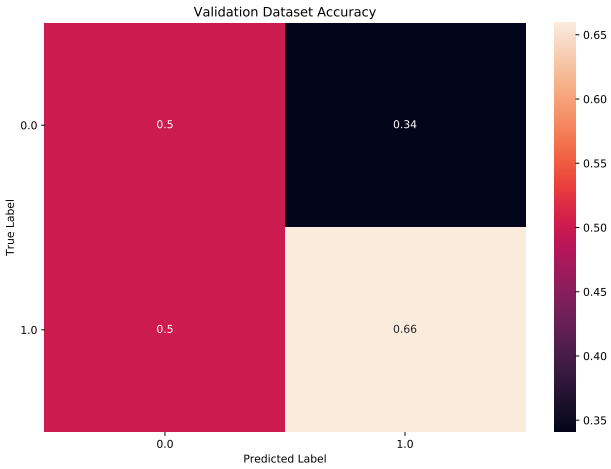


Fig. 4. Best True Positive Validation Accuracies.

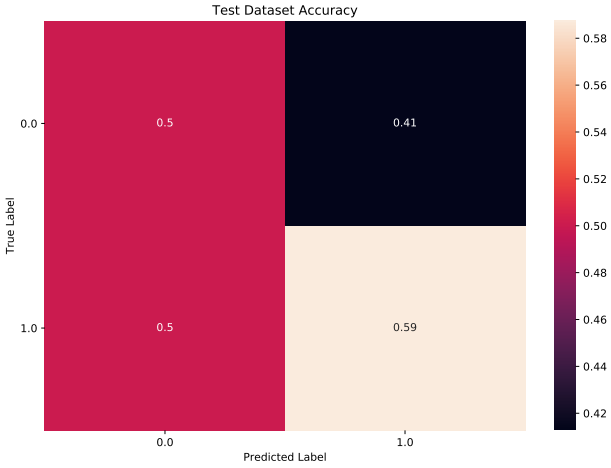


Fig. 5. Best True Positive Validation Accuracies.

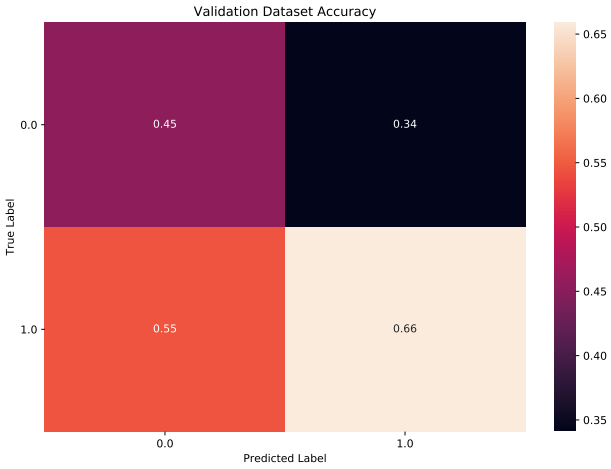


Fig. 6. Anomaly Validation Accuracies.

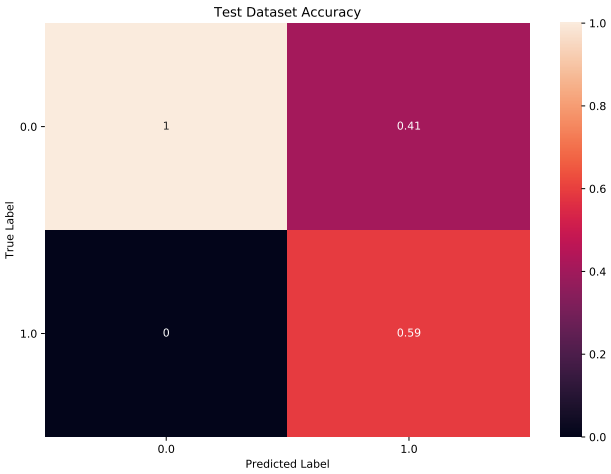


Fig. 7. Anomaly Test Accuracies.

full-size batches, which led to the decision to not utilise overly-large batch-sizes. Indeed, should one used 128 size batches, one would not be able to predict the small out-of-sample dataset without duplicating data points. Instead cutting off the last few observations such that the batch size was full, was deemed preferential.

This time, the main hyperparameters explored contained the number of stacked LSTM layers, the number of neurons in each of these layers, the probability or even presence of dropout between each of these layers, as well as the number of neurons used in the Dense Layer to make sense of the LSTM layers' output before finally outputting a prediction. The same optimizers as in the TCN were compared, along with the same activation functions.

C.2 Results

The training of the LSTM remained substantially more troublesome than that of the TCN. The model would consistently converge to the solution of predicting only the positive class. Perhaps indeed this makes sense but remains a troublesome result and nearly entirely invalidates the necessity of a model. The Receiver Operating Characteristic Plots of both the highest area under the curve validation and test sets are portrayed in figures 8 and 9. Which emphasize the lack of generalizibility of such a solution and ineffectiveness of the resulting LSTM model.

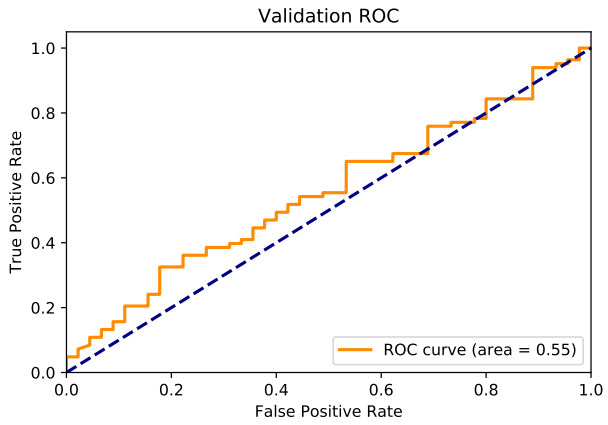


Fig. 8. LSTM ROC Validation Set.

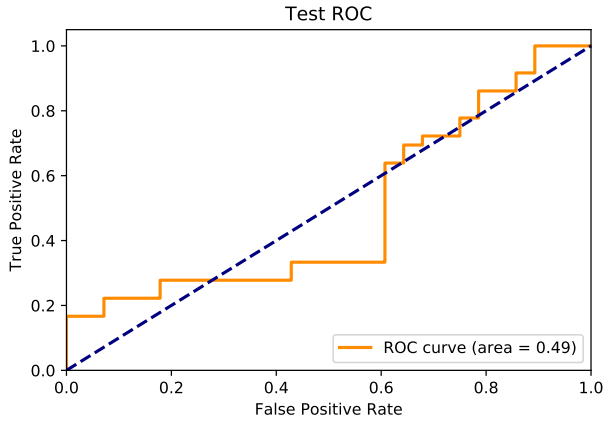


Fig. 9. LSTM ROC Test Set.

IV. Miscellanea and Concerns

A. Future Work

This research can be extended by adding more data and check the prediction accuracy. Appending the following data would maybe show different results:

- Modelling with more indices such as Dow Jones Industrial Average or Nasdaq Composite;
- Fetching up-to-date posts from Reddit.

Note that together with this supplementary data, removing multidimensionality further in the prepossessing step is to be further investigated. Using Principal Component Analysis (PCA), a statistical procedure employing orthogonal transformation to convert a set of correlated variables to a set of uncorrelated variables.

More modelling approaches remain to be explored,

- Recursive Neural Network (RNN): This type of NNs are mainly used when context is important — when decisions from past iterations or samples can influence current ones. This can thus be of interest in this present work, as every posts have internal influence to its parent post.
- Generative adversarial networks (GAN): Generator tries to generate some data, and the discriminator, receiving sample data, tries to tell generated data from samples — the model could be used to create new plausible samples, useful in training further tickers.

In the following subsections, the authors also develop some research path that should be further studied, particularly concerning the subsection IV-D.

B. Suggestions

B.1 Corpus Analysis

Further normalisation of the subReddit chat texts using social media corpora, is possible. Unique post sentences using a part-of-speech tagger are defined, from it unrecognised words could be extracted. These are considered as revealing peculiarities of chat language of such a platform. Indeed, every post or even blog environment creates a new lexicon and dynamics of speech. One does not address similarly his friends on Facebook or his followers on Instagram. From a list of such words, one can further define a precise sentiment analysis[21].

B.2 BERT

In this research pre-trained sentiment analysis models are investigated and used, however it is possible to construct a model using PyTorch BERT model. Thus its architecture as a multi-class classification can be further adjusted. Fine-tuning a model for this single project, leveraging BERT's large-scale language knowledge, could improve the score_compound resulting from the sentiment analysis.

B.3 LDA

The idea behind Linear Discriminant Analysis is to reduce the number of dimensions to a single

axis. It tries to minimize the variance and maximize the distances between classes, calculating the global mean from both classes and computes the mean vector. Along with the calculation of the covariance matrix, with it it can compute a scatter matrix to create discriminant functions.

C. Similar Applications

Similar research projects have found application of deep neuronal models in the prediction of various indicators based on Sentiment Analysis from social media. One of the most representative examples, is the use of similar methodology to transcribe the sentiment of famous people and relate their content to one asset movement[26]. One is concerned by a weight-influenced prediction, on posts that generate financial market movement. The sentiment of the whole population is not sought but the analysis is done on the influence from posts of famous celebrities or structures to particular tickers. For example, America National Bank or Financial Times tweeting about the flaws in the Tesla Free-cash-flow could lead to a market movement. In other words, classical investment news would be done in a quantitative aspect. Such a model would be dynamically changing with a continuous back-testing machine learning algorithm. To better represent such approach, it is typically to account only for a ticker exposed to high news fluctuation, such as Apple or Tesla[27]. Several's actors would be defined for each ticker, as one experimentally perceives recurrent influences in the ticker valuation.

Predicting one future stock trend with sentiment analysis using classic non quantifiable data such as financial news articles about a company has shown good results before. The only assumption made is that news articles have impact on stock market or more precisely investors. Results with prediction model having 80% accuracy has be been found[28;29].

D. Intraday

The hypothesis that a similar research instead performed on intraday data is discussed in this section. The authors thinks that the poor results come from the stability of stock indices as SPY or DOW. The true valuation of a stock is quickly recovered and tiny-investors do not influence the daily return of an index. This latter being composed of many different assets and for the most usual indices of hundreds of different ticker, this means that the WSB subReddit should not have enough influence on sufficient number of individual ticker to show a impact on the overall portfolio of the indices. It has been many time proved that such volatility dynamics are in a steady-state[30].

Nevertheless, similar methodology applied on intraday remains to be investigated. The idea comes from the GameStop short-squeeze. The following procedure would be changing from the previous data preprocessing and featuring: Create a lexicon of common names for a particular ticker, to identify a particular comment discussing about Tesla or its indirect holders. For example Telsa would be Tesla, TSLA, Elon Musk, Elon, Tlo, ... ; Keep and perform only sentiment analysis on those latters posts; Potentially aggregate data in time window (i.e. every 15 minutes); Run the model to predict an intraday movement of this ticker. One can spectulate that following a heavy discussion on the forum about it, a direct consequence can be seen in the ticker intraday return.

This latter approach would identify tiny investors behaviour on the market. In some sense, one can identify a similar short-squeeze that happened for GameStop with a bot and surf on the wave with the remainder users.

V. Conclusion

The present work examined the ability of public sentiment to influence financial markets. First, the posts and comments from the subReddit forum *WallStreetBets*are gathered through already existing databases. After a variety of data pre-processing steps, a multitude of feature engineering possibilities are attempted, before curating two different Neural Network architectures to tackle classifying the assembled dataset. Both, Long Short-Term Memory Neural Network and Temporal Convolutional Neural Network, lead to poor results when trying to predict a test set. The final accuracies of 50% and 59% respectively for the negative and positive returns, obtained by the Temporal Convolutional Network on the testing set, do not appear significantly different from the 50% accuracy which would be obtained by flipping a weighted

coin with probability-weights of the of positive and negative returns respectively. This result discredits the model's performance and ability to predict the market's movement from these assembled sentiment dataset. However, those investigations revealed future possible research opportunities and improvements. In order to further assess the ability of such a platform to predict asset returns, an intraday prediction focused on a particular ticker could show good result for extreme events prediction. Moreover, these *tiny-investors* are active, operate in volumes and are growing day by day on the *WallStreetBets* subReddit.

Appendix A
Github Code

The code can be found in the following Github repository:
github.com/Amin-Debabeche/ML_Finance

References

[1] F. B. Oriani and G. P. Coelho, "Evaluating the impact of technical indicators on stock forecasting," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2016, pp. 1–8.

[2] J. Huang, J. Chai, and S. Cho, "Deep learning in finance and banking: A literature review and classification," *Frontiers of Business Research in China*, vol. 14, no. 1, p. 13, Jun. 2020. [Online]. Available: https://doi.org/10.1186/s11782-020-00082-6

[3] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970, publisher: [American Finance Association, Wiley]. [Online]. Available: https://www.jstor.org/stable/2325486

[4] C. R. Stephens, H. A. Benink, J. L. Gordillo, and J. P. Pardo-Guerra, "A New Measure of Market Inefficiency," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 1009669, Aug. 2007. [Online]. Available: https://papers.ssrn.com/abstract=1009669

[5] Z. Umar, I. Yousaf, and A. Zaremba, "Comovements between heavily shorted stocks during a market squeeze: Lessons from the GameStop trading frenzy," *Research in International Business and Finance*, vol. 58, p. 101453, Dec. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S027553192100074X

[6] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, number: 1. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[7] "r/wallstreetbets subreddit stats (wallstreetbets)." [Online]. Available: https://subredditstats.com/r/wallstreetbets

[8] Honnibal, Matthew and Montani, Ines, "{spaCy 2}: Natural language understanding with {B}loom embeddings, convolutional neural networks and incremental parsing," 2017, to appear. [Online]. Available: https://spacy.io/

[9] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes," in *2019 12th International Conference on Information Communication Technology and System (ICTS)*, Jul. 2019, pp. 49–54.

[10] S. Kunal, A. Saha, A. Varma, and V. Tiwari, "Textual Dissection of Live Twitter Reviews using Naive Bayes," *Procedia Computer Science*, vol. 132, pp. 307–313, Jan. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918309165

[11] K. Ahmed, N. E. Tazi, and A. H. Hossny, "Sentiment Analysis over Social Networks: An Overview," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2015, pp. 2174–2179.

[12] D. Appelbaum, H. K. Duan, T. Sun, and H. Hu, "Business News Headlines and the Prophetic Vision of Bankruptcies: An Application of Natural Language Processing," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3906091, Aug. 2021. [Online]. Available: https://papers.ssrn.com/abstract=3906091

[13] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP," p. 6.

[14] A. Panchbhai and S. Pankanti, "Exploring Large Language Models in a Limited Resource Scenario," in *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, Jan. 2021, pp. 147–152.

[15] "Data Mining," in *Mining of Massive Datasets*, A. Rajaraman and J. D. Ullman,

Eds. Cambridge: Cambridge University Press, 2011, pp. 1–17. [Online]. Available: https://www.cambridge.org/core/books/mining-of-massive-datasets/data-mining/E5BFF4C1DD5A1FB946D616D619B373C2

[16] "S&P 500 (^GSPC) Charts, Data & News - Yahoo Finance." [Online]. Available: https://finance.yahoo.com/quote/%5EGSPC/chart/

[17] N. Cetorelli, B. Hirtle, D. P. Morgan, S. Peristiani, and J. A. C. Santos, "Trends in Financial Market Concentration and their Implications for Market Stability," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 975307, Mar. 2007. [Online]. Available: https://papers.ssrn.com/abstract=975307

[18] P. Embrechts and M. Hofert, "Statistics and Quantitative Risk Management for Banking and Insurance," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 493–514, Jan. 2014. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-statistics-022513-115631

[19] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools - Revised Edition*. Princeton University Press, May 2015, google-Books-ID: l2yYDwAAQBAJ.

[20] H. Sorensen, D. Jones, M. Heideman, and C. Burrus, "Real-valued fast Fourier transform algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 6, pp. 849–863, Jun. 1987, conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

[21] D. Deksne, "Chat Language Normalisation using Machine Learning Methods:," in *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*. Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2019, pp. 965–972. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007693509650972

[22] T. J. Sefara, "The Effects of Normalisation Methods on Speech Emotion Recognition," in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, Nov. 2019, pp. 1–8.

[23] Z. Liao and G. Carneiro, "On the importance of normalisation layers in deep learning with piecewise linear activation units," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–8.

[24] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv:1803.01271 [cs]*, Apr. 2018, arXiv: 1803.01271. [Online]. Available: http://arxiv.org/abs/1803.01271

[25] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," p. 25.

[26] K. Joshi, B. H. N, and J. Rao, "Stock Trend Prediction Using News Sentiment Analysis," *International Journal of Computer Science and Information Technology*, vol. 8, no. 3, pp. 67–76, Jun. 2016. [Online]. Available: http://aircconline.com/ijcsit/V8N3/8316ijcsit06.pdf

[27] P. Hofmarcher, S. Theußl, and K. Hornik, "Do Media Sentiments Reflect Economic Indices?" p. 6.

[28] Y. Shynkevich, T. McGinnity, S. Coleman, and A. Belatreche, "Predicting Stock Price Movements Based on Different Categories of News Articles," in *2015 IEEE Symposium Series on Computational Intelligence*, Dec. 2015, pp. 703–710.

[29] R. Schumaker and C.-N. Huang, "Sentiment Analysis of Financial News Articles," Jan. 2009.

[30] M. M. Rounaghi and F. Nassir Zadeh, "Investigation of market efficiency and Financial Stability between S&P 500 and London Stock Exchange: Monthly and yearly Forecasting of Time Series Stock Returns using ARMA model," *Physica A: Statistical Mechanics and its Applications*, vol. 456, pp. 10–21, Aug. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378437116002776