ツイートを利用した家で楽しめる視聴コンテンツの発見

谷口　愛依

新型コロナウイルスの影響により外出自粛が強いられたが，多くの人が家での時間を有意義にするため，家での過ごし方に工夫を凝らし，その様子を「おうち時間」というキーワードとともに Twitter 上に投稿するようになった．また，家で過ごすにあたり，視聴コンテンツの需要が急増したが，どのような視聴コンテンツが実際に提供され，利用されているのかが不明である．そこで本研究では，Twitter のツイートから，投稿したユーザが家でどのような視聴コンテンツを利用しているのかを明らかにする手法を提案することを目的とする．

提案手法では，「おうち時間」が含まれるツイートのうち,「視聴コンテンツを利用している」ツイート，「視聴コンテンツを提供している」ツイート,「それ以外」のツイートに，事前学習済み言語モデルの一つである BERT をファインチューニングすることで分類する．そして,「視聴コンテンツを利用している」もしくは「視聴コンテンツを提供している」と分類されたツイートから，視聴コンテンツのタイトル名を明らかにする．本研究では，タイトル名候補語として，ツイート中からカギ括弧と二重カギ括弧で囲まれている語句，ハッシュタグの語句を抽出した．ハッシュタグの語句は，一般固有名詞に限定する．形態素解析器は MeCab を，辞書として mecab-ipadic-NEologd を使用した．

BERT を用いたツイートの分類精度の評価は, SVM による分類を比較対象とし, 5 分割交差検証により，評価を行った．実験データは，日本において緊急事態宣言が発令されていた 2020 年 3 月 1 日から 6 月 30 日までに横浜市民が，キーワードとして「おうち時間」を含めて投稿したツイート 2,752 件を対象とした．実験の結果, BERT を用いた分類の F 値，再現率，適合率，正解率は，それぞれ 0.713, 0.695, 0.762, 0.708 の値を得て，すべての値において SVM による分類を上回る精度を得た．

タイトル名の抽出では，それぞれの手法ごとの正解タイトルとの一致度によって，その精度を確認した．実験データは，「視聴コンテンツを利用している」もしくは「視聴コンテンツを提供している」ツイートのうち，正解タイトルが含まれるツイート 772 件を対象とした．実験の結果，カギ括弧と二重カギ括弧で囲まれた手法と，ハッシュタグの語句を手がかりとした手法の一致度は，それぞれ，0.766，0.435 となった．

（指導教員　関洋平）