ニューラルネットワークによるTwitterエンゲージメントの予測

石川県立金沢泉丘高等学校　岩本 周也　篠地 佑宣　清水 慶人　白井 元己

研究の流れ

①TwitterAPIを用いてツイートを収集する
②ツイートを絵文字の消去などの前処理にかける
③前処理済みのツイートを分かち書きする
④Doc2Vecを用いてツイートをベクトル（数値データ）に変換する
⑤ベクトルとフォロー数・フォロワー数・曜日・時刻のデータを使ってモデルを作成し、検証・考察する

モデル作成の下準備

利用システム

TwitterAPI

Twitter社から提供されている特定のワードを含んだツイートを収集することができるもの

Doc2Vec

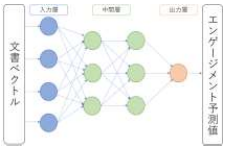
ライブラリ「Gensim」で使用できる文章をベクトルに変換することのできる機能[1]

BERT

Google社から提供されている文章を双方向から学習させた事前学習モデル

Neural Network

入力した値に、重みをかけて線形変換した値を次の層へ送る、という計算操作がネットワーク状に連なったもの。評価関数はMSE、最適化関数はAdamを用いた。



イメージ図

エンゲージメント

今回の研究では、以下のように言葉を定義する

エンゲージメント＝
いいね数＋リツイート数

パターン

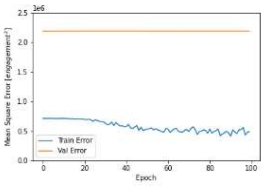
ニューラルネットワークに入力するデータの組み合わせ

パターン①：テキスト
パターン②：テキスト・フォロー数・フォロワー数・曜日・時刻
パターン③：フォロー数・フォロワー数・曜日・時刻

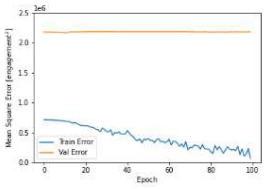
モデルの作成と考察

「**オススメ**」という言葉の含んだツイートを22029件収集し、モデルを学習した。ベクトルは400次元とした。

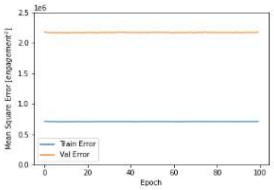
パターン①



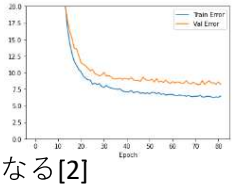
パターン②



パターン③

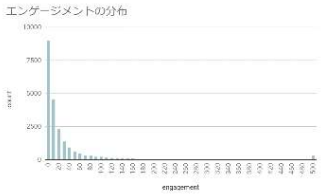


・縦軸：MSE　横軸：epoch
・オレンジ：検証データ
青色:訓練データ
・正しく学習できているときのグラフは右上のような推移になる[2]



・学習しても未知のデータである検証データに対して予測の誤差が減少していない（収集データの調査へ）
・ベクトルの次元数が大きすぎて入力が多くなり予測がうまくいっていない可能性がある（モデルの改善へ）
・パターン①とパターン②より、テキスト以外のデータが予測に影響を及ぼしていると考えられる
・パターン③では訓練データでも学習が進まないためテキストは確実に予測に必要であるとわかる
➡ただし、一般的な感覚ではテキスト以外のデータとエンゲージメントの間だけでも相関性がありそう（モデルの改善へ）

収集データの調査



22029ツイート中
エンゲージメントが500以上のものは**313ツイート**のみ

エンゲージメントが極端に多いものが外れ値となっている

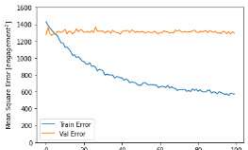
モデルの改善

実験①：エンゲージメントの範囲を250以下・次元数を100に変更
実験②：TwitterAPIがpopularと判断したエンゲージメントの多いツイート154件・BERTでEmbeddingを次元数は768

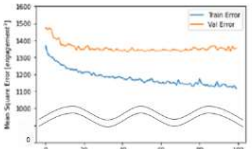
追加実験の結果

実験①

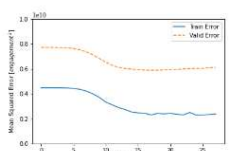
パターン②



パターン③



実験②



・実験①のパターン②では検証データに減少傾向がみられないため学習がうまくいっていない。
・パターン③では両方とも減少傾向がみられる。
・実験②では両方とも減少傾向がみられる。

エンゲージメントが少ないテキストなしのデータとエンゲージメントの多いテキストありのデータに関して予測するモデルを作成できたといえる。

今後の展望

・より高精度なモデルの作成
・キーワードに縛られない一般的なモデルの作成
・多くのエンゲージメントを得る方法の理論化と実社会で広告への応用

参考文献

[1]Tomas Mikolov. Distributed Representations of Sentences and Documents. https://cs.stanford.edu/~quocle/paragraph_vector.pdf(参照2021-11-10).
[2]Google."回帰：燃費を予測する" https://www.tensorflow.org/tutorials/keras/regression?hl=ja(参照2021-11-10)
他多数