新興市場を対象とした市況情報の抽出

Extraction of market analysis information for emerging markets

神田 裕輝[1]　高野 海斗[1]　酒井 浩之[1]　北島 良三[2]　中川 慧[3]

Kaito Takano[1]　Yuki Kanda[1]　Hiroyuki Sakai[1]　Ryozo Kitajima[2]　Kei Nakagawa[3]

[1]成蹊大学
[1]Seikei University

[2]東京工芸大学
[2] Tokyo Polytechnic University

[3]野村アセットマネジメント株式会社
[3] Nomura Asset Management Co., Ltd.

Abstract: In this study, we propose a method to extract market information of emerging markets from newspaper articles. Here, we use Bloomberg articles as an information source to extract market information of emerging markets where is expanding in recent years. The extracted market information is useful as reference information when creating a market analysis report for the target emerging market. However, there are many articles that describe market conditions in large markets such as the Nikkei Stock Average and the Dow Jones Industrial Average, but there are few articles that mention one emerging market, such as India, China and Taiwan. It is only described in one part of an article that mentions several emerging markets. Therefore, in this study, we extract market information about one target emerging market from articles where information about market conditions of multiple emerging markets is mixed in one article. Furthermore, we select important sentences from the sentences extracted using the topic model and create a monthly report.

1　はじめに

近年，証券市場における個人投資家の比重が増加しており，個人投資家に対して投資判断の支援をおこなう技術の必要性が高まっている．その一例として，経済新聞記事をテキストマイニングの技術を用いて解析し，経済市場を分析する研究などが行われている[1][2][3]．また、投資判断のための情報源としてファンドの運用報告書に記載されるマーケットレポートが用いられる．マーケットレポートには世界経済や金融市場の動き等が記載されており、それらはファンドの運用担当者が、株価が大きく動いた日を調べ、その前後に起きたイベントを確認し、内容をまとめている．しかし、この作業は運用担当者にとって大きな負担となっている．したがって、経済テキストを機械的に解析し、マーケットレポートを自動で作成することができれば、ファンドの運用担当者の負担を減らすことができると考えられる．

本研究では，Bloomberg 記事を情報源として，近年市場が拡大しつつある新興市場を対象に分析を行い，市況情報を抽出する手法を提案する．本研究で抽出したい市況情報の例を以下に示す．

> ４日のインド株式市場では指標のＳ＆Ｐセンセックス指数が４営業日続落。不良債権が増えることへの懸念から銀行株が売られた。アクシス銀行はここ２日での下げが１年４カ月で最大となった。発電設備メーカーのバーラト重電機は１カ月ぶり大幅安。…..

関連研究として，経済テキストを用いて市況分析コメントを生成する研究がある．酒井らは日経新聞記事を用いて，日経平均株価について述べた記事やそれに関連のある記事を抽出し，そこから重要キーワード，重要な要因文を推定し，市況分析コメントを生成する手法を提案している[4][5]．しかし文献[4][5]の手法は日経平均について言及している複数の記事からレポートを生成するものであり，また，1 つの記事中で日経平均について言及している記事を対象としている．一方，Bloomberg においては１つの記事中で，例えばインド市場のような１つの新興市場について言及している記事はほぼなく，複数の新興市場の市況について言及している記事の１部に

記述があるだけである．そのため，新興市場の市況情報をまとめた市況分析レポートを作成するには，複数の新興市場の市況について言及している記事を探し，さらに，その中から対象としている新興市場の市況について述べている文を抽出し，それらをまとめる必要がある．従って，日経平均株価やダウ平均株価の市況分析レポートの作成に比べると多くの労力を必要とする．

そこで本研究では Bloomberg 記事を対象として，記事の 1 部のみに対象としている新興市場の記述がある状況下で，複数の記事より，対象新興市場についての有益な情報を抽出することを試みる．図 1 に，本研究において想定しているタスクの概要を示す．

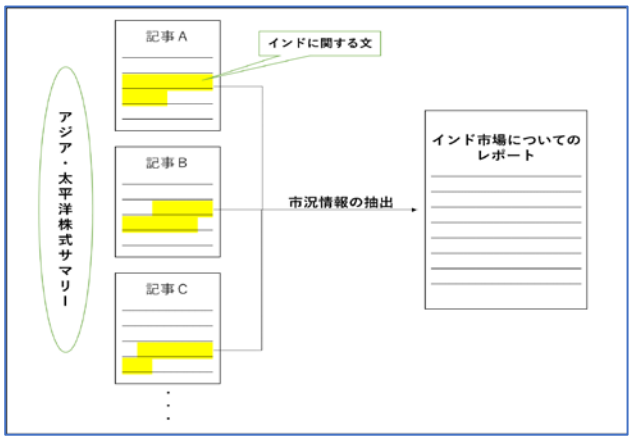

図 1　本研究のタスク概要

本研究では，対象とする新興市場をインド市場と定め，複数の新興市場の市況情報について記述されている「アジア・太平洋株式サマリー」からインド市場の市況情報のみを識別し，抽出する．

2　提案手法

本節では，Bloomberg 記事より対象の新興市場の株価に影響を与える情報を取得する手法の説明を行う．本研究では対象とする新興市場を「インド」と定めて研究を行った．

2.1　手法概要

複数の新興市場の市況情報についてまとめて記述されている「アジア・太平洋株式サマリー」には，インド市場以外にも中国市場や韓国市場，台湾市場といった複数の市場についての市況状況がある．そのため，文単位でインド市場の市況について述べている文であるかどうかの分類をする必要がある．手法の概要を以下に示す．

Step1：Bloomberg 記事から複数の新興市場について言及している記事（「アジア・太平洋株式サマリー」）を抽出．

Step2：抽出した記事から，最近傍法のための学習用データとして，対象としている新興市場（本研究ではインド市場）の市況について述べた文を抽出する．.

Step3：取得したデータから語の TF･IDF 値を計算し，語を要素，TF･IDF 値を要素値とした単語ベクトルを生成する

Step4: 学習用データとテスト用データとの文間類似度を求め，最近傍法にてインド市場の市況情報について述べている文であるかどうかを分類する

Step5：抽出した文を時系列順に並べ，市況情報をまとめる．

2.2　新興市場に関する記事の抽出

本研究では，Bloomberg 記事の 2015 年 1 月から 2017 年 9 月の約 3 年分の記事を用い，使用する記事数は新規記事 4,338,113 記事である．

新興市場（インド市場）について言及している記事を抽出するために，記事本文に対象とする国名と経済や市場に関する単語を含む記事を抽出した．内容を確認したところ，その多くが「アジア・太平洋株式サマリー」の記事であった．そのため，特定の市場（インド市場）に限定したレポートを生成するには，この記事から特定の市場の市況情報について言及した文のみを抽出する必要がある．そこで，この記事中の“インド株式市況”の前後に注目して，重要な文の抽出を試みる．なお，抽出された「アジア・太平洋株式サマリー」は 712 記事である．

2.3　学習データの作成

前節で抽出された「アジア･太平洋株式サマリー」712 記事を用いて，インド市場の市況情報に関する文を抽出するために，最近傍法により文をインド市場に関する文かそうでないかを判定する．まず，学習用データとテスト用データを作成する．ここで，“インド株式市況”を含む文とその次の文は，インドについて言及していると想定し学習用データの正例とする．この時に明らかに他の国の市場について言及されている場合は正例データから除くこととする．一方で，“インド株式市況”を含む文のそれ以前の文はインド以外について言及していると想定し学習用データの負例とする．以下に学習データの例を示す．

・正例

・【インド株式市況】 6 日のインド株式相場は 6 営業日続落。タタ・モーターズの利益が市場予想に届かず、企業業績に照らした株価水準に警戒感が広がった。
・【インド株式市況】21 日のインド株式相場は反落。タタ・スチールの四半期赤字が予想を上回る規模だったことから、金属株が売られた。

・負例

・【中国・香港株式市況】中国本土の株式相場は 3 営業日続落。
・原油安で利益が押し上げられ、消費者の需要も高まるとの見方が材料視された。
・ペトロチャイナ（601857CH）は 6 %高と、4 年ぶりの高値を付け、上海総合指数の上昇に大きく寄与した。

テスト用データは正例とした 2 文以降の文とする．この時抽出した文は，正例が 1403 文，負例が 9696 文，テスト用データが 6896 文である．また，以降の市況情報の抽出の計算において文のつながりを考慮し，正例の 2 文を繋げ 1 つの文として想定することにより，正例を 706 文とする．

2.4 市況情報の抽出

単語の出現頻度に基づいた重みを計算するために式(1)より TF･IDF 値を算出する．使用する単語は半角の単語を除いた 2 語以上の単語とする．

$$TF \cdot IDF(t,d) = TF(t,d) \cdot \log\left(\frac{|N|}{df(t,N)}\right) \quad (1)$$

N：学習用データの文の総数（今回は 10402 文）
$TF(t,d)$：文$d \in N$において，単語 t の出現頻度
$df(t,N)$：文集合 N のうち単語 t を含む文の数

算出した TF･IDF 値をベクトルの要素値として，式(2)よりベクトル間の余弦を求め，文間類似度 sim を計算する．

$$\mathrm{sim}(V_d, V_t) = \frac{V_d \cdot V_t}{\|V_d\| \cdot \|V_t\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (2)$$

ここで，V_d, V_tはそれぞれ学習用データ，テスト用データに含まれる単語を要素とした単語ベクトル（$V_d = \{x_1, x_2, \cdots, x_n\}$, $V_t = \{y_1, y_2, \cdots, y_n\}$），$x_i, y_i$はそれぞれ学習用データ，テスト用データに含まれる単語の TF･IDF 値を表す．

テストデータの文と学習データとの文間類似度を求め，学習データにおける最も類似度が高い文が正例であったのであれば，その文をインド市場の市況情報についての文であるとして抽出する．全テスト用データ 6896 文を最近傍法で判定した結果，インド市場の市況情報に関わる文として抽出できたのは 2132 文であった．以下に，抽出された文の一部を示す．

・世界最大の製油所を所有するリライアンス・インダストリーズは 10 カ月ぶり安値を付けた。
・インド石油ガス公社（ＯＮＧＣ）は構成銘柄の中で最もきつい値下がり。
・インド最大の民間銀行、ＩＣＩＣＩ銀行は続落。
・タタ・スチールが大きく下げ、金属株指数は 3 週間ぶりの低水準に落ち込んだ。

2.5 評価

テスト用データの 100 文をインドに関する文かそうでないかを最近傍法によって判定した．テスト用データ 100 文を，正例・負例を含めた学習用データ 10402 文とのそれぞれの文間類似度を算出し，その中で最も値が大きい文と同じ分類にすることとした．ここで，文中に“インド”を含むものは，文間類似度の値によらずインド市場に関係すると判定する．分類の内容としてインド市場に関係すると判定した場合 1，インド市場とは関係ないと判定した場合は-1 とする．また，それらの分類分けが正しいかどうかを人手にて○×で評価した．表 1 に判定結果の例を示す．

表 1. 文間類似度の判定結果の例

	テストデータの文	分類	評価
①	英高級車メーカーのジャガー・ランドローバーを傘下に置くタタ・モーターズは３日続伸	1	○
②	政府の１日声明によれば，道路建設に充てる資金捻出のため燃料税がリットル当たり２ルピー引き上げられた	−1	×

③	【韓国株式市況】韓国総合株価指数は前日比1.1%高の1904.65	−1	○
④	銀行株と公益事業株の上げが目立った	1	×

この評価より，最近傍法による全体の精度は 76%であることがわかった．

2.6 レポートの作成

最近傍法により抽出した 2132 文と，その文を含む記事における正例の文を繋げ，それを時系列順に並べて，月ごとのインド市場の市況情報のレポートを作成する．以下に例として抽出した 2015 年 2 月の市況情報の一部を示す．正例データの文と，最近傍法により抽出された市況情報の文を組み合わせることによって，繋がりのあるわかりやすい文章になっている．

２日のインド株式市場では指標のＳ＆Ｐセンセックス指数が前週末に続き下落。消費関連銘柄などが売られた。インド準備銀行（中央銀行）は３日に金融政策決定会合を開く。たばこのＩＴＣは続落。センセックス指数は前週末比 0.2%安の 20122.27 で終了。インド中銀の３日の金融政策決定を前にブルームバーグ・ニュースがまとめたエコノミスト調査では、41 人中 31 人が 7.75%での政策金利据え置きを予想。４日のインド株式市場では指標のＳ＆Ｐセンセックス指数が４営業日続落。不良債権が増えることへの懸念から銀行株が売られた。アクシス銀行はここ２日での下げが１年４カ月で最大となった。発電設備メーカーのバーラト重電機は１カ月ぶり大幅安。.....

3 考察

最近傍法による分類であるが，全体の精度は 76%であった．まず，表 1 の①のようにインドに関する文が正しく判定されたものは 47 文中 24 文あり，これらの多くは，文中にインドに関する固有名詞を含んでいたため，正しく判定されたと考えられる．

表 1 の②のようにインドに関わる文を誤って判定したものは 23 文あった．この時，表 1 の例のようにどの国にも共通する単語が多くみられ，インドに関する固有名詞が含まれていたが負例と同じ分類に判定された文もあったため，df 値の取得方法を改める必要があると考えられる．

表 1 の③のようにインドとは関係のない文で正しく判定されたものは 53 文中 52 文あり，これらは例のようにインド以外の国名が含まれることや，インドに関する固有名詞が含まれていないことが多かったため正しく判定できたのだと考えられる．

表 1 の④のようにインドとは関係のない文を誤って判定したものは 1 文あった．これらはどの国にも関係しうる文であったため，判定が困難であったと考えられる．

また今回，最近傍法との比較のため深層学習による市況情報の抽出も試みた．使用した学習用データとテスト用データは，最近傍法の学習データの作成方法と同様である．ただし，各記事の正例の文を繋げ，正例が 706 文，負例が 9696 文，テスト用データが 6896 文とした．深層学習モデルは，Long short-term memory (LSTM) と Multilayer perceptron (MLP) を組み合わせ，図 2 で示すモデルを用いた．

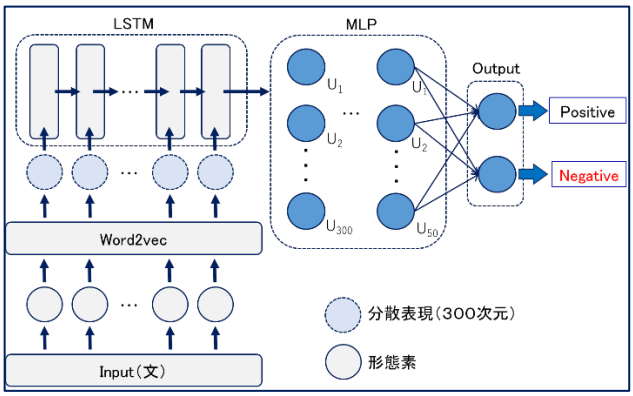

図 2　深層学習モデル

処理の概要であるが，はじめに学習データの文を形態素解析し，Word2vec により形態素の分散表現を得る．それを LSTM に入力し，出力されたベクトルを MLP の入力としてモデルを学習する．MLP の中間層は，ノード数 300 が 3 層，ノード数 150 が 3 層，ノード数 50 が 3 層の計 9 層とした．出力層は 2 要素である．また，epoch 数は 3，中間層の活性化関数を ReLU とし，出力層の活性化関数をシグモイド関数とした．なお，分散表現を得るための Word2vec の学習データとして，10 年分の経済新聞記事を使用した．

全テスト用データ 6896 文を判定した結果，インド市場の市況情報に関わる文として抽出されたのは 315 文でり，また，その精度は 72%であった．この結果は最近傍法の結果である 83%と比べて低いものであり，Bloomberg 記事を対象とした本研究においては最近傍法が有用である結果であった．

さらに手法比較にくわえて，本提案手法を他の新興国市場を対象としても正しく市況情報抽出が可能かどうかを検証するために，対象を「中国・香港市

場」として追加実験を行った．その結果抽出された例を以下に示す．

- 上海総合指数は前営業日に当たる昨年 12 月 31 日と比べ 3.6%高の 3350.52 と、終値ベースで 09 年 8 月 6 日以来の高値を付けた
- ペトロチャイナ（601857CH）は 6 %高と、4 年ぶりの高値を付け、上海総合指数の上昇に大きく寄与した

テスト用データ 100 文を最近傍法により判定し，その結果を人手で評価したところ，精度は 77%であった．この結果から，本手法では高い精度を維持しながら他の新興国を対象としても同様の処理が可能であり，汎用性の高い手法と言える．

4　トピックモデルによるレポートの生成

2 章では最近傍法によって抽出した文と同記事の正例に当たる文を繋げることによって，市況情報をまとめたレポートを作成した．ただし，抽出した文は 2132 文あり，レポートを作成にするには文量が多い．そのため，レポートに採用する文を最近傍法によって抽出した文の中から選別する必要がある．そこで本節では，重要な文の選別方法とレポートの作成方法について説明する．

4.1　トピックモデルの概要

本研究では Python のライブラリのうち gensim の LDA を用いてトピックモデルを生成する．図 3 に本手法の流れを示す．

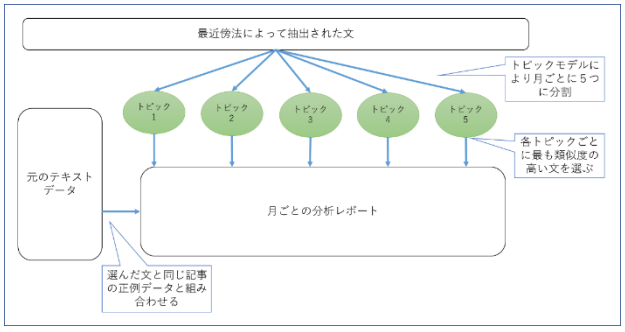

図 3　トピックモデルを用いたレポート生成

抽出した文から重要な文を選別する方法として，まず最近傍法の際に用いた学習用データの文をデータセットとして，gensim によってトピックモデルを生成する．この時のモデル生成では，数字を除く 2 語以上の名詞，動詞，形容詞を対象とし，トピック数を 5 に設定した．これによって似た単語を持つ文で 5 つに分割できる．表 2 に生成されたトピックの例を示す．そして生成されたモデルを最近傍法によって抽出した文に月ごとに適応することで，抽出した文を 5 つのトピックに分割する．次に，各トピックで最近傍法の際に求めた文間類似度が最も高い文をレポートに採用する．つまり，各月で 5 文選ばれることになる．そして，選んだ文と同じ記事の正例データの 1 文（記事のうちインドに言及している最初の文）と繋げ，最大合計 10 文のレポートを作成する．

表 2．生成されたトピックの例

トピック 1		トピック 2		トピック 3	
単語	生起確率	単語	生起確率	単語	生起確率
銘柄	0.028	指数	0.120	株式	0.131
下げ	0.026	中国	0.068	市況	0.052
発表	0.022	セン	0.056	インド	0.052
なっ	0.022	ハン	0.054	相場	0.041
こと	0.018	本土	0.051	香港	0.039
利益	0.016	前日	0.042	市場	0.038

4.2　トピックモデルの評価

最近傍法によって抽出した 2132 文を月ごとにモデルにかけ，5 つのトピックに分割した後，レポートを作成した．表 3 に例として 2015 年 3 月のトピックモデルによる結果を示す．

表 3　2015 年 3 月の各トピックにおける単語の生起確率

トピック 1		トピック 2	
単語	生起確率	単語	生起確率
会社	0.006	インド	0.010
ぶり	0.006	ルピー	0.008
最大	0.005	銀行	0.005
週間	0.005	ドル	0.005
インド	0.005	高値	0.005
セン	0.004	上昇	0.005
大幅	0.004	付け	0.004
セックス	0.004	バラード	0.004
トピック 3		トピック 4	
単語	生起確率	単語	生起確率
セックス	0.017	ぶり	0.009
セン	0.016	大幅	0.004

前日	0.016	セン	0.004
終了	0.010	セックス	0.004
引け	0.005	終了	0.004
銀行	0.005	週間	0.004
ぶり	0.004	HDFC	0.004
インド	0.004	住宅	0.003

トピック 5	
単語	生起確率
メーカー	0.004
続落	0.004
最大手	0.003
営業	0.003
ぶり	0.003
なっ	0.003
電機	0.003
発電	0.003

・トピック 1 に属する文

・指標のＳ＆Ｐ・ＢＳＥセンセックスは週間ベースで 4 週続伸となった
・世界最大の製油所を所有するリライアンス・インダストリーズは 1.4%安
・国内最大のエンジニアリング会社ラーセン・アンド・トゥブロは 2 週間ぶりの大幅下落、インドステイト銀行は 5 営業日続落となった
・インド最大のエネルギー探査会社、インド石油ガス公社（ＯＮＧＣ）は 5 カ月ぶりの大幅上昇

・トピック 2 に属する文

・ルピー建て 2024 年 7 月償還債の利回りは 2 ベーシスポイント（ｂｐ、１ｂｐ＝0.01%）上昇の 7.74%
・４日にはインド準備銀行（中央銀行）による緊急利下げを受けて、初めて 3 万台を付けた
・通貨ルピーとインド国債も値下がりした
・構成銘柄の中で発電設備のバーラト重電機の下落率が最大だった

・トピック 3 に属する文

・インド株の指標であるＳ＆Ｐ・ＢＳＥセンセックスは前週末比 0.3%高の 29459.14 で終了
・センセックスは前日比 0.7%安の 29380.73 で終了
・センセックスは前日比 0.2%高の 29448.95 で終了した
・指標のＳ＆Ｐ・ＢＳＥセンセックスは前営業日比 2.1%安の 28844.78 で終了

・トピック 4 に属する文

・住宅金融でインド最大手のハウジング・デベロップメント・ファイナンス（ＨＤＦＣ）は 2 カ月ぶりの大幅下落
・英高級車メーカーのジャガー・ランドローバーを傘下に置くタタ・モーターズと、エンジニアリング会社ラーセン・アンド・トゥブロも買われた
・住宅金融のハウジング・デベロップメント（ＨＤＦＣ）は 3 日ぶりに値上がり
・英高級車メーカーのジャガー・ランドローバーを傘下に置くタタ・モーターズは 6 週間ぶりの大幅下落

・トピック 5 に属する文

・日用品大手ヒンドゥスタン・ユニリーバは 1 カ月ぶり高値に上昇
・発電設備メーカーのバーラト重電機は 3 営業日ぶりに下落
・一方、契約者数でインド最大手の携帯電話サービス会社ブハルティ・エアテルは続伸した
・携帯電話サービス企業ブハルティ・エアテルを中心に電気通信株が下げた

単語の生起確率と各トピックの文を見ると，トピックごとに特徴があることが分かる．トピック 1 では生起確率の上位に“最大”があり，トピック 1 の文を見るとインドの大手企業について言及している文が多い傾向にある．同様にトピック 2 では通貨や銀行について，トピック 3 ではインド株の指標について，トピック 4 では自動車や住宅金融メーカーについて，トピック 5 ではその他メーカーについて言及している傾向にあった．このようにトピックモデルによって抽出した文を 5 つのトピックに分割した後に，各トピックから 1 文をレポートに採用する文として選ぶことによって，情報が重複しないレポートを生成できる．以下に作成したレポートを示す．

・2015 年 3 月のインドの市況情報のレポート

２日のインド金融市場では株価指数が３営業日続伸。インド株の指標であるＳ＆Ｐ・ＢＳＥセンセックスは前週末比 0.3%高の 29459.14 で終了。
５日のインド株式相場は上昇。指標のＳ＆Ｐ・ＢＳＥセンセックスは週間ベースで４週続伸となった。４日にはインド準備銀行（中央銀行）による緊急利下げを受けて、初めて３万台を付けた。
10 日のインド株式市場では指標のＳ＆Ｐ・ＢＳＥセンセックスが続落し、1 カ月ぶりの安値を付けた。住宅金融でインド最大手のハウジング・デベロップ

メント・ファイナンス（ＨＤＦＣ）は２カ月ぶりの大幅下落。
インド株式市場では指標のＳ＆Ｐ・ＢＳＥセンセックスが下落。発電設備メーカーのバーラト重電機は３営業日ぶりに下落。

下線部が選別した文を示す．選んだ文と同記事の正例の 1 文と繋げることによって読みやすく分かりやすい文章になっている．また，抽出した文は時系列順に表示し、同じ記事の文を抽出した場合はそれらを繋げることで違和感のない内容のレポートになっている．

また，各トピックから文を抽出せずに 1 つのトピックからいくつかの文を抽出すれば，1 つのトピックに関するレポートを生成することもできる．例えば，金融に関するレポートを生成したければ，金融に関するトピックから文を抽出することで可能になると考えられる．

しかし，トピック間で明確に分割できていない文もあり，似た文が異なるトピックに属していることも多い．これを改善するために，トピックモデルを生成した際のデータを見直すことや，特徴のある単語には重み付けする必要があると考えられる．

4 まとめ

本稿では Bloomberg 記事をもとに，複数の記事からインド市場のような 1 つの新興市場の市況情報を抽出する手法について述べた．本手法では新興市場のようなニッチなものは記事の一部に記述されていることに注意し，最近傍法によって必要な情報を抽出した．そしてトピックモデルによって抽出した文から重要な文を選別し，市況情報をまとめたレポートを作成した．本手法の精度は 76%となり，比較手法として行った深層学習の手法よりも高い精度を得ることができた．また，他の新興市場を対象として市況情報を抽出することによって，本手法の有用性・汎用性を確かめられた．

今後の課題として，最近傍法とトピックモデルの精度を向上することや，文章にする上で要因文を推定することが考えられる．

参考文献

[1] 蔵本貴久，和泉潔，吉村忍，石田智也，中島啓浩，松井藤五郎，吉田稔，中川裕志: 新聞記事のテキストマイニングによる長期市場動向の分析，人工知能学会論文誌, Vol. 28, No. 3, pp. 291-296, 2013.

[2] 和泉潔，後藤卓，松井藤五郎: 経済テキスト情報を用いた長期的な市場動向推定，情報処理学会論文誌, Vol. 52, No. 12, pp. 3309-3315, 2011.

[3] 高野海斗，酒井浩之，北島良三: 有価証券報告書からの事業セグメント付与された業績要因文・業績結果の抽出，人工知能学会論文誌, Vol. 34, No. 5, pp1-22, 2019.

[4] 酒井浩之，坂地泰紀，和泉潔，松井藤五郎，入江圭太郎: 関連記事を用いた市況分析コメントの自動生成，第 22 回 金融情報学研究会, pp. 61-66, 2019.

[5] 酒井浩之，坂地泰紀，和泉潔，松井藤五郎，入江圭太郎: 経済テキストからの市況分析コメントの自動生成，第 20 回 金融情報学研究会, pp.44-49, 2018.