

BERT と LDA による口コミデータに基づく評価指標の分析 Evaluation Metric Analysis of Review Data Using BERT and LDA

林 尚也[†] 浦野 昌一[†]
Hayashi Naoya Urano Shoiti

1. はじめに

近年、観光産業は個人旅行や体験型観光の増加、旅行者の年齢層の拡大など、多様化が進んでいる。特にインターネットの普及により、旅行者が宿泊施設や観光地を選ぶ際に口コミの果たす役割が大きくなっている。口コミは、実際の体験に基づいた具体的な情報を提供し、旅行者は自身の価値観やニーズに近い他者の意見を参考にすることで、満足度の高い旅行を計画できるようになっている。本研究では、こうした背景を踏まえ、旅行者の年代による評価基準や価値観の違いを口コミデータから明らかにすることを目的とする。従来はアンケートによる定性的な分析が中心であり、口コミデータを用いた定量的・構造的な分析は十分に行われていない。年齢層ごとの観光ニーズをデータから抽出することは、よりきめ細かな観光サービス設計に資するものである。トピックモデリングは、大量のテキストから潜在的な話題を自動的に抽出する手法であり、自然言語処理分野で広く活用されている。代表的な手法である Latent Dirichlet Allocation (LDA)^[1]は、単語の共起に基づいてトピックを推定するが、文脈を考慮しないため、同じ単語が異なる意味を持つ文脈で用いられていても、同一トピックとして処理されてしまうという課題がある。本研究では、この課題を解決するために、事前に BERT^[2]によって得られた文脈ベースの単語・文書埋め込みを用い、それを LDA に組み込むことで、文脈に即した意味的一貫性のあるトピック抽出を目指す。これにより、年代別評価基準の分析を実現する。

2. TF-IDF を用いた LDA

本研究では比較対象として、従来広く用いられている LDA を用いたトピックモデリングを実施した。具体的には、各口コミ文に対して TF-IDF ベクトルを算出し、これを入力として LDA を適用するという手順をとった。この手法は語の出現頻度に基づく BoW (Bag-of-Words) モデルに依拠しており、語彙的特徴に基づくトピック抽出が可能であるが、語間の文脈的関係性を捉えることは困難である。そのため、本研究ではこの手法をベースラインとして、文脈情報を取り入れた手法との比較を行った。

3. 提案手法

文脈的特徴と語彙的特徴の両面を反映したトピック抽出を実現するため、本研究では BERT による文書埋め込みと TF-IDF ベクトルを統合し、それを LDA に適用する手法を提案する。

まず、各口コミ文に対して事前学習済みの BERT モデルを用いて文書レベルの埋め込みベクトルを取得し、次元削減のために主成分分析 (PCA) を適用した。次に、対応する TF-IDF ベクトルとともに、それぞれを別個に全結合層に入力し、非線形な特徴抽出を行った。この変換処理は、

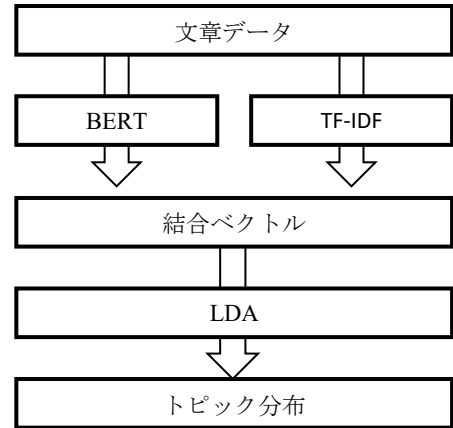


図 1 BERT と TF-IDF の結合と LDA

ニューラルネットワーク構造により BERT と TF-IDF の特徴空間を整合させ、融合することを目的としている。

続いて、2 つの変換後ベクトルを連結し、文脈の意味情報と語の出現頻度情報を統合した特徴ベクトルを構築した。この統合ベクトルは、従来の BoW や TF-IDF ベースの LDA では考慮されにくい意味的特徴を保持しつつ、語彙的な重要語の影響も反映した表現となっている。

最後に、得られた統合特徴量群を入力として LDA を適用することで、口コミ文書群に潜在するトピック構造を抽出した。この手法により、従来型 LDA よりも意味的一貫性と解釈性の向上が期待できるトピック分類を可能とする。本手法の処理フローを図 1 に示す。

4. シミュレーション

年代別の口コミにおける評価指標の傾向を把握するため、異なるトピックモデリング手法による比較分析を行った。まず、ベースラインとして TF-IDF ベクトルを入力とした従来型の LDA を用いた。次に、BERT による文書埋め込みと TF-IDF ベクトルを統合し、LDA に入力する手法を実施した。なお、各手法においてトピック数は 4 から 7 まで変化させ比較を行った。

4.1 使用データセット

京都駅周辺の宿泊施設における口コミを対象とした。京都は 17 件のユネスコ世界遺産を有し、日本国内で最多の登録数を誇る。年間観光客数は約 7,578 万人にのぼり、国内外から多くの旅行者が訪れている。なかでも京都駅周辺は複数の交通手段が集中する観光拠点であり、宿泊および観光の動向を把握するうえで適したエリアである。そこで口コミデータは、じゃらんの Web サイト^[3]より取得を行っ

[†] 明治大学大学院先端数理科学研究科 Meiji University Graduate School of Advanced Mathematical Sciences

た。対象施設は、じゃらんで設定可能な価格帯ごとにおすすり度の高い上位5件を抽出し、価格帯間で重複しないよう選定した。各施設から過去1年分の口コミ本文および投稿者の年齢層に関するデータを収集し、合計で5,396件の口コミデータを取得した。

4.2 評価方法

本研究では、モデルの性能評価において定量的評価と定性的評価の2つを行った。

まず定量的評価として、トピックモデルの品質を測る指標の一つである CV コヒーレンススコア^[4]を算出した。CV コヒーレンススコアは、トピック内の語彙が意味的にどれほどまとまりを持っているかを測定する指標であり、単語同士の共起頻度に基づく情報と、語彙の類似度に基づく情報を組み合わせて算出される。各トピックに対して、上位単語を抽出し、単語間の共起頻度を算出した。その後、共起情報から得られる単語ベクトル間のコサイン類似度を計算し、トピック内の全単語ペアに対して意味的な一貫性の指標として統合した。

トピックごとにこのスコアを算出し、それらの平均値をもってモデル全体のコヒーレンススコアとし、異なるトピックモデル間の比較評価を行った。

一方で、トピックモデルの実用性においては、単に数値的な指標だけでなく、人間がトピックをどれだけ直感的に理解・解釈できるかが重要であると考えた。そこで、本研究では主観的評価として、目視によるトピックの「解釈のしやすさ」や「代表単語の妥当性」に関する評価も実施した。

このように、コヒーレンスによる客観的指標と、人間の解釈に基づく主観的評価の両面からモデルの性能を総合的に検証することで、意味的精度と実用面のバランスを評価した。

4.3 シミュレーション結果

表1に各条件におけるシミュレーション時のモデル精度比較を示す。

シミュレーション結果から、多くの条件でBERT埋め込みとTF-IDFを統合した手法が、従来のLDAに比べてコヒーレンススコアで高い値を示し、より意味的に一貫したトピック抽出が可能であることが定量的に確認された。一方で、BoWを直接用いていないため、トピックごとの代表単語による直感的な理解が難しく、従来LDAの強みである「解釈のしやすさ」は一部失われている。

代表文ベースでの分析では、トピック分類としての精度向上は見られず、むしろ文章の長さや形式といった文体や種類の違いに近い分類結果となった。これにより、トピック構造の明確化には顕著な効果が確認できなかった。

以上より、意味的な精度は向上したものの、実用面で重要な可読性や解釈性の向上には課題が残ると考えられる。

表1 コヒーレンススコアによる精度比較

トピック数	4	5	6	7
LDA (TF-IDF)	0.481	0.470	0.456	0.418
LDA (結合)	0.505	0.498	0.501	0.464

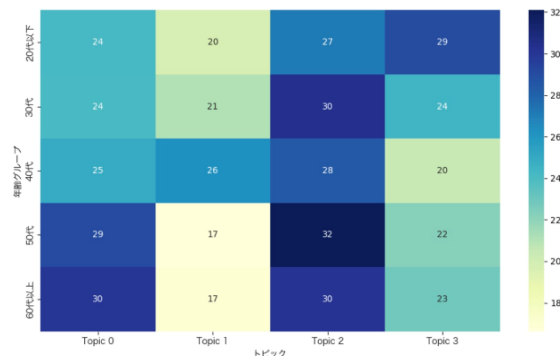


図2 年齢グループ内のトピック割合

4.4 年代別分析

年齢層によって口コミの注目点や評価の傾向には明確な違いが見られた。20代を中心とする若年層は、風呂や施設の快適さ、朝食、もてなしなどの「体験的な価値」に着目し、ポジティブな口コミが多くみられた。また、リピーターとしての利用は少なく、宿泊施設を“特別な体験”として一時的に楽しむ傾向がうかがえた。30代では家族連れでの利用が多く、朝食や家族向けサービスへの評価が高く、快適性と利便性のバランスを重視する姿勢が顕著だった。

一方で、40代以降の中高年齢層では、立地やアクセス、荷物預かり、スタッフ対応などの「実用的な価値」に対する評価が増え、不満点についても積極的に言及する傾向が見られた。彼らはリピーターや同行者との利用が多く、宿泊施設に対して“安定した品質”を強く求めていることが推察される。

このように、年齢層によって重視する観点や期待するサービスが異なり、ターゲットに応じたサービス設計の必要性が示唆された。

図2に例としてトピック数4の時の年齢ごとのトピック割合を示す。それぞれトピック0はサービス、トピック1は子連れ、トピック2はリピーター・同行者、トピック3は清潔さ・快適さとラベル付けすることができる。

5. おわりに

BERT埋め込みとTF-IDFの統合により、意味的なトピック抽出には一定の効果が見られたが、直感的な解釈性には課題が残った。これは、文全体の意味を捉えるBERTと、語の出現分布に基づくLDAの特性の違いによると考えられる。また、年代別に口コミを分析した結果、重視される観点到の違いが見られ、世代ごとの価値観を反映した観光戦略の必要性が示唆された。

参考文献

- [1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol.3, No.4 (2003).
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv preprint arXiv:1810.04805, (2018).
- [3] 宿・ホテル予約-旅行ならじゃらん net, <https://www.jalan.net/>, 2024年7月29日閲覧
- [4] Michael Röder, Andreas Both, Alexander Hinneburg, “Exploring the Space of Topic Coherence Measures”, Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM), Vol.8, No.8 (2015).