7W-01

ツイートデータを活用した 意見抽出モデルの構築と精度改善

間明拓海† 櫻井義尚‡

明治大学大学院先端数理科学研究科† 明治大学総合数理学部‡

1. はじめに

近年，インターネット環境やスマートフォンが普及したことによってSNSの利用者数が増加し，膨大な数の投稿がSNS上に発信されている．SNSに発信されている投稿の中には，企業がマーケティングを行なっていく上で重要となる意見が存在しており，SNS上の意見を企業が収集し分析を行う「ソーシャルリスニング」という手法の重要性が高まっている．しかし，Twitter中の意見は含まれる割合が少なく，ソーシャルリスニングのために膨大な数の投稿の中から手動で意見の抽出を行うことは現実的ではない．そのため，インターネットからの意見抽出の研究が行われているが，機械学習によりその抽出モデルを構築する場合，ランダムに取得したツイートをアノテーションすることで教師データを作成すると不均衡データとなるため，大量の教師データを作成することが難しい．

本研究では．意見の「大量の教師データを作成することが難しい」という問題を解決するために，機械的にアノテーション可能なバズツイートデータと転移学習を活用した意見抽出モデルの構築と精度向上を試み，その有用性を検証する．

2. 関連研究

インターネットからの意見抽出に関する研究としては，以下の2つの研究が挙げられる．

① 立石ら[1]は，「良い」，「好き」のような物事に対する肯定または否定の評価を表す表現の辞書である評価表現辞書を用いることによって意見抽出を行うシステムを提案した．

② 野崎ら[2]は，辞書フィルタを活用して段階的にサンプリングするアノテーション手法であるPSSAを用いて意見の不均衡データであるという問題を緩和することによって教師データとして使用できるようにし，教師あり機械学習モデルを用いて意見抽出を行うことができるシステムを提案した．

意見抽出には，評価表現辞書のような辞書を用いる方法と機械学習を用いる方法があるが，立石ら[1]のような評価表現辞書だけによる意見抽出では機械学習を用いた場合と比べて精度が劣る．そのため，機械学習を用いた意見抽出に関する研究が多く行われているが，野崎ら[2]の研究から分かるように，機械学習を用いた意見抽出では意見が含まれる割合が少ない不均衡データであることから生じる「大量の教師データを作成することが難しい」という問題にどのように対応するかが重要となる．この問題に対して，野崎らは教師データ作成における改善を試みている．本研究では,機械的にアノテーション可能なバズツイートデータと転移学習を活用した意見抽出モデルの構築と精度向上を試みる．

3. 提案手法

本研究では，自己教師あり学習を用いてバズツイート分類モデルを構築し，そのモデルをソースモデルとして転移学習を行うことで意見抽出モデルの構築を行う．(図1)

本章では提案する意見抽出モデルの詳細について述べる．

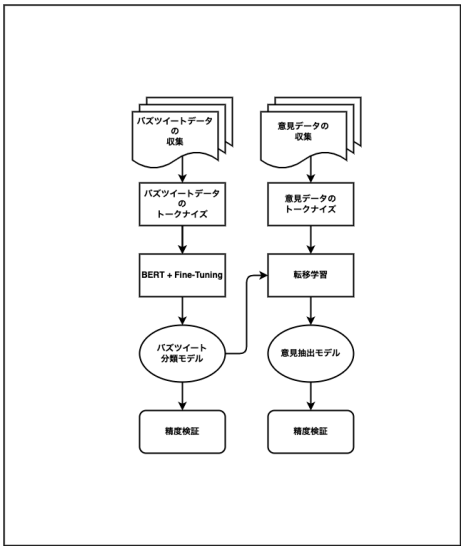

図1 提案する意見抽出モデルのイメージ図

Improving the accuracy of opinion extraction models using tweet data

†TAKUMI MAMYO, YOSHITAKA SAKURAI, Meiji University

3.1 バズツイートデータの収集

Twitter API を用いて，特定のトピックに関連した以下に示すバズツイートの条件に当てはまるツイートの検索を行い，バズツイートデータを収集する．

【バズツイートの条件】
・いいね数とリツイート数の合計が 100 以上
・フォロワー数が 1000 人以下

3.2 バズツイートデータのトークナイズ

SentencePiece を用いることによってバズツイートデータをサブワードに分割する．SentencePiece は，テキストをサブワードに分割することができるトークナイズモデルであり，形態素解析に比べて扱う語彙数と未知語を少なくすることができる．

3.3 BERT と Fine-Tuning を用いたバズツイート分類モデルの構築

日本語 Wikipedia で事前学習を行なった BERT 事前学習モデルに 768 ユニット 1 層とクラス分類用の 2 ユニット 1 層からなる全結合層を追加する．このモデルに対し訓練用バズツイートデータセットを入力して BERT 事前学習モデルの NSP-Dense 層と追加した層のみ学習を行うことによって Fine-Tuning を実装し，転移学習のソースモデルとなるバズツイート分類モデルを構築する．さらに，構築したモデルにテスト用バズツイートデータセットを入力することによって精度検証を行う．

3.4 意見データの収集

意見データとしては，野崎らにより構築された Twitter からの意見抽出のためのデータセット[2]を用いる．

3.5 意見データのトークナイズ

ツイートデータと同様に，SentencePiece を用いることによって意見データをサブワードに分割する．

3.6 バズツイート分類モデルと転移学習を用いた意見抽出モデルの構築

バズツイート分類モデル(クラス分類用の全結合層を除いたもの)に 768 ユニット 1 層とクラス分類用の 2 ユニット 1 層からなる全結合層を追加する．このモデルに対し訓練用意見データセットを入力して追加した全結合層のみ学習を行うことによって転移学習を実装し，バズツイート分類モデルと転移学習を用いた意見抽出モデルを構築する．さらに，構築したモデルにテスト用意見データセットを入力することによって精度検証を行う．

4. おわりに

本研究では，機械学習を用いて意見抽出を行う場合に，意見が含まれる割合が少ない不均衡データであることから生じる「大量の教師データを作成することが難しい」という問題に対応したバズツイートと転移学習を活用した意見抽出モデルを提案した．

今後の課題として，提案したモデルを用いて実際に意見抽出を行い，その精度と有効性の検証を行う必要がある．

参考文献

[1]立石健二，石黒義英，福島俊一．”インターネットからの評判情報検索.” 情報処理学会研究報告自然言語処理(NL)2001.69(2001-NL-144)(2001):75-82

[2] 野崎雄太,櫻井義尚.「Twitter からの意見抽出モデル構築のための教師データ作成手法.」研究報告数理モデル化と問題解決(MPS)2020-MPS-127.9(2020):1-6.

[3] 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).