レビューデータ分析におけるトピック抽出モデルに感情の要素と文脈を考慮した単語分散表現モデルである BERT を導入したモデルの拡張

「Topic Sentiment Joint Model with Word embeddings BERT」

東北大学　経済学研究科　C0EM1023　長谷川 一旗

2020 年 8 月 30 日

概要

マーケティングにおいて、オンライン上に散見されるレビュー文から、消費者の商品に対する感情や消費者がどのような点に関心を持っているかについて分析することは非常に重要視されている。したがって、テキストから書き手の感情を分析する感情分析やテキストからトピックと呼ばれる書き手の関心事を抽出するトピックモデルといった手法は、近年の主要な研究分野になっており、組み合わせたモデルが登場するなど研究が盛んになっている。加えて、こうした感情分析やトピックモデルに対して、単語の分散表現と呼ばれる今までの Bag-of-words より柔軟な言語表現を導入することで、性能や精度の向上を目指したモデルも存在してきている。しかしながら、これら自然言語処理分野において bag-of-words などに代表される言語表現において、文脈を考慮することのできないという点が問題視されていた。その問題点を解決した BERT という文脈を考慮した単語分散表現の獲得に成功したモデルの登場により、自然言語処理分野では近年大きな変化を遂げている。したがって、本研究では感情分析とトピックモデルを組み合わせたモデルに対して、BERT から得られた文脈を考慮した単語分散表現を導入することで、感情分析の精度やトピック抽出での意味的結束性の向上が見られるかを検証していく。

目次

1 Introduction

高度な情報化社会の成立により、テキストデータが急速に蓄積されつつある現代社会において、テキストから隠れた知見を発見するためのテキストマイニングは非常に多くの分野で活用されている。

マーケティング分野においては、E コマース事業やソーシャルメディアの急速な発展に伴って、オンライン上に散見されるレビューを分析し、消費者の感情やトピックを抽出し理解することが、企業にとって非常に重要視されつつある。したがって、テキストから書き手の感情を解析する手法やテキストから書き手の関心ごとであるトピックを抽出するトピックモデルといった手法は、その重要性から近年の主要な研究分野の一つとなっており、様々な改良モデルが考案されている。Joint Sentiment Topic Model(JST)[1] では、Latent Dirichlet Allocation(LDA)[2] にセンチメント層を導入することで、テキストから感情とトピックを同時に抽出することを可能にし、より実用的なモデルへと拡張している。Gaussian LDA for Topic Models with Word Embeddings[3] では、自然言語処理分野において、分散表現 (Word Embeddings) と呼ばれる単語を高次元のベクトルで表現する技術を LDA に導入することで、意味的に近く解釈のしやすいトピックの抽出が可能となったと同時にモデルの性能の向上も示されている。Joint Sentiment Topic Model(JST) では、学習コーパスの少ない場合にモデル性能の低下することが問題として指摘されており、この問題点の解決するために分散表現を導入した Topic Sentiment Joint Model with Word Embeddings[5] というモデルも存在する。このモデルでは、外部の大規模コーパスを用いた分散表現の導入によって、モデルの性能が大幅に向上したことが示されている。しかしながら、これらのモデルに利用されている分散表現は、Word2Vec[6] という文脈を考慮していない分散表現である。文脈を考慮しない分散表現では、異なる意味合いを持つ単語に対してそれぞれの意味を区別できない分散表現になってしまうため、多義語のある文章に対して十分な性能を発揮できないという問題点が指摘されている。近年の自然言語処理分野では、こうした問題点に対応するために深層学習を利用することで文脈を考慮した単語分散表現の獲得を目指す研究が盛んに行われており、ELMo[7] や BERT[8] といったモデルは既存の分散表現モデルと比較して、高い性能を誇ることが知られている。よって、既存の分散表現である Word2Vec を利用している様々なタスクやモデルにおいて、こうした文脈を考慮した分散表現を置換してあげることにより性能の向上が見込まれるが、それらを検証した論文は多くは出てきていません。

したがって、本稿では分散表現を利用したセンチメントとトピックの同時抽出モデルである Topic Sentiment Joint Model with Word Embeddings(TSWE) に導入する分散表現として、文脈考慮型の分散表現を導入したモデルを提案し、その有効性について実験を通して検証を行っていく。

2 Related Researches

2.1 Latent Dirichlet Allocation(LDA)

LDA は、Blei ら [2] によって提案された文書の確率的生成モデルである。LDA では、文書には複数のトピックが存在すると仮定し、文書中に含まれる単語にトピックを割り当てる。トピックは、文書から観測できないため、文書内の単語の共起情報からトピックを推定する。図 1 (a) に示す LDA の生成過程をまとめると以下の通りである。

1. for $k = 1$ to K
   (a) Draw word distribution ϕ_k $\sim$Dir(β)
2. For each document d
   (a) Draw a topic distribution θ_d $\sim$Dir(α)
   (b) For each word index n from 1 to N_d
      i. Draw a topic $z_{dn} \sim$ Multi(θ_d)
      ii. Draw a word $w_{dn} \sim$ Multi($\phi_{z_{dn}}$)

LDA は、自然言語処理分野だけでなく多種多様な分野に応用可能なモデルであり、改良も容易であることから様々な改良モデルが存在する。その中でも、

センチメント層を導入することで感情推定とトピック推定を同時に行うことを可能にしたモデルである JST[1] モデルを 2.2 節、分散表現を導入することでその性能を向上させただけでなく、未知語に対しても対応可能になっている Gaussian LDA[3] モデルを 2.3 節で紹介する。2.4 節では、本稿での提案手法の基礎にもなった JST モデルに分散表現を導入することで、より性能を向上させた TSWE[5] モデルの紹介を行う。

2.2 Joint Sentiment/Topic Model for Sentiment Analysis(JST)

JST は、Lin ら [1] によって提案された文書からセンチメントとトピックの抽出を目的としたモデルである。JST では、文書単位での感情分類を達成するために、文書層とトピック層との間にセンチメント層を導入することで通常の LDA モデルを拡張した。JST では、従来の LDA と異なり、S 個のセンチメントラベルが文書と紐づけられており、そのもとでトピックがセンチメントラベルに紐づけられている。したがって、単語はセンチメントラベルとトピックの両方に紐づけられることになる。図 1 (b) に示す JST の生成過程をまとめると以下の通りである。

1. For each topic-sentiment pair (k,l)
   (a) Draw word distribution $\phi_{k,l} \sim \mathrm{Dir}(\beta)$
2. For each document d
   (a) Draw a sentiment distribution $\pi_d \sim \mathrm{Dir}(\gamma)$
   (b) For $l = 1$ to S
      i. Draw a topic distribution $\theta_{d,l} \sim \mathrm{Dir}(\alpha)$
   (c) For each word index n from 1 to N_d
      i. Draw a sentiment label $l_{dn} \sim \pi_d$
      ii. Draw a topic $z_{dn} \sim \theta_{d,l_{dn}}$
      iii. Draw a word $w_{dn} \sim \phi_{z_{dn},l_{dn}}$

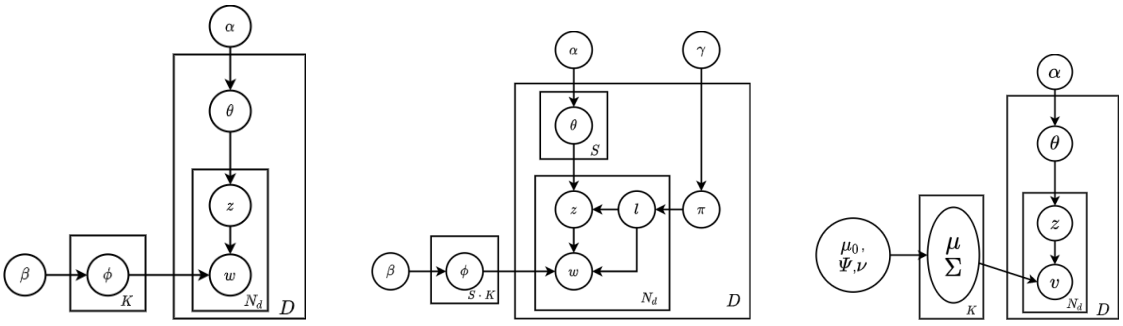
映画のレビューデータセットを利用した評価実験では、他のラベル付きアプローチと比較して、文書レベルでの感情分類において非常に優れた性能を示している。加えて、抽出されたトピックからは確かに一貫したトピックとなっていることも示している。しかしながら、このモデルでは単語を Bag-of-words として表現しているため、単語の順序を考慮出来ていないということが指摘されている。

2.3 Gaussian LDA for Topic Models with Word Embeddings

分散表現学習の登場により、トピックモデルと分散表現を組み合わせ、より意味的に一貫したトピックが生成されやすくなることを目指した研究が、Gaussian LDA[3] である。このモデルは、Hu ら [4] が提案した LDA におけるトピックを生成する分布を多次元ガウス分布にするというモデルに単語の分散表現を組み合わせたもので、Das ら [3] によって提案された。連続空間上に埋め込まれた単語ベクトルに対して、トピック k を同一空間上での多次元ガウス分布とした。これによって、トピック毎の単語分布が連続分布となり、この分布から単語ベクトルが生成される過程がモデル化されている。図 1 (c) に示す Gaussian LDA の生成過程をまとめると以下の通りである。

1. for $k = 1$ to K
   (a) Draw topic covariance $\mathbf{\Sigma_k} \sim \mathcal{W^{-1}}(\mathbf{\Psi}, \nu)$
   (b) Draw topic mean $\mu_\mathbf{k} \sim \mathcal{N}(\mu, \frac{\mathbf{1}}{\kappa}\mathbf{\Sigma_k})$
2. For each document d
   (a) Draw a topic distribution $\theta_d \sim$Dir(α)
   (b) For each word index n from 1 to N_d
      i. Draw a topic $z_{dn} \sim$ Categorical(θ_d)
      ii. Draw $v_{dn} \sim \mathcal{N}(\mu_{\mathbf{z_{dn}}}, \mathbf{\Sigma_{z_{dn}}})$

このモデルでは、従来の LDA と異なり単語を Bag-og-words として表現するのではなく、分散表現を用いて表現することで、分散表現が類似しているつまり意味的に近しいと思われる単語を同じトピックに割り当てやすくなり、トピック内の意味的統一性が向上し、自己相互情報量 (PMI) も上昇することが報告されている。しかしながら、このモデルで利用されている分散表現は Word2Vec[6] と呼ばれる手



(a)LDA (b)JST (c)Gaussian LDA

図 1　グラフィカルモデル

法を用いて作成されたものである。この手法では、Bag-of-words と同じく単語の順序を考慮できていないため、文脈を考慮した結果にならず多義語などの場合に推定がうまくできない可能性が残されている。

2.4 Topic and Sentiment Model with Word Embeddings(TSWE)

TSWE は、JST に外部の大規模コーパスから学習させた分散表現を導入した Joint Topic Sentiment Model で、Fu ら [5] によって提案されたモデルである。TSWE では、学習コーパスが短く小さい場合に、それまでの Topic Sentiment Joint Model で問題視されていた、単語の共起情報を使用することによって生じる分布の推定がうまくいかない問題を単語の分散表現を導入することにより解決を試みた研究である。TSWE では、JST におけるディリクレ多項式成分を Sentiment-Topic-Word のディリクレ多項式成分と Word Embeddings 成分の混合成分に置換している。このモデルでは、Word Embeddings 成分から単語を生成する確率を以下のように定めている。

$$MulT(w_i|\nu_k\omega^{\mathrm{T}}) = \frac{\exp(\nu_k\omega_{w_i})}{\Sigma_{w_i'\in W}\exp(\nu_k\omega_{w_i'})} \quad (1)$$

図 2 に示す TSWE の生成過程をまとめると以下の通りである。

1. For each topic-sentiment pair (l,k)
   (a) Genearte the word distribution of the sentiment-topic pair $\phi_{l,k} \sim \mathrm{Dir}(\beta)$
2. For each document d
   (a) Draw a distribution $\pi_d \sim \mathrm{Dir}(\gamma)$
   (b) For $l = 1$ to S under document d
      i. Draw a topic distribution $\theta_{d,l} \sim \mathrm{Dir}(\alpha)$
   (c) For each word index n from 1 to N_d
      i. Draw a sentiment label $l_{dn} \sim \mathrm{Multi}(\pi_d)$
      ii. Draw a topic $z_{dn} \sim \mathrm{Multi}(\theta_{d,l_{dn}})$
      iii. Draw a binary indicator variable $s_{dn} \sim \mathrm{Ber}(\lambda)$
      iv. Draw a word $w_{dn} \sim (1 - s_{dn})\mathrm{Multi}(\phi_{z_{dn},l_{dn}}) + (1 - s_{dn})MulT(\nu_{z_{dn}}\omega^{\mathrm{T}})$

しかしながら、このモデルでも Gaussian LDA[3] 同様、分散表現として Word2Vec[6] で学習させた分散表現を用いており、依然として文脈を考慮できないという問題点が残されている。本研究では、この問題点を解決するために利用する分散表現を BERT[8] と呼ばれる手法で学習させた文脈を考慮できる分散表現を用いて、その性能の向上を図る。こうした点を踏まえ、次節以降では、単語分散表現を獲得する手法として有名な Word2Vec[6] や文脈考慮型分散表現の作成をはじめて可能にした ELMo[7]、その ELMo の性能を大幅に向上させただけでなく、汎化性能をも高めた BERT[8] について基本的な概念を説明していく。

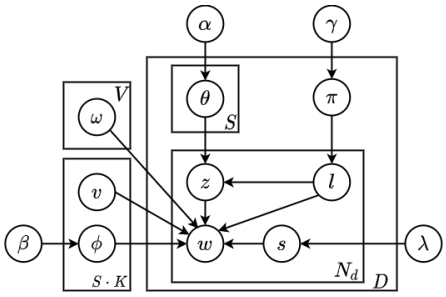


図 2　TSWE のグラフィカルモデル

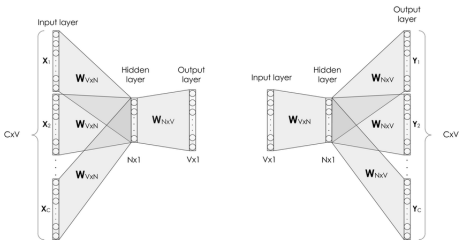


図 3　Word2Vec における CBOW モデル (左) と Skip-gram モデル (右)[*1]

2.5　Word2Vec

Word2Vec は、Mikolov ら [6] が提案したニューラルネットワークを用いて、単語の分布表現を獲得する手法である。つまりは単語を低次元の密なベクトルで表現したものを学習する手法である。従来、自然言語処理のタスクではテキストデータを扱う際に、bag-of-words や LSI[9]、pLSI[10] で圧縮したベクトルを用いてきた。しかしながら、これらの手法は次元数が高次元になってしまい計算効率が悪くなってしまう問題点や圧縮した際の精度的な問題点が残されてしまっていた。こうした問題を解決するため、Mikolov らは単語の意味は単語の周辺の単語によって決定されるという分布仮説の下、テキスト中の各単語を周辺単語から予測するというタスクを設定し、このタスクを大規模なテキストデータからニューラルネットワークによって学習させることで各単語に対する概念ベクトルを獲得した。Word2Vec には、図 3 左で示される、周辺の単語から対象とする単語が現れる確率を最大するように学習させる、Countinous Bag-of-Words(CBOW) と呼ばれる手法と図 3 右で示される、対象の単語を入力とした際の周辺の単語予測のエラー率が最小になるように学習させる、Skip-gram という手法が存在するが、どちらにせよ最終的に単語の分布表現が生成される。これにより、単語を意味空間上に対応させることができ、意味的に近い単語の分類や単語を意味的に計算することが可能になった。

しかしながら、この手法では対象とする単語とその周辺にどのような単語が存在するかという点しか考慮できず、語順によって異なる意味に捉えられるような単語をうまく表現できないという問題点がある。

コンピュータの計算能力の向上に伴い、深層学習という多層なニューラルネットワークの学習が容易になったことを受け、自然言語処理分野でも凄まじい発展が見られた。次節で紹介する深層学習を用いた分散表現を獲得する手法を紹介する。

2.6　ELMo

ELMo は、Matthew ら [7] によって提案された、深層学習を用いることで文脈を考慮した単語の分散表現を獲得する手法である。ELMo は、LSTM[12] と呼ばれる時系列データを学習することのできるリカレント・ニューラルネットワーク (RNN) の一種を用いた、多層の双方向 LSTM による単語レベルの言語モデルである。Word2Vec と同様に対象単語の予測確率が高くなるように学習させる点は変わらないが、文頭から対象単語までの順方向と文末から対象単語までの逆方向という双方向の情報を用いて予測させている点が ELMo の特徴になっている。これにより、今まで考慮出来ていなかった文脈を踏まえた分散表現の獲得に成功したのである。加えて、ELMo では、双方向型言語モデルの中間層によって単語を表現していることも注目すべき点である。

しかしながら、厳密な双方向言語モデルにおいては、予測する単語の先の単語、例えば順方向の場合は対象単語から文末までの単語を事前に知っている

[*1] Tutubalina ら [11] より引用

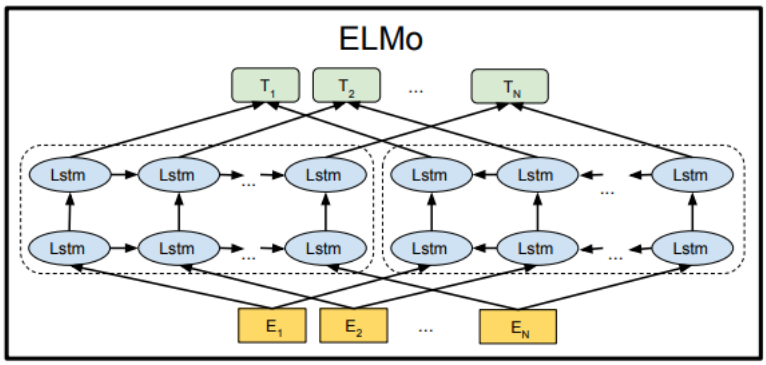


図 4　ELMo

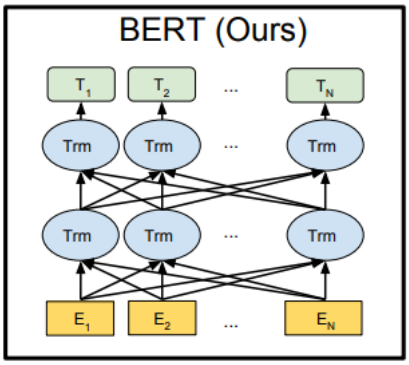


図 5　BERT

ことになる。これは、予測においては事前に答えを知っている状態に当たるためうまく予測モデルを作成できないという問題点が生じる。この問題を解決するために、ELMo では順方向の情報を用いたモデルと逆方向の情報を用いたモデルを別々に学習させた後、統合する手法が採用されているが、この別々にモデルを学習させている点が浅い双方向モデルと呼ばれ、実際に文脈を捉えているか疑問視されている点でもある。図 4 にグラフィカルモデルを示す。

2.7　BERT

BERT は、Devlin ら [8] によって提案された、Transformer[13] という深層学習の手法を用いた様々な自然言語処理タスクに応用可能な事前学習モデルである。BERT では、Masked Language Model(MLM) と Next Sentence Prediction(NSP) という二つのタスクを用いて学習させることで、双方向言語モデルにおける学習時に予測対象の先の単語が見えてはいけないという制約を克服した。MLM は、入力文における 15% の単語に対し、確率的に 3 つの処理を施した状態でその処理を行う前の単語を予測させるというタスクである。3 つの処理とは、選択された 15% のうち、80% を [MASK] に置換するマスク変換処理、10% を別単語に置換する処理、そして残りの 10% は何もせずそのままにする処理となっている。これによって、単語レベルでの学習が可能となっている。しかしながら、MLM だけでは文レベルでの学習ができないため、次に紹介する NSP というタスクを用いて、文レベルでも学習を行うことで広範な自然言語処理モデルとして機能している。NSP は、2 つの文を入力として与え、その 2 文が隣り合っているかどうかを当てるタスクである。NSP では、文の片方を 50% の確率で他の文に置換し、それらが隣り合っているか、隣り合っていないかを判別することで文レベルでの学習を可能としている。

以前から自然言語処理タスクにおける精度の向上には、言語モデルによる事前学習が有効である考えられていた。この言語モデルでは、事前学習で得られた分散表現を特徴量として扱う特徴量ベースという手法と事前学習済みのモデルの最後の部分の重みを再学習させることで新しいタスクにも適用可能にするファインチューニングという手法が存在する。ELMo は、特徴量ベースであったためタスクに応用する際にそのタスクごとにアーキテクチャを再定義する必要があった。一方、BERT ではファインチューニングを採用しているため、タスクごとに大きくパラメータを変更する必要が無く応用の幅が広いという点も特徴の一つである。図 5 にグラフィカルモデルを示す。

3 Topic and Sentiment Model with BERT

3.1 Topic and Sentiment Model with BERT

本研究では、文脈を考慮できないという問題点の解決のために、文脈も学習させた分散表現を既存のトピックモデルに導入することで、その性能の向上を目指す。本研究では、マーケティングへの応用という観点からテキストからトピックを抽出することだけでなく、センチメントの抽出も重要だと考え、分散表現を利用したセンチメントとトピックの同時抽出モデルである Topic Sentiment Joint Model with Word Embeddings(TSWE)[5] に、Word2Vec で学習された分散表現ではなく BERT で学習させた文脈を考慮した分散表現を導入することで、その性能の向上を目指す。また、文脈を考慮した分散表現の導入による性能の向上を測るべく文書レベルでの感情分類とトピック抽出による評価実験を行い、性能について検討していく。なお、本研究で用いるパラメータは表 1 に示す。

TSWE における負の対数尤度は、L_k は以下のように設定される。

$$L_k = \mu \parallel \nu_k \parallel_2^2 - \sum_{w_i \in W} N^{k,w_i} \left(\nu_k \omega_{w_i} - \log \left(\sum_{w_i' \in W} \exp(\nu_k \omega_{w_i'}) \right) \right) \tag{2}$$

そして、L-BFGS[14] 手法を適用して、L_k を最小化するトピックベクトル ν_k を導出する。

3.2 Generative process

本研究では、TSWE モデルを基礎とするため、改めてモデルの生成過程について以下に示す。

1. For each topic-sentiment pair (l,k)
   (a) Genearte the word distribution of the sentiment-topic pair $\phi_{l,k} \sim \mathrm{Dir}(\beta)$
2. For each document d
   (a) Draw a distribution $\pi_d \sim \mathrm{Dir}(\gamma)$
   (b) For $l = 1$ to S under document d
      i. Draw a topic distribution $\theta_{d,l} \sim \mathrm{Dir}(\alpha)$
   (c) For each word index n from 1 to N_d
      i. Draw a sentiment label $l_{dn} \sim \mathrm{Multi}(\pi_d)$
      ii. Draw a topic $z_{dn} \sim \mathrm{Multi}(\theta_{d,l_{dn}})$
      iii. Draw a binary indicator variable $s_{dn} \sim \mathrm{Ber}(\lambda)$
      iv. Draw a word $w_{dn} \sim (1 - s_{dn})\mathrm{Multi}(\phi_{z_{dn},l_{dn}}) + (1 - s_{dn})MulT(\nu_{z_{dn}}\omega^{\mathrm{T}})$

なお、$MulT$ の定義は式 (1) に示している。

表 1　パラメータ記号とその説明

パラメータ	説明
D	文書数
d	文書インデックス
N	総単語数
N_d	文書 d に含まれる単語数
V	全文書に現れる単語の種類 (語彙数)
W	文書集合
$w_{dn} = w_i$	文書 d に含まれる n 番目の単語
S	センチメントのラベル数
l	センチメントラベル
K	トピック数
z	割り当てられたトピック
ν	トピックベクトル
ω	単語の分散表現
θ	トピック分布
ϕ	単語分布
π	センチメント分布
α, β, γ	ディリクレ分布のハイパーパラメータ
s	二値指標変数
λ	ベルヌーイ分布のハイパーパラメータ

3.3 Gibbs sampling

事後分布を解析的に求めることは困難であるため、ギブスサンプリングを用いて事後分布を推定していく。サンプリングに必要な条件付確率は以下の通りである。

$$P(z_i = k, l_i = l | w, z^{\neg i}, l^{\neg i}, \alpha, \beta, \gamma, \lambda, \nu, \omega) \propto \left((1-\lambda) \cdot \frac{N^{\neg i}_{l,k,w_i} + \beta}{N^{\neg i}_{l,k} + V\beta} + \lambda \cdot MulT(w_i | \nu_k \omega^{\mathrm{T}}) \right) \cdot \frac{N^{\neg i}_{d,l,k} + \alpha}{N^{\neg i}_{d,l} + K\alpha} \cdot \frac{N^{\neg i}_{d,l} + \gamma}{N^{\neg i}_{d} + S\gamma} \quad (3)$$

また、推定に用いられるパラメータ π, θ, ϕ は、式 (4),(5),(6) で示される。

$$\pi_{d,l} = \frac{N_{d,l} + \gamma}{N_d + S\gamma} \quad (4)$$

$$\theta_{d,l,k} = \frac{N_{d,l,k} + \alpha}{N_{d,l} + K\alpha} \quad (5)$$

$$\phi_{l,k,i} = (1-\lambda) \cdot \frac{N^{\neg i}_{l,k,w_i} + \beta}{N^{\neg i}_{l,k} + V\beta} + \lambda \cdot MulT(w_i | \nu_k \omega^{\mathrm{T}}) \quad (6)$$

4 Experiments

本節では、文脈を考慮した分散表現の導入による性能の向上を測るべく、文書レベルでの感情分類とトピック抽出による評価実験を行い、性能について検討していく。

4.1 Experimental setup

4.1.1 Using Word Embeddings

本研究では、分散表現として Word2Vecと BERT のオープンソースを利用する。これらのモデルは、大規模コーパスに対して学習させる場合、時間とコストが非常にかかるためこれらの事前学習済みのモデルを利用する。

*2 https://code.google.com/archive/p/word2vec/
*3 https://github.com/google-research/bert

4.1.2 Experimental datasets

本研究では、感情分類用のデータセットとして Large Movie Review dataset(IMDb データセット)[15] と呼ばれる大規模な映画レビューのデータセットを利用する。このデータセットは、ポジティブかネガティブかの二値分類用データで、2 万 5 千件の訓練用データと 2 万 5 千件のテスト用データの合計 5 万件のデータで構成されている。データの前処理として、アルファベット以外の文字は削除し、アルファベットはすべて小文字にする処理も行う。加えて、TSWE モデルに倣い、学習済み分散表現に出現しない単語や出現頻度が 2 以下または 15 以上の単語は、ストップワードとして学習には使用しない。また、コピーされたと思われるような繰り返されたレビューについても同様に削除する。

4.1.3 Hyper Parameter Settings

事前分布のハイパーパラメータ α, β, γ は、TSWE モデルに倣い以下のように設定する。

$$\alpha = \frac{50}{K} \quad (7)$$

$$\beta = 0.05 \quad (8)$$

$$\gamma = \frac{0.05A}{S} \quad (9)$$

なお、A は平均的な文書長とする。

4.1.4 Evaluate Metrics

感情分類の評価指標としては、先行研究 [1, 5] で用いられている Accuracy を利用する。Accuracy は、分類問題で一般的に用いられることの多い評価指標である。定義については、以下の式 (10) に示す。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

なお、TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative である。

次に、トピック抽出の評価指標としては、perplexity と自己相互情報量 (PMI) を利用する。perplexity は、言語モデルの評価指標として良く用いられる指標

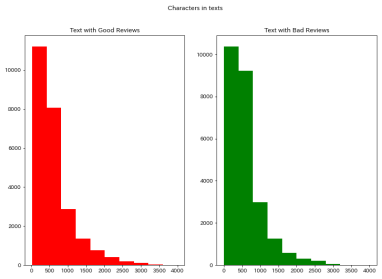


図 6　Number of characters in texts

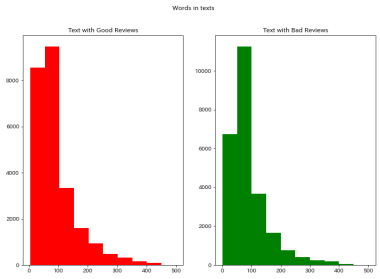


図 7　Number of words in each text

で、言語モデルの複雑さを表す。PMI は、Newman ら [16] によって提案されたトピックの意味的拘束性を表す指標である。perplexity と PMI について、以下の式 (11),(12) で示す。

$$\text{perplexity} = \exp\left(-\frac{\sum_{d=1}^{D}\log p(w_d)}{\sum_{d=1}^{D}N_d}\right) \tag{11}$$

$$\text{PMI}(w_i, w_j) = \log\frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{12}$$

4.2　Experimental Results and Analysis

本来なら、感情分類の結果およびトピック抽出の実験結果を載せ、それらについて考察、検討するところではあるが、実験の進捗が芳しくないためデータの記述統計の結果の図 6 9 に載せ、そちらについて少し言及していく。図 6 は、レビュー文内の文字数についての統計である。これを見るに、0 から 1000 文字以内の比較的短いレビュー文がレビュー全体でも多いことが分かる。全体を通して、Good/Bad に大きな差は見られないと思われる。図 7 は、各レビュー文内の単語数についての統計である。これを見るに、0 から 100 単語以内の比較的短いレビューが多いことが分かる。全体を通して、Good/Bad に大きな差は見られないと思われるが、0 から 100 単語以内のレビューにおいては肯定的のレビューより否定的なレビューのほうが若干単語数が多いことが分かる。これは、否定的な意見のほうが追求したいポイントが明確なためだと思われる。図 8 は、一般的なストップワードを除いた頻出単語の上位 20 単語を示したものである。これを見るに、映画のレビューということもあり、1 位 2 位には、映画に関連する単語が上がっている。しかしながら、その他の単語を見ると何かのトピックに限定されるような単語ではないことが分かる。したがって、ストップワードの調整に関しては非常に重要だと思われる。図 9 は、各テキストにおける平均単語長についての統計である。これを見るに、平均単語長は、6 ぐらいであることがわかる。全体を通して、Good/Bad に大きな差は見られないと思われる。

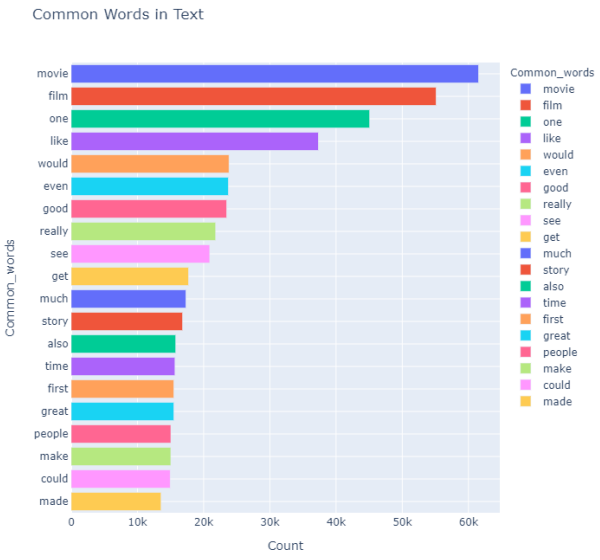


図 8　Common Words in Text

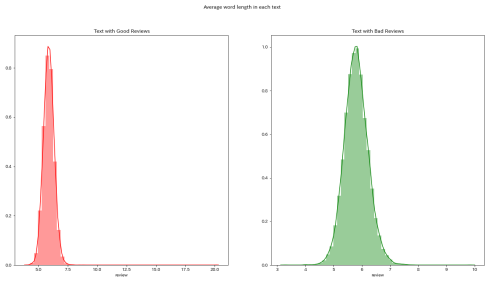


図 9　Average word length in each text

5　Conclusions and Future Work

本研究では、文脈を考慮できていないというこれまでの自然言語処理タスクにおける問題点を解決した分散表現モデルを既存のトピックモデルに導入することでその性能の向上を目的とした以下の実験を行う予定であった。センチメントとトピックの同時抽出モデルに分散表現を取り入れたモデルに対して、文脈が考慮された分散表現を生成できる手法 BERT を導入することで、感情分類の精度向上と、トピック内の意味的拘束性の向上について実験により検証していく予定である。しかしながら、実験コードの作成に少々時間がかかってしまったため、本格的な実験には入ることができなかった。したがって、今後は実際の実験を通してモデルの性能が向上するかどうかを検証していきたい。

参考文献

[1] Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. International Conference on Information and Knowledge Management, Proceedings, 375 - 384.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(4 - 5), 993 - 1022.

[3] Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian LDA for topic models with word embeddings. ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 1, 795 - 804.

[4] Hu, P., Liu, W., Jiang, W., & Yang, Z. (2012). Latent topic model based on Gaussian-LDA for audio retrieval. Communications in Computer and Information Science, 321 CCIS, 556 - 563.

[5] Fu, X., Wu, H., & Cui, L. (2016). Topic sentiment joint model with word embeddings. CEUR Workshop Proceedings, 1646, 41 - 48.

[6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 1 - 9.

[7] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Lin-

guistics: Human Language Technologies - Proceedings of the Conference, 1, 2227 - 2237.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference (Vol. 1).

[9] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211 - 240.

[10] Hofmann, T. (1999). Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, 51(2), 50 - 57.

[11] Tutubalina, Elena & Nikolenko, Sergey. (2017). Demographic Prediction Based on User Reviews about Medications. Computacion y Sistemas. 21. 227-241. 10.13053/CyS-21-2-2736.

[12] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory, Neural Computation, Vol. 9, No. 8, 1735 - 1780

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-December(Nips), 5999 - 6009.

[14] Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. In Mathematical Programming, (Vol. 45), 503-528.

[15] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)

[16] David Newman, Sarvnaz Karimi, and Lawrence Cavedon.(2009) External evaluation of topic models. pages 11-18, December.