6S-09

インフルエンサーツイート分類モデルと転移学習を用いた意見抽出システムの構築

間明拓海† 櫻井義尚‡
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」という問題を解決するために，少量の教師データからでも高精度なモデルを構築することができる転移学習という手法を用いた意見抽出システムを提案する．

2. 関連研究

インターネットからの機械学習を用いた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 川島ら[1]は，表現辞書とn-gram判定，半教師あり学習の手法の1つであるDistant Supervisionを用いて半自動的に教師データを収集し，Support Vector Machineを用い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．

川島ら[1]，野崎ら[2]の研究から分かるように，機械学習を用いて意見抽出を行う場合には，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」という問題をどのように解決するかが重要となる．

3. 意見抽出システム

本研究では，複雑なアノテーションの必要がないインフルエンサーツイート分類モデルと転移学習を用いた意見抽出システムを提案する．(図1)

本章ではシステムの詳細について述べる．

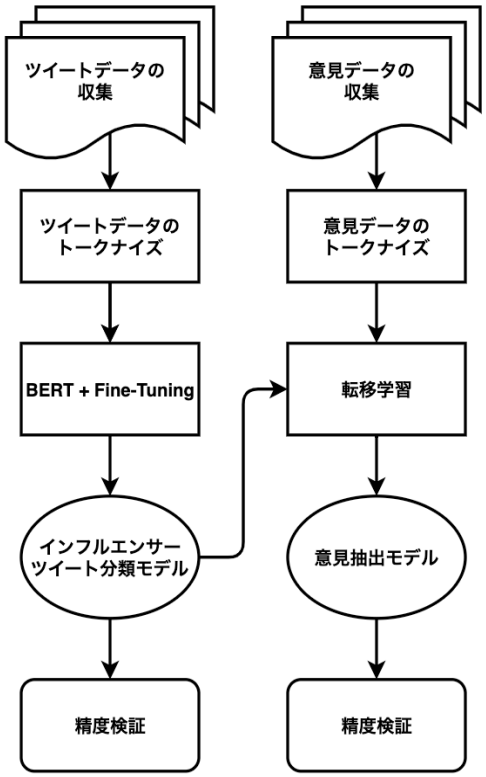

図1 提案した意見抽出システムのイメージ図

Construction of opinion extraction system by influencer tweet classification model and transfer learning
†TAKUMI MAMYO, YOSHITAKA SAKURAI, Meiji University

3.1 ツイートデータの収集

インフルエンサーの条件に当てはまるユーザーのツイートの検索と特定キーワードを含むツイートの検索を行い，ツイートデータを収集する．

3.2 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ツイートデータをサブワードに分割する．SentencePiece とは，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 を用いたインフルエンサーツイート分類モデルの構築

SentencePiece と日本語 Wikipedia で事前学習を行なった BERT 事前学習モデルに過学習を防ぐためのドロップアウト層と 2 値分類を行うための全結合層を追加したモデルに対し訓練用ツイートデータを入力し，Fine-Tuning を行うことによってインフルエンサーツイート分類モデルを構築する．さらに，構築したインフルエンサーツイート分類モデルにテスト用ツイートデータを入力することによってインフルエンサーツイートモデルの精度検証を行う．

3.4 意見データの収集

教師データとしては，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．以下これを意見データと呼ぶ．

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 インフルエンサー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

過学習を防ぐためのドロップアウト層と 2 値分類を行うための全結合層以外の重みを固定した状態のインフルエンサーツイート分類モデルに訓練用意見データを入力し，転移学習を行うことによって意見抽出モデルを構築する．さらに，構築した意見抽出モデルにテスト用意見データを入力することによって意見抽出モデルの精度検証を行う．

3.7 転移学習を行う際の学習済みモデルとしてインフルエンサーツイート分類モデルを用いる理由

① インフルエンサーは「フォロワー」の数，インフルエンサーツイートは「いいね」や「リツイート」の数によって機械的に決定することができるため，複雑なアノテーションの必要がない．

② インフルエンサーが発信する情報は重要度が高いものが多いため，インフルエンサーツイート分類は意見抽出と関連性が高いと考えられる．

4. おわりに

本研究では，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」という問題に対応したインフルエンサーツイート分類モデルと転移学習を用いた意見抽出システムを提案した．

今後の課題として，提案したシステムを用いて実際に意見抽出を行い，その精度を検証する必要がある．また，従来の意見抽出手法と比較し，本研究の有効性を検証する．

参考文献

[1] 川島崇秀,佐藤哲司,神門典子.「半教師あり学習を用いた要望ツイートの抽出手法の評価.」マルチメディア,分散,協調とモバイルシンポジウム 2016 論文集 2016(2016):38-43.
[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.