

畳み込みニューラルネットワークを用いた アナリスト往訪記録における景況感判定

Business Confidence Prediction for Analyst Report using Convolutional Neural Networks

高山 将丈^{*1}
Shota TAKAYAMA

小澤 誠一^{*1,2}
Seiichi OZAWA

廣瀬 勇秀^{*3}
Takehide HIROSE

飯塚 正昭^{*3}
Masaaki IIZUKA

^{*1} 神戸大学大学院 工学研究科
Graduate School of Engineering, Kobe University

^{*2} 神戸大学 数理・データサイエンスセンター
Center for Mathematical and Data Sciences, Kobe University

^{*3} 三井住友 DS アセットマネジメント株式会社
Sumitomo Mitsui DS Asset Management Company, Limited

To decide valuable companies to be invested, investment trust and fund management companies, which manage funds deposited from investors, have collected information on company's budget status and plans. However, the number of visit reports are usually too large even for skilled fund managers to easily derive reliable business outlooks and investment decisions. In this research, to alleviate fund managers' and analysts' commitment for the investigation and analysis, we propose a machine learning system that can support them to make accurate predictions on business outlook from collected visit reports. We attempt to predict business confidence for specific companies and industries using CNN that is expected to have good readability and robustness for polarity perturbation. As a result, we obtain 81.4% in classification accuracy for analysts' reports provided by the Sumitomo Mitsui DS Asset Management Company, Limited. It has 5.7% better accuracy than the best baseline model using Word2Vec and SVM.

1. はじめに

投資家の資金を預かり、その運用を行う運用会社では、投資対象の企業を決定するため、各企業にアナリストがヒアリングして財務状況や将来計画などの情報を収集している。この調査分析結果は往訪記録やアナリストレポートとしてまとめられ蓄積されているが、その量が膨大であるため、たとえ熟練したファンドマネージャであっても、これらから適切な投資判断を導き出すことは容易でない。

この業務を支援するため、アナリストレポートに対して深層学習を用いて景況感判定を行う試みが行われている[小林 2017]。小林らの手法では、アナリストレポートから景況感判定の根拠となる表現を得るため、学習を行う前にキーとなる表現を人手で与える。次いで、得られた特徴量を 12 層の階層型ニューラルネットワークを伝播させ、判定を行う。しかしこの手法には、個々のアナリストの書き方の違いによって、有用表現が異なるため、分析者が分析対象文書の特徴を十分に把握している必要がある点や、勾配消失及び発散、過学習などが懸念されるなどの問題が存在する。

一方で、深層学習、特に畳み込みニューラルネットワーク[LeCun 1998](CNN)を用いて文書分類を行う研究も行われている[Kim 2014]。この研究では学習済みの単語ベクトルを単語数分並べ、(単語ベクトル長×単語数)の 2 次元ベクトルとして文書ベクトルを作成した後、CNN を用いて文書の分類を行っている。

本研究の最終目標は、投資先企業を決定するファンドマネージャをサポートするシステムの開発である。時々刻々と状況が変化する投資業界において実用的なシステムを開発するためには、即時性が必要となる。本研究では、CNN で文書分類を行う

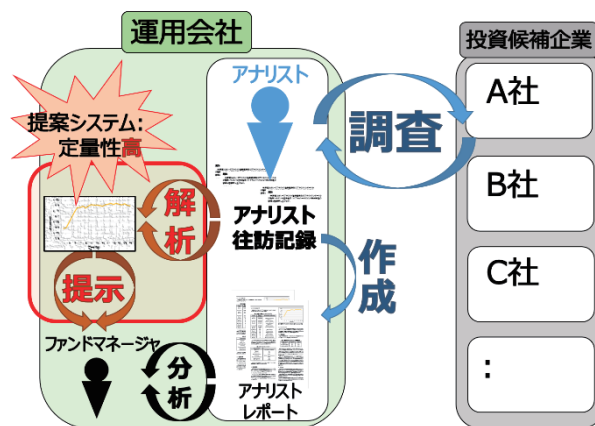


図 1 提案システムの概念図

Kim らの研究[Kim 2014]をベースに、往訪記録やアナリストレポートなどの金融文書の特徴を活かした新たな機械学習システムを提案する(図 1)。

2. アナリスト往訪記録について

アナリストは投資候補企業に訪問したり説明会に参加したりして情報を収集する。調査結果の正式な報告に先立ち、まずアナリスト往訪記録と呼ばれる文書に簡潔にまとめる。継続的に調査を行い、十分に情報を収集したところで、アナリストレポートと呼ばれる報告書にまとめる。このアナリストレポートを 1 つの情報源としてファンドマネージャは投資先企業を選定する。

しかし、この投資先企業を選定フローには 3 つの問題がある。1 点目は、膨大な数に上るアナリストレポートを担当部署の各員が目を通し、投資先企業を決定しているということである。アナリストレポートは図表も文章量も多いため、各ファンドマネージャが複数のレポートを通読するのに相応の時間が必要となる。2 点目は、アナリストが調査先企業の情報を入手してから、アナリ

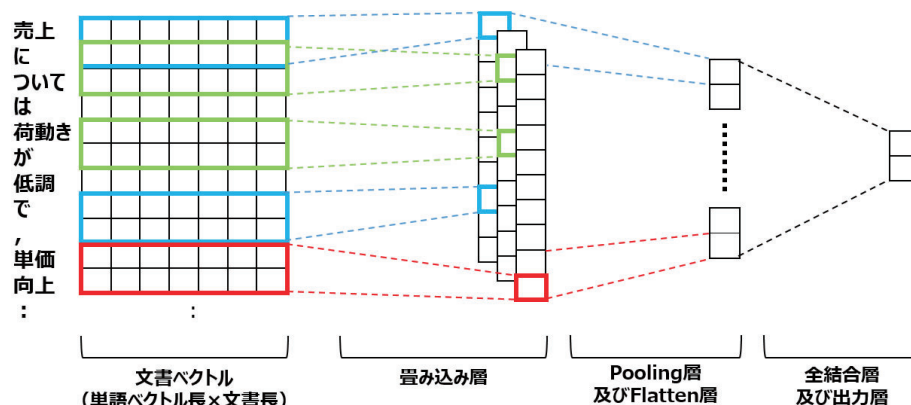


図 2 CNN を用いた文書分類モデルの概念図

ストレポートを発行するまでに時間がかかるということである。アナリストは十分な情報を収集してからレポートを発行するため、判断の遅れが大きな機会損失につながる可能性がある。3 点目は、アナリストレポートには、アナリストが十分な自信や裏づけを持った内容しか掲載しないということである。投資判断を左右する情報源を提供するため、アナリストは自らの提出したアナリストレポートの説明責任を負っている。したがってアナリストの率直な意見や経験に基づく勘など、業績判断に有用だが裏打ちの少ない内容はアナリスト往訪記録のみに記載され、アナリストレポートには記載されない傾向がある。その結果として投資機会を逃してしまうケースが少なからず存在する。

これらの問題点を踏まえて、人間が読むには多くの負担がかかるが、アナリストの本音やニュアンスといった有用な情報が含まれていると考えられ、速報的に発行されるアナリスト往訪記録を分析し、定量的な出力を行うモデルを構築することで、これらの問題点を解決すると共に、有用なシステム開発の一助になると考えている。

3. 提案手法

本論文では文書の判定に CNN を用いた。CNN の模式図は図 2 の通りである。CNN を用いた理由の 1 つ目として、CNN の特徴である位置不変性に注目したからである。アナリスト往訪記録では、公平性を期した報告の必要性から、一文書内に極性が混在するという特性がある。また 2 つ目として、モデルの学習結果の可視化が容易だと考えたからである。運用会社は投資家に対し自身の投資の根拠について説明する責任を持つ。そこで可読性と極性揺らぎへのロバスト性をあわせもつと期待される CNN を用い、往訪記録に対する景況感判定を試みた。

3.1 判定を行うネットワーク

先行研究として既に CNN を用いて文書分類を行う研究がなされている[Kim 2014]。Kim らは極性判定タスクでの検証に映画レビューデータセットを使用している。しかしこのデータセットに比べて、今回の解析対象のアナリスト往訪記録には 1 文書内に極性が混在する文章が多い。これは自らの調査結果をポジネガ双方向からの視点で述べる必要があるからである。したがって局所表現ではなく、文書全体の傾向を見る必要があると予想した。すなわち、プーリング層にはデータの削除によってデータサイズを小さくする MaxPooling ではなく、全体的な特徴を残したままデータを伝播できる AveragePooling がより適切だと考えた。

3.2 文書ベクトルの作成

アナリスト往訪記録及びアナリストレポートについて適当な前処理を行った後、McCab[工藤 2004]を用いて分かち書きを行った。活用を行う品詞は全て原型に変換した。辞書として NEologd[Sato 2015]を用いた。

次に、単語ベクトルを作成した。上記で作成したコーパスを元に Word2Vec[Mikolov 2013]で各単語ベクトルの学習を行った。本論文では単語ベクトルの学習に用いたコーパスにはアナリスト往訪記録及びアナリストレポートを用いた。アナリストが用いる語彙は特徴的であるためこのドメインに特化したモデルの構築という目的においては、大規模なコーパスを用いるより、少ないながらもドメインに特化したコーパスを用いる方が良いと考えたからである。

最後に文書ベクトルを作成した。作成対象の文章に出現する順番に単語ベクトルを並べてゆき、(Word2Vec の次元数×各文書の単語数)の 2 次元配列として文書ベクトルを作成した。固定長のベクトルとしてモデルに入力するため、今回は各文書で 300 単語を上限として文書の前方から取り出した。また 300 単語に満たない文書においては、文書ベクトル後部をゼロパディングし、サイズを統一した。

4. 性能評価

4.1 使用したデータ

本研究では、三井住友 DS アセットマネジメント株式会社(旧、大和住銀投信投資顧問株式会社)の協力のもと、アナリスト往訪記録及びアナリストレポートを使用した。各データ数の詳細は表 1 の通りである。

表 1 使用した金融文書の構成

文書種類		数量
アナリスト往訪記録	ポジティブ	838
	ネガティブ	573
	ニュートラル	1,834
	ラベルなし	12,261
アナリストレポート		21,892

事前学習として Word2Vec の学習を行う必要があるが、今回はアナリスト往訪記録及びアナリストレポートの全文章を用いた。また提案手法の評価には、入力した文書から景況感判定を行うためラベル付きのデータが必要である。今回の評価ではアナリ

スト往訪記録のうち、ポジネガのラベルをもつ 1,411 文書のみを対象として評価を行った。

4.2 Word2Vecについて

今回は Word2Vec の学習に Python モジュールである gensim[Řehůřek 2010]を用いた。Word2Vec の学習に用いたパラメーターは表 2 の通りである。本研究では一文書内にポジネガが混在する特性を考慮し、文ごとに学習を行った。

表 2 Word2Vec のパラメーター

Parameters	
Vector size	500
Window	10
Min_count	1
iter	50

4.3 CNNについて

まず、本実験において用いた CNN の構成を表 3 に示す。またこのネットワークのパラメーターを表 4 に示す。

表 3 検証に使用した CNN のモデル

Name	Shape	Operation performed
Input	500×300×1	-
Conv-1	1×296×32	500×5 convolution 32 filters
Activate-1	1×296×32	ReLU activation
Pool-1	1×74×32	1×4 AveragePooling
Conv-2	1×74×32	1×5 convolution 32 filters
Activate-2	1×74×32	ReLU activation
Pool-2	1×18×32	1×4 AveragePooling
Flatten	576	-
Dropout-1	576	50 % dropout
Dence-1	32	Fully connected
Activate-3	32	ReLU activation
Dropout-2	32	50 % dropout
Dence-2	2	Fully connected
Activate-4	2	Softmax activation
Output	2	Binary classification

Input 層に入力されるデータが各文書ベクトルである。畳み込み層と Pooling 層を交互に 2 層ずつ、その後 2 層の全結合層の出力層で二値分類を行い、景況感判定を行うモデルである。

表 4 CNN のパラメーター

Parameters	
Optimizer	Adam[Diederik 2014]
Epochs	25
Batch sizes	100
Random Sampling	True
Validation	10-fold cross validation
Learning rate(Adam)	0.001
Beta_1(Adam)	0.9
Beta_2(Adam)	0.999

モデルの評価では 10-fold cross validation を行った。1 epoch 終了ごとにテストデータで評価を行い、精度を記録した。

このモデルの学習は 25 epochs 行い、最終 10 epochs 分の結果を平均し、精度とその分散を求めた。

4.4 比較手法

今回は提案手法の有効性を比較するための baseline として後述の 2 手法においても同様に景況感判定を行った。10-fold cross validation を行い、平均値をモデルの精度とした。

(1) Word2Vec+SVM[Boser 1992]

Word2Vec+SVM では、提案手法と同様に学習した単語ベクトルを文書内で加算平均することで文書ベクトルとした。この文書ベクトルを SVM に入力し景況感判定を行った。

SVM には Python モジュールである scikit-learn[Pedregosa 2011]を用いた。

(2) Doc2Vec+SVM

Doc2Vec+SVM では提案手法と同様のデータセットから学習した Doc2Vec を用いて文書ベクトルを作成した。文単位でベクトルを作成し、文書内で加算平均することで文書ベクトルとした。この文書ベクトルを用いて(1)と同様に評価を行った。

4.5 実験結果

今回の実験結果は表 5 のようになった。また提案モデルの Epoch 毎の Accuracy 変化は図 3 のようになった。

表 5 提案手法と baseline の精度比較

Model Name	Accuracy[%]	分散
提案手法	81.4	13.7
Word2Vec+SVM	75.7	17.0
Doc2Vec+SVM	66.8	6.19

比較手法のうち、提案手法が一番良い精度を得られた。baseline として比較したモデルでは Word2Vec+SVM が最も良い精度であった。十分なデータが存在していたため、単語ベクトルが十分に学習できていたためであると考えられる。また Doc2Vec では分散が最も小さい値となったが、Accuracy が最も低くなった。

また Epoch 毎の精度変化であるが、7epoch 程度で学習が一旦安定化していることがわかる。そして徐々に Accuracy が上昇した後細かな上下を繰り返している。これは過学習抑制を狙い、Dropout 層を 2 層使用しており、ランダムに結合重みが更新されているためであると考えられる。このまま学習を続けることで過

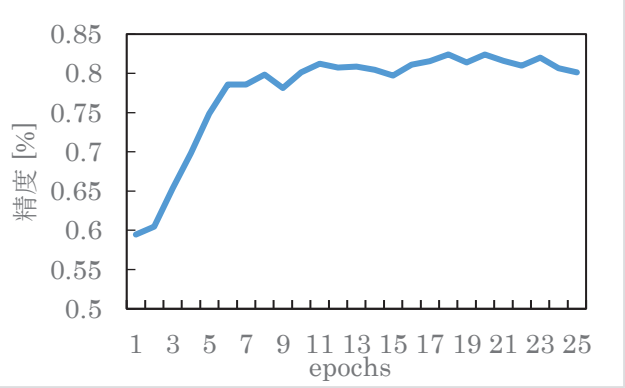


図 3 エポック毎の精度変化

学習が発生することが想定されるため、どのタイミングで学習を終え、実際の業務のシステムに組み込んで使用するかが今後の課題である。

5. まとめ

本研究では、運用会社の運営を補助するシステム開発の一助として、アナリスト往訪記録における文書の景況感判定に有効なモデルの提案を行った。その結果ラベルの付いたアナリスト往訪記録において 81%程度の性能が得られた。

今後の展望としては、今回の研究で分かち書きの際に使用した汎用的な辞書ではなく、アナリスト特有の語彙に対応した辞書の作成を考えている。

アナリスト往訪記録を定量的に分析できるモデルの構築が出来たことから、このモデルを礎に業務補助を行うシステムの開発へと進んでいくことが出来るであろう。

参考文献

- [Boser 1992] Bernhard E. Boser., Isabelle M. Guyon, Vladimir N. Vapnik: A training algorithm for optimal margin classifiers, COLT '92 Proceedings of the fifth annual workshop on Computational learning theory, Pages 144-152, ACM , 1992.
- [Diederik 2014] Diederik P. Kingma, Jimmy Lei Ba, Adam: A Method for Stochastic Optimization, ICLR 2015, 2015
- [Kim 2014] Yoon Kim: Convolutional Neural Networks for Sentence Classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics , 2014.
- [LeCun 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner: Gradient-Based Learning Applied to Document Recognition , Proceedings of the IEEE 86(11), Pages 2278-2324, IEEE , 1998.
- [Mikolov 2013] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space, In Proceedings of ICLR Workshops Track, 2013.
- [Pedregosa 2011] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al.: Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830 , JMLR, Inc. and Microtome Publishing , 2011.
- [Řehůřek 2010] Radim Řehůřek and Petr Sojka: Software framework for topic modelling with large corpora, THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS, University of Malta , 2010.
- [Sato 2015] Toshinori Sato, Taiichi Hashimoto and Manabu Okumura: Operation of a word segmentation dictionary generation system called NEologd (in Japanese), Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL) , Information Processing Society of Japan, 2016.
- [小林 2017] 小林 和正, 酒井 浩之, 坂地 泰紀, 平松 賢士: アナリストレポートからのアナリスト予想根拠情報の抽出と極性付与, 人工知能学会研究会資料, 一般社団法人情報処理学会, 2017.
- [工藤 2004] 工藤 拓, 山本 薫, 松本 悠治: Conditional Random Fields を用いた日本語形態素解析, 情報処理学会研究報告自然言語処理(NL), 一般社団法人情報処理学会, 2004.