高頻度注文情報の符号化と深層学習による短期株価予測

田代　大悟[1*]　和泉　潔[1]
Daigo Tashiro[1]　Kiyoshi Izumi[1]

[1] 東京大学大学院工学系研究科
[1] Graduate School of Engineering, The University of Tokyo

Abstract: Predicting the price movements of stocks based on deep learning and high frequency data has been studied intensively in recent years. Especialy, limit order book which describes supply-demand balance of the market is used as feature of a neural network, however, these methods do not utilize the properties of market orders. On the other hand, order encoding method of our prior work can take advantage of these properties. In this paper, we apply some types of convolutional neural network(CNN) architectures to order-based features to predict the direction of mid-price movements. The results show that smoothing filters which we propose to employ over embedding features of orders improve accuracy. Furthermore, inspection of embedding layer and investment simulation are conducted to demonstrate the practicality and effectiveness of our model.

1　序論

めまぐるしく変動している金融商品の価格を予測することは可能であるのか．という問いに対して，実務家だけでなく，様々な分野の学者の間で多くの理論研究，実証研究が行われてきた．効率的市場仮説の提唱や実証研究によるそれへの反証などを経て，最近では情報工学の分野，特に機械学習，深層学習を用いた手法が増えつつある．これらの手法は，データの背後に潜む規則や知識を発見するパターン認識の能力を備えており，市場予測において一定の成果を上げている．

市場予測に関する研究が多く行われている中，市場の様相も大きく変化した．市場の電子化と高速化に伴い，アルゴリズム取引や高頻度取引 (High Frequency Trading: HFT) といった機械的な取引が台頭し，市場で観測される注文データ，取引データはサンプリング頻度が非常に高く，また膨大化している．これらは「高頻度データ」と呼ばれ，有効な利用が期待されている．

高頻度データに対しても，深層学習を用いた研究が行われている．特に，板を用いたものが多い [1][2]．板の注文の価格と量を売り (アスク)，買い (ビッド) それぞれ数本分を入力として，ニューラルネットワークによる仲値の動向予測を行い，既存の機械学習を上回る精度を上げている．しかしこれらの手法には，成行注文に関してより重大な課題が存在する．それは，板のベストアスクまたはベストビッドの数量が減少したとき，それが成行注文によるものかキャンセル注文によるものか区別がつかない，というものである．成行注文とは，即時約定かつコストを支払う注文でトレーダーの強い意思を表したものであり，マーケットへのインパクトも大きい．さらに，成行注文とリターンとの相関も強いため，キャンセル注文のもつ意味，情報とは異なる．これをニューラルネットワークに識別させるには，注文系列自身をモデルの入力とすれば良い．

一方，指値注文やキャンセル注文も価格への影響を持つと言われており，これを無視することはできない [3][4]．そこで本研究では，すべての注文タイプを含めた高頻度注文系列と，深層学習を用いた短期の価格動向予測を行う．まず注文の符号化手法について説明し，予測モデルとして価格予測に特徴的な注文を捉えるよう CNN を改良した A-CNN(Average Convolutional Neural Network) と，その課題を踏まえて拡張した A-CNN+を提案する．本研究の目的は，この手法によるアルゴリズム取引の支援である．本研究の達成により，その運用パフォーマンスを向上することができると考える．

2　注文の符号化手法

注文の特徴には，価格，数量，時刻は数値情報 (量的変数)，売り買いの別，注文タイプはカテゴリ情報 (質的変数) というように，質の異なる変数をもって表され

*連絡先：東京大学大学院工学系研究科システム創成学専攻和泉研究室，〒 113-8654 東京都文京区本郷 7-3-1，E-mail: m2016dtashiro@socsim.org

る．本節では，これらをニューラルネットワークへの入力とするために用いる手法，注文の符号化手法について説明する．注文の符号化は 2 つの段階に分けられる．まず，i) 注文に属する変数に対してすべて質的変数側に変換するカテゴリ化．ii) その後，すべての変数を一つの質的変数に変換する 2 つの処理である．まず，i) の処理を行う意図について説明する．

注文タイプ

売り買いの別，成行注文，指値注文，キャンセル注文の注文タイプに関しては，すでにカテゴリ化された質的変数であり，以降，本研究では一つにまとめて注文タイプとして扱うこととする．

価格

質的変数をニューラルネットワークの入力とする場合，正規化を行うことが一般的である．しかし，注文の価格の分布は非対称で多峰性を持っており，標準偏差を求めるのは難しい．そこで，絶対価格ではなく仲値からの距離という相対価格を用いるとする．この度数分布はべき性を持ち，平均と標準偏差に意味をなさない．同じく標準化が難しいことから，数値情報ではなくそれをカテゴリ化して用いる．また，相対価格だと，仲値から遠い指値注文は価格への影響が小さいものとして大きく一つとして扱うことが可能であるということから，カテゴリ化の方が正規化に優っていると考えている．

時刻差

高頻度データを分析するにあたり，どのような投資家がどのような注文を出したのかと識別することは難しい．しかし，注文市場参加者の思惑は様々であり，発注者の情報を注文に関連付けることにより市場予測の精度を高めることができると考えられる．そこで，その指標として一つ前の注文との時刻差を情報に加える．これは例えば，ミリ秒オーダー間隔での注文は機械的なトレーダーによる注文だと識別し，発注者がどのような取引を行っているのかというおおよその分類情報を加えることを意図している．

次に ii) について説明する．複数のカテゴリを統一し一つの符号として表す目的として，ニューラルネットワークへ入力する注文を均質にするということが挙げられる．また，同じ価格であっても，指値注文のそれとキャンセル注文のそれでは意味が異なる．さらに，ニューラルネットワークの第一層と注文の関係を一意にすることができ，後に行う埋め込み層の分析を容易にするという利点もある．

Figure 1 には，具体的な注文の符号化手法を挙げる．価格の影響の少ない仲値から遠い価格は粒度を粗くカテゴリ化する．時刻差カテゴリは，500 ms を境に機械的な注文か否か，さらに機械的な注文の中でも 20 ms を境に高速な注文か比較的低速な注文なのかを識別できることを期待して設定した．

$$
\begin{pmatrix} \mathrm{MarketOrder}^{\mathrm{ask}} \\ \mathrm{MarketOrder}^{\mathrm{bid}} \end{pmatrix} \times \begin{pmatrix} \sim 20 \\ 20 \sim 500 \\ 500 \sim \end{pmatrix}
$$

$$
\begin{pmatrix} \mathrm{LimitOrder}^{\mathrm{ask}} \\ \mathrm{LimitOrder}^{\mathrm{bid}} \\ \mathrm{Cancel}^{\mathrm{ask}} \\ \mathrm{Cancel}^{\mathrm{bid}} \end{pmatrix} \times \begin{pmatrix} \sim 1 \\ 1 \sim 2 \\ 2 \sim 3 \\ 3 \sim 5 \\ 5 \sim 7 \\ 7 \sim 10 \\ 10 \sim \end{pmatrix} \times \begin{pmatrix} \sim 20 \\ 20 \sim 500 \\ 500 \sim \end{pmatrix}
$$

注文タイプカテゴリ　　価格カテゴリ (円)　　時刻差カテゴリ (ms)

Figure 1: 本研究で用いた注文の符号化.

3　深層学習による価格予測

3.1　データセット

データは FLEX-FULL ヒストリカルデータを用いた．銘柄は全 20 銘柄を用いる．期間は各々2013 年 7 月 1 日から 2014 年 6 月 30 日までの 1 年間のうち，東京証券取引所の営業日 245 日分を使用した．それぞれの銘柄に対し，予測対象となる価格は約定価格と仲値，各予測対象に 2 クラス，3 クラスの分類問題を設定した．

本節では，アルゴリズム取引での運用を考慮して価格動向予測のデータセットの作成を行う．価格予測モデルの入力となる注文系列 (イベントドリブン) と出力である価格動向を決定するタイミングを，アルゴリズム取引と同じ時間スケール (タイムドリブン) に沿ったものにする．具体的には，符号化済みの一つの注文系列を 30 秒毎で間隔で区切りを設定し，複数の注文系列を獲得する (Figure 2)．このとき，分割された注文系列 $\mathcal{S}^j$ は可変長となるが，その取り扱いは予測モデル側で解消する．

$\mathcal{S}^j$ に対応する価格の動向クラス d^j は，間隔の終点から 30 秒後の価格を対象とする (Figure 2)．価格の上がり/下がりを予測する 2 クラス分類問題，上り/下り/変動しないを予測する 3 クラス分類問題を設定した．

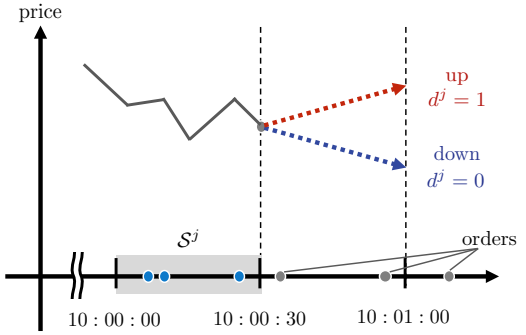

Figure 2: 系列 $\mathcal{S}^j$ のサンプリングと 2 クラス分類におけるラベリング.

3.2 CNN による予測モデル

CNN は画像認識の分野だけでなく，自然言語処理の分野においても，文書分類といったタスクで成功を収めている [5][6]．本研究で注文時系列に対して CNN を用いる理由として，CNN が位置に対して不変性を有している点が挙げられる．

CNN は畳み込みの後に系列方向に最大プーリングを行うため，指値注文やキャンセルが系列の後方に集中したとしても，成行注文を認識し，価格動向のシグナルとなる特徴やパターンを掴むのに有利だと考えられる．このような理由から，本研究では，局所的な注文系列を畳み込み，パターンを抽出する CNN を用いたモデルを用いる．

注文 x_t の埋め込みベクトル $\boldsymbol{x}_t$ を用いてパディングによって n で統一された系列 $\mathcal{S}$ は，$\boldsymbol{x}_{1:n} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ と表現する．これの局所的な行列 $\boldsymbol{x}_{i:i+h}$ を畳み込み，活性化関数を適用することによって新たな特徴 c_i を得る．ストライド幅 1 として，$(\boldsymbol{x}_{1:h}, \boldsymbol{x}_{2:h+1}, \ldots, \boldsymbol{x}_{n-h+1:n})$ に対してこの畳み込みを行うと，次のような新たな特徴ベクトル $\boldsymbol{c} = [c_1, c_2, \ldots, c_{n-h+1}]^T \in \mathbb{R}^{n-h+1}$ を得る．その $\boldsymbol{c}$ 最大プーリングを行うことによって，一つのフィルタから得られる一つの特徴量 $\hat{c} = \max(\boldsymbol{c})$ を得る．複数の畳み込みフィルタに対して，この処理を行う．k_{conv} 個のフィルタによる畳み込み演算と最大プーリングによって得られる特徴量 $\hat{\boldsymbol{c}} = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{k_{\text{conv}}}]^T$ を全結合層へ入力し，その出力をソフトマックス関数による変換後，出力クラスを得る．畳み込みフィルタのサイズを複数にするのは，その大きさによって，パターンを抽出するための系列の長さを変えるためである．つまり，大きいサイズのフィルタであれば，より多くの注文の関係性を捉え，反対に小さいフィルタであれば，局所的な関係性を捉えようとする．

3.3 注文の埋め込みの平均化を利用した予測モデル

本項では，注文の埋め込みの平均化を利用した予測モデル (Average Convolutional Neural Network: A-CNN) を提案する．金融市場での時系列データでは，一定の価格トレンドが観測されても，注文単位などのミクロ構造での明確なパターンは少ない．テキストデータとは異なり，注文の時系列では，畳み込みフィルタが捉える局所的な領域での，注文の相互作用が小さいと考えられる．そこで，一定数の注文の平均をとることで，これらの影響を小さくすることを考える．記号列に対して平均を取ることはできないので，埋め込み行列の局所的な範囲を対象とした平均化を行う．

埋め込み行列 $\boldsymbol{x}_{1:n}$ に対して平均プーリングを，窓幅 $1 \times l_{\text{pool}}$ で適用する．プーリング前の注文時系列方向上下それぞれのパディングサイズ l_{pad} によるパディングを行う．プーリング後の特徴行列は，床関数を $\lfloor x \rfloor$ とすると，$\boldsymbol{x}_{\text{pool}} \in \mathbb{R}^{e \times \lfloor \frac{l_{\text{pad}}+n}{l_{\text{pool}}} \rfloor}$ と表すことができる．その後は前項の CNN と同様に順伝搬および学習を行う．

また平均化による効果は，成行注文の相互作用を捉えるために一役買う．前項の CNN では，成行注文の間隔が，畳み込みフィルタのサイズより大きい場合に，成行注文の関係の特徴を抽出できない．しかし平均化を加えることで，注文系列に対するフィルタが捉える長さを $l_{\text{pool}} \times h$ に広げることができる．このようにすることで，出現頻度の低い成行注文間の相互作用を捉えることができる．しかし，成行注文がプーリングの窓 l_{pool} 内に複数存在する場合に，それらを平均化してしまうデメリットが生じる．

3.4 A-CNN の拡張モデル

本項では前項の A-CNN をさらに拡張したモデル (以下 A-CNN+) を提案する．Figure 3 にその図を示す．注文の埋め込み行列から，窓幅の異なる平均プーリングを行うことにより，複数の平均化行列を作る．そしてそれぞれの平均化行列に対して畳み込みフィルタを用意し，畳み込みと最大プーリングをとった後，ベクトルを連結し，上述の CNN と同様に全結合層へと入力する．

この操作の目的は，畳み込みフィルタのサイズを複数持たせるのと同様で，注文系列の文脈の大小に多様性を持たせるためである．つまり，平均プーリングのサイズが大きいほど，畳み込みフィルタはよりグローバルから特徴を抽出し，小さいほどローカルから特徴抽出を行う．ここでは，注文系列の中で価格変動に有用なパターンが異なる大きさで存在すると仮説を置いている．グローバルでは価格時系列でいうモメンタムのような大きなトレンドが，ローカルではよりミクロな注文の相互作用といったパターンがあると考えている．この複数のプーリングの窓幅と複数の畳み込みフィルタの窓幅を設定することで，注文間の相互関係の捉える幅を相乗的に変化させ，あらゆる長さのパターンに対応させる．また A-CNN での，プーリングの窓内に出現する複数の成行注文の平均化まで行ってしまう課題を解決する．

3.5 結果

各銘柄 1 年間のデータのうち，学習用データ，検証用データ，評価用データとして，古い順に 7 : 1.5 : 1.5 に分割した．20 銘柄の各実験，手法ごとに、学習用データを用いたモデルの学習とパラメータ探索を行い，検証用データ用で最も評価の高いモデルを選択する．評

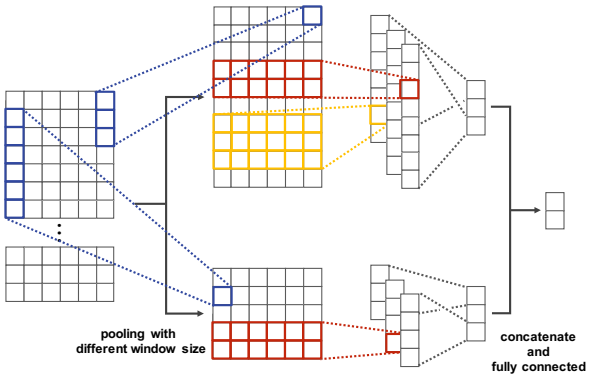

Figure 3: A-CNN+の構造.

価では，評価用データにおける各クラスに対する F1 値をそれぞれ求め，平均を取ったもので行う．

実験結果の一つを Figure 4 に示す．これは仲値の 2 クラス予測実験のモデルの評価用データでの F1 値である．すべての銘柄で提案手法がベースラインを上回る．A-CNN と A-CNN+では，実験や銘柄によって優劣が異なる．

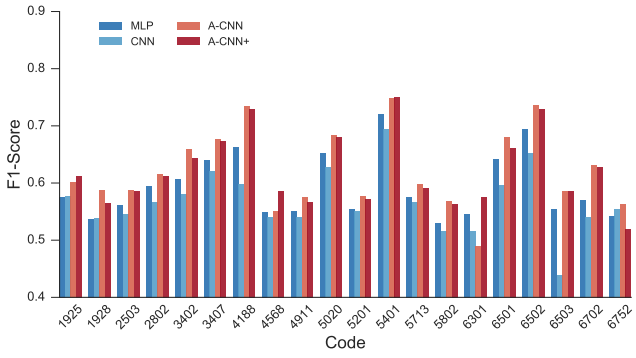

Figure 4: 仲値予測の 2 クラス分類問題における各手法の F1 値．ベースラインとしてロジスティック回帰，非線形 SVM(Support Vector Machine)，MLP(Multi Layer Perceptron)，3.2 節の CNN を用いた．ここでは，ベースラインの中で最もスコアの高かった MLP，CNN と，提案手法である A-CNN と A-CNN+を示す．

3.6 埋め込み層の分析

本節では，埋め込み層の分析を行う．各注文に対応したベクトルは各注文を表現したベクトルであり，そのノルムはニューラルネットの発火の強さであると考えることできる．Figure 5 は，2 クラス分類問題の，対象を約定価格とした実験と，仲値とした実験それぞれの，ノルムの平均を示している．約定価格予測では成行注文の重みが突出して高いのに対して，仲値予測では仲値に近いところに出された注文やキャンセルは成行注文と同等の重みを置かれている．

これは，仲値の予測が，約定価格の予測と比較して指値注文，キャンセル注文の影響を受けやすいためである．仲値は，ベストアスク，ベストビッド付近の指値注文とキャンセル注文が入ることにより変動するが，約定価格は成行注文が入ることでしか変動しない．仲値も成行注文によって変動するので，成行注文を含めすべての注文のタイプに反応するようなモデルになっていると考えられる．この分析結果は，約定価格と仲値の性質から見て納得のいくものとなっており，提案手法が意図の通りに動作していることを示している．

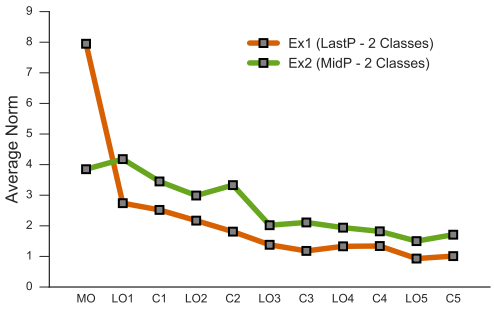

Figure 5: A-CNN モデルにおける埋め込みベクトルのノルム平均の詳細．成行注文と，各価格帯における指値注文，キャンセル注文の埋め込みベクトルのノルムを，すべての銘柄の対して平均を取る．MO は成行注文，LO は指値注文，C はキャンセル注文を指す．LO1 や C2 の数字は価格カテゴリを表す．1 は Figure 1 における価格カテゴリ ～ 1 を，2 はカテゴリ 1 ～ 2 を，3 はカテゴリ 2 ～ 3，4 はカテゴリ 3 ～ 5 を，5 はカテゴリ 5 ～ 7 である．LastP は約定価格を，MidP は仲値を指す．

4 投資シミュレーション

ニューラルネットワークの出力層にはソフトマックス関数を用いており，確率を出力する．モデルの出力するこの確率が高い場合のみ採用し投資行動をとることは，精度を高め，取引一回あたりのパフォーマンスを上げることができると考えられる．そこで，出力の確率の最大のものに閾値を設け，投資行動を決定することを考える．

この出力の確率が高いほど，うまく特徴を捉え信頼性の高い出力となることを裏付けるものととして，Figure 6 を示す．これらは，各評価用データにおいて，出力に閾値を設けた場合の Precision である．出力が各閾値を超えたサンプル数の中で Precision を算出する．閾値を高く設定するほど，評価値が高くなり，その閾値を超える確率を出力する回数が減少する傾向が見られる．

投資テストを行うにあたり，収益の高さによって 2 クラス分類問題と 3 クラス分類問題の比較を行う．モデルは，A-CNN と A-CNN+のうち検証用データで最も F1 値が高かったものを用いる．

コストは，取引を行う度にかかるものとする．簡単のためスプレッドを均一の 1 円とし，取引一回にかかるコストとして，ハーフスレッドの 0.5 円とする．さ

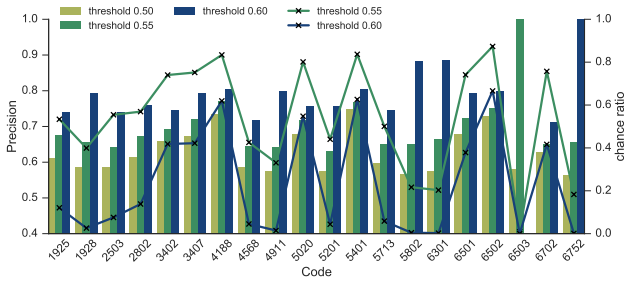

Figure 6: 出力に閾値を設けた場合の Precision と予測機会の割合 (2 クラス分類)．棒グラフ: Precision(左軸)．折れ線グラフ: 出力の確率が閾値を超えたサンプル数の割合 (右軸)．閾値は (0,50, 0.55, 0.60) とした．サンプル数の割合は評価用データのサンプル数を 1 としたときの割合を示している．閾値が 0.5 の場合は 1 となる．

らに，すべての銘柄で運用を行ったとして，このポートフォリオから得られる利益を算出する．

結果を Figure 7 に示す．閾値を変更しつつ投資テストを収益を計算した．それぞれの分類問題いずれもの特徴として，閾値を大きくしていくとある点を境に収益がプラスに転ずることがわかる．さらに大きくしていくと，収益が最大となる点を通り，減衰し 0 に近く．収益がプラスとなっている領域は，一回の取引でハーフスプレッドのコストを上回る領域である．閾値を大きくしすぎると一回の予測の精度を高めることができるものの，取引の回数が小さくなるため収益の和は減衰する．

これから，本研究で提案した注文の符号化手法と，埋め込み行列の平均化を用いた提案手法が実用的であることがわかる．そして両者を比較すると，より大きな収益を上げることができ多くの閾値でプラスの収益を上げることのできる 2 クラス分類の方が良い結果だと言える．これは 2 クラス分類問題の方が，3 クラス分類問題より学習が容易であったためだと考えられる．

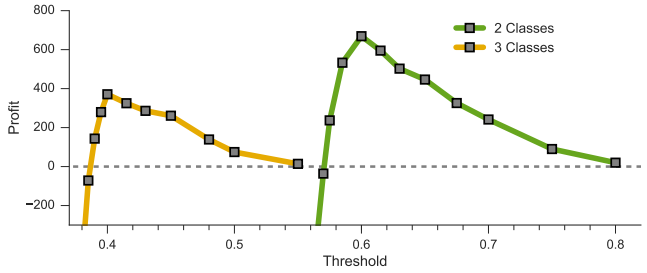

Figure 7: 閾値を変化させテストを行ったときの収益の推移．コストを設定した上で，すべての銘柄に対して得られる収益の合計．

5　まとめと今後の課題

本研究では，株価の短期動向予測のため，注文の符号化手法を提案し，価格動向予測に有用な特徴的な注文を捉えるのに適するように CNN を改良した．

大規模な高頻度データから複数のタスクを設定し実験を行い，提案した手法が他の手法より優れていることを示した．さらに，埋め込み層による分析を行い，その意図が機能していることを確認した．また，すべてのタスクで成行注文に強く反応するモデルであることを，仲値の予測問題では，仲値に近い指値注文とキャンセル注文も成行注文と同等に注目していることを発見した．

最後に，提案手法の実務的な有用性を示すために投資テストを行った．成行注文による仲値の売買を行い，プラスの成績を上げることで，実用面でも耐えうるモデルであることを示した．さらに，出力の確率に閾値を設定し，2 クラス分類のモデルを 3 クラス分類に適用することで，3 クラス分類のモデルをパフォーマンスの面で上回った．

今後の課題として，板の情報との組み合わせが挙げられる．板の情報は流動性という面で，注文系列に不足する情報を補足することができる点で，さらなる精度向上が見込まれる．

参考文献

[1] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, Using Deep Learning to Detect Price Change Indications in Financial Markets, *European Signal Processing Conference* (2017).

[2] M. F. Dixon, Sequence Classification of the Limit Order Book using Recurrent Neural Networks, *Journal of Computational Science* (2017).

[3] Z. Eisler, J. P. Bouchaud, J. Kockelkoren, The price impact of order book events: market orders, limit orders and cancellations, *Quantitative Finance*, vol. 12, pp. 1395-1419 (2012).

[4] R. Cont, A. Kukanov, and S. Stoikov, Price impact of order book events, *Journal of Financial Econometrics*, vol. 12, pp. 47-88 (2014).

[5] Y. Kim, Convolutional Neural Networks for Sentence Classification, *Empirical Methods in Natural Language Processing*, pp. 1746-1751 (2014).

[6] R. Johnson, and T. Zhang, Effective Use of Word Order for Text Categorization with Convolutional Neural Networks, *North American Chapter of the Association for Computational Linguistics* (2014).