トピックモデルによる市場変動要因の抽出

川原一修 [1]
[1]Japan Digital Design
{takanobu.kawahara}@japan-d2.com

概要

本稿では日々大量に生じる金融ニュースから市場の変動に影響を与えたトピックをトピックモデルを使用して効率的に抽出する手法を考察した。先行研究と比べてトピックの抽出に文章のコンテキストまで用いる点に新規性があり、比較実験でコンテキストを使用する手法の優位性を示したあと、実際に市場変動に影響を与えたトピックを抽出し人間の目で見ても違和感のないことを確かめた。

1 はじめに

金融アナリストが市場分析を行う際に、過去にあった、類似した環境での市場動向について調査し将来を予想する上での参考とすることが多々ある。しかし、金融市場は様々な要素で変動し、市場の値動きだけから過去の相場がどのような要素によってドライブされていたのか特定することは難しく、金融アナリストは過去の金融ニュースなど多量の文献調査に時間をさくケースがままある。そこで本稿では、市場関連ニュースのヘッドラインを使用してどのようなニュースが市場の動向に影響を与えていたのか検出する手法について考察した。日々大量に流れてくるニュースをトピックモデルによって大別し、その後シンプルな回帰モデルで市場変動に影響を与えていたトピックを効率的に検出する手法を提案する。時々の市場をドライブしていたトピックを明らかにすることで、金融アナリストの文献調査の補助情報として利用することを考えた。

2 先行研究

2.1 トピックモデル

文章の集合から、これらの文章の生成過程に潜んでいる潜在トピックを抽出しようとする研究はトピックモデルとしてこれまで盛んに研究されてきた。とりわけ、Latent Dirichlet Allocation(LDA)[1] や Non-Negative Matrix Factorization(NMF) [2] は実務でも広く利用されるようになった。これらの手法は大量の文章の中からトピックを効率的に抽出する上で有用な手法ではあるが、文章中に各単語が何回登場したかを数え上げて作成した行列 (Bag of Words, BoW) をベースとした手法であるため、単語同士のつながりが形成する文章のコンテキストを必ずしも活用しきれないことが欠点としてあげられる。

これに対して、特に Bidirectional Encoder Representation from Transformers(BERT) [3] の登場以来、文章の分散表現を使用してトピックの抽出に文章のコンテキストまで活用しようとする研究が行われている。節 3.1 で解説する BERTopic[3] の他にも、Sentence-BERT [4] による分散表現と BoW を同時に活用し、Valuational Auto Encoder を用いて潜在トピックを抽出する Combined Topic Model(CTM)[5] や,Doc2Vec [6] によって事前学習された分散表現からクラスターを探知することでトピックを抽出する Top2Vec[7] などでは、従来手法である LDA などと比べてトランスフォーマーベースの手法のパフォーマンスが優位だったとの報告がなされている。

2.2 トピックモデルを使用したマーケット変動要因の抽出

マーケットの変動を説明するトピックを抽出する研究としては [8] が挙げられる。[8] では、S&P500 指数の構成銘柄の価格リターンやボラティリティを説明するトピックを抽出する手法が研究された。手法としては複数の LDA のアンサンブルによるトピック抽出と相対的なトピックウェイトによるターゲット変数の回帰であり、アンサンブルの手法に様々な工夫が凝らされている。トピックのセンチメントを株価の将来変動の予想に活用した研究としては [9] や [10] があるが、これらの研究と異なり [8] では、将来の市場変動を予測することを研究のモチベーションにしているのではなく、変動の要因となっているトピックを検出することをモチベーションにしている点が本稿に近しい。[8]、本稿とも

に市場変動とテキスト情報で同時点のデータを使用して分析を進めていることをここで強調する。本稿との違いについては [8] では LDA ベースの手法を用いており、文章のコンテキストまで取り入れて分析を行うことの効果の検証は本稿で新規の観点だと考える。

3 理論

3.1 BERTopic

本節では BERTopic[11] について概観する。BERTopic では LDA などの従来手法と異なり、文章の生成に特別なモデルを想定しない。BERTopic のトピック抽出方法は要約すると以下の流れになる。

1. Sentence-BERT[4] による分散表現の抽出
2. UMAP[12] による次元削減
3. HDBSCAN[13] によるクラスタリングとトピックの抽出
4. c-TF-IDF によるトピックの代表単語の抽出

c-TF-IDF は通常ドキュメント単位で計算される TF-IDF をクラスター単位に拡張した手法であり、各単語 t のクラスター c 内でのウェイト W は以下の様に計算される

$$W_{t,c} = tf_{t,c} \cdot log(1 + \frac{A}{tf_t})$$

右辺第一項はクラスター内での単語の出現頻度 tf であり、通常の TF-IDF で inverse document freaquancy に相当する第 2 項は全クラスターの平均単語数 A を各々のクラスターに単語 t が登場する頻度で除した値となっている。このウェイトが高い単語ほどクラスター内での重要度が高い単語と解釈される。

3.2 トピックの評価指標

本稿ではトピックモデルの精度評価に Topic Coherence(TC) と Topic Diversity(TD) の 2 つの指標を採用した。TC の計算方法にはいくつかのバリエーションがあるが、本稿では normalized pointwise mutual information(NPMI) [14] を利用した。具体的には 2 つの単語 w_i, w_j の NPMI は以下の式で計算される。

$$NPMI(w_i, w_j) = \left(ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \right) \Big/ -lnp(w_i, w_j) \quad (1)$$

$$TC_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} NPMI(w_i, w_j)$$

式 1 の $p(.)$ はコーパスの中での単語の出現確率で、単語ペアが同時に出現する確率が高いほど NPMI の値は大きくなり、-1 から 1 の間の値を取る。TC_t はトピック t で上位 k 単語の NPMI の平均値であり、TC は全てのトピックについて TC_t の平均をとって計算される。使用する単語数 k はハイパーパラメータであり本稿では 10 を用いた。TC は人間の評価したトピックの精度との相関が高いとの研究 [15] があることから広く利用されるようになった。TD は各トピックにユニークな単語の割合で、0 から 1 の値をとる。広範なトピックを探知できていれば値は 1 に近づき、トピックの重複が多ければ値は 0 に近づく [16]。

4 データ

金融ヘッドラインデータ　本稿では Bloomberg 社のイベントドリブンフィードサービスから Textile News のヘッドラインを取得して利用する。本データセットにはニュースを識別する ID、ニュースがフィードされた時刻、言語、ヘッドラインなどの情報が提供されている。ニュースのヘッドラインのみを分析の対象とし本文は使用しない、また言語が英語のヘッドラインのみ使用する。ニュースによっては事後的にアップデートされるものがあり、同じ ID を持つニュースの中でフィード時刻が最新のレコードのみを使用してニュースの重複を回避した。使用したデータの期間と件数は以下の通りである。

1. トピックモデルの比較
   - 期間: 2022 年 12 月 5 − 7 日、13、14 日
   - 件数: 27,937
2. マーケット変動要因の抽出
   - 期間： 2018 年 1 月 1 日〜2022 年 6 月 30 日
   - 件数：7,232,088

本稿では以後このデータセットを金融ヘッドラインデータと呼ぶことにする。

市場データ　代表的な市場指数として S&P500 指数を利用する。日次で終値を Bloomberg から取得して利用する。

5 実験手法

本稿では 2 つの実験を行った。1 つ目の実験は金融ヘッドラインニュースに有効なトピックモデルの探索を行うための比較実験であり、複数のトピックモデルをデータに当てはめて得られたトピックの品質を比較する。2 つ目の実験はトピックモデルに

よって得られたトピックのセンチメントが市場変動に対して説明力があるのか検証した上で、実際に検出されたトピックを例示する。

5.1 トピックモデルの比較

本節では複数のトピックモデルについて金融ヘッドラインデータでの性能を比較する。比較したモデルは LDA,NMF,CTM,Top2Vec,BERTopic-FinBERT,BERTopic の 6 種類である。BERTopic-FinBERT については BERTopic のプロセスのうち 1 番の分散表現の抽出に、FinBERT [17] の [cls] トークンの出力に相当するベクトルを利用した。6 種類の手法の中で、CTM,Top2Vec,2 種の BERTopic が文章の分散表現を利用した比較的新規の手法である。クラスタリングの品質はトピック毎に抽出した代表単語の TC と TD で計測した。すべてのモデルでトピック数をハイパーパラメータとして与えることができる。本実験では 10,20,30,40,50 と 5 種類のトピック数について乱数シードを変えながら 3 回実験を繰り返し、全 15 回の試行の精度指標の平均を報告している。

5.2 市場変動の回帰モデル

前出の BERTopic を使用して抽出したトピックが市場の変動に対して説明力があるか、シンプルな回帰モデルを構築して検証する。

$$r_{t,t-1} = \beta_1 r_{t-1,t-2} + \beta_2 r_{t-1,t-6} + \beta_3 r_{t-1,t-11} + \sum_{i=1}^{K} \beta_{i+3} sentiment_{i,t} \quad (2)$$

$r_{t,t-i}$ は $t-i$ 時点の引け値から t 時点の引け値までの価格リターンであり、最初の 3 項は前日までの 1 日、5 日、10 日リターンを説明変数として使用しており、$sentiment_{i,t}$ はトピック i の t 時点でのセンチメントである。使用するトピック数 K を 0 とした場合と 0 より大きな数としたときにモデルの説明力に違いがあるかを検証する。回帰は Lasso 回帰として行い、制約項の係数はハイパーパラメータとして 5fold のクロスバリデーションにより決定する。精度指標としては R^2 値、平均絶対誤差、平均 2 乗誤差を測定する。モデルはデータを 3 ヶ月毎に区分し、それぞれの期間でトピックの抽出と、回帰モデルの推定を行う。ヘッドラインの取得時間が UTC21 時以降の場合には翌日のニュースとして取り扱う、本稿では市場指数として S&P500 指数を使用しており、米国での取引が活発となる時間を意識して時間を区分した。ヘッドラインのセンチメントの計測には FinBERT を利用する。FinBERT では入力文章の極性が positive,neutral,negative の 3 クラスに判定される共に、クラスに属する可能性をスコアとして出力する。得られたセンチメントが positive の場合にはスコアをセンチメントスコアとして利用し、neutral の場合には 0、negative の場合にはスコアに − 1 を乗じてヘッドラインのセンチメントスコアとした。各トピック i の日次センチメントスコアは、時点 t で得られたヘッドラインのうち、当該トピックに属するヘッドラインのセンチメントスコアの平均として計算する。

表 1 モデル毎のトピックの品質

	TC	TD
LDA	-0.12153	0.73961
NMF	-0.03005	0.60773
CTM	-0.08247	0.92868
Top2Vec	-0.55158	0.84149
BERTopic-FinBERT	0.01972	0.80226
BERTopic	0.04179	0.85132

6 実験結果

6.1 トピックモデルの比較

各モデルにより抽出されたトピックの精度指標を表 1 にまとめた。TC については BERTopic の精度が最も良好で次いで BERTopic-FinBERT が良好であった。LDA や NMF の様な BoW をベースとした手法と比較してトピックの精度が高くなっているのは [11] の報告と傾向が一致しており、文章のコンテキストを活用することが得られたトピックの精度に有利に働く可能性があるという点は金融ヘッドラインデータに関しても同様であった。TC でみた精度が最も低いのは Top2Vec であり、複雑な潜在ベクトルの分布に対して、クラスターのセントロイドを使用したトピックの抽出が [11] の指摘通り有効に機能していない可能性がある。最後に BERTopic-FinBERT と BERTopic の比較では、直観的には金融コーパスでのファインチューニングが行われた FinBERT と使用することで精度が向上することが期待されたが、結果としては Sentence-BERT を使用したモデルの精度が比較的良好だった。文全体の分散表現を獲

表 2　回帰モデルの精度指標

	R^2	MAE	MSE
$K = 0$	0.02661	0.00815	0.00016
$K = 50$	0.27649	0.00694	0.00012
$K = 100$	0.22507	0.00727	0.00013

表 3　各トピックの回帰係数と代表単語

	topic			
	1	2	3	4
coef	0.032	0.01052	0.00962	0.00124
word0	500	tanker	ukraine	treasuries
word1	sp	lng	russian	orders
word2	nasdaq	oil	kyiv	ecb
word3	100	meg	ukrainian	curve
word4	futures	store	zelenskiy	block
word5	600	sign	troops	curves
word6	index	pictures	russia	goods
word7	leads	iea	captured	italian
word8	decline	dates	war	factory
word9	tech	hook	soldier	extend

得するように訓練した Sentence-BERT を利用することがヘッドラインのクラスター作成には有利だった可能性があり、金融コーパスで文章全体の分散表現を獲得するように事前学習したモデルを使用することでさらなる精度の向上を望めることから今後の課題として記載する。

6.2　マーケット変動要因の抽出

6.2.1　精度指標

金融ヘッドラインデータを用いて、式 2 のモデルを推定し、モデルの精度指標を表 2 にまとめた。使用するトピック数は K を 0,50,100 と 3 種類実験している。表中の数字は、データ期間を 3 ヶ月毎の区分に分けた 18 セットでそれぞれモデルを推定した結果の平均値である。K を 0 として、過去の価格変動だけで市場変動をモデリングしたケースではマーケット変動に対する説明力はほどんどなく R^2 値も 0 に近い値になっている、使用するトピックの数を増やすと R^2 の値は大きくなり、2 つの誤差指標も $K = 0$ の場合と比べて低下している。トピックのセンチメントがマーケットの変動の少なくとも一部を説明していると考えられる。

6.2.2　抽出されたトピック

次に、実際に推定されたモデルと抽出された単語の具体例を示す。表 3 に 2022 年 4 月 1 日から 6 月 30 日のデータを使用したモデルについて一部のトピックセンチメントに対する回帰係数と各トピックの代表単語の上位 10 単語を示す。

推定されたモデルでは 24 個の説明変数の係数が 0 と異なった、うち値が正であったものは 7 個も含め、全てがトピックセンチメントであった。表 3 では紙面に収めるために 4 つのトピックのみについて記載する。1 番目のトピックは SP500 や NASDAQ などの単語が並び主に米国の株式市場に関するニュースと思われる。当日の金融市場の動向をタイムリーにニュースとして配信するベンダーが存在し、これらのニュースがトピックとして抽出されている。市場動向ニュースを事後的に抽出することに実務的な意味があるのかは議論の余地があるが、その他のトピックにも特徴的な単語が並ぶ。3 番目のトピックでは”ukraine”、”russia”などの単語が並びロシアのウクライナ侵攻の情勢に金融市場が敏感になっていたことが示される。4 番目のトピックには”ecb”,”treasuries”, ”curve”、などの金利市場に関わる単語が並び、各国中央銀行の動向や金利市場のニュースが株式市場にも影響を与えていた可能性が示唆される。

7　結論

本稿では金融アナリストが過去の市場変動の要因となった事象を調べる際に、ニューステキストとトピックモデルを使用して効率的に市場変動の要因となったトピックを抽出する手法を提案し実証実験を行った。実験は 2 段階に分けて行われ、1 つ目の比較実験では金融ニューステキストに複数のトピックモデルを適用し、トランスフォーマーベースの手法が LDA などの従来手法と比べてトピック抽出のパフォーマンスが優位であることを明らかにした。2 つ目の実験では BERTopic を使用して実際にトピックを抽出し、回帰分析を通じて市場の変動を説明するトピックが検出できるか検証した。トピックを説明変数として使用したモデルはそうでないモデルと比べて R^2 などでみて精度が向上しており、市場の変動を説明するトピックが検出できていると考える。実際に検出されたトピックも人間の直観にあう内容となっており本稿で提案した手法に一定の有用性があると考える。

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. **J. Mach. Learn. Res.**, Vol. 3, No. null, pp. 993–1022, mar 2003.

[2] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. **Neural Computation**, Vol. 23, No. 9, pp. 2421–2456, 2011.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[5] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 759–766, Online, August 2021. Association for Computational Linguistics.

[6] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, **Proceedings of the 31st International Conference on Machine Learning**, Vol. 32 of **Proceedings of Machine Learning Research**, pp. 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[7] Dimo Angelov. Top2vec: Distributed representations of topics. **CoRR**, Vol. abs/2008.09470, , 2020.

[8] Paul Glasserman, Kriste Krstovski, Paul Laliberte, and Harry Mamaysky. Choosing news topics to explain stock market returns. In **Proceedings of the First ACM International Conference on AI in Finance**, ICAIF '20, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Thien Hai Nguyen and Kiyoaki Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1354–1364, Beijing, China, July 2015. Association for Computational Linguistics.

[10] Jaydeep Soni Ivalio Dimov. Trading esg news using topic codes factors, 2021.

[11] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. 2022.

[12] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[13] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In **2017 IEEE International Conference on Data Mining Workshops (ICDMW)**. IEEE, nov 2017.

[14] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. **Proceedings of GSCL**, Vol. 30, pp. 31–40, 2009.

[15] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 530–539, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[16] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 439–453, 2020.

[17] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.