ツイート感情分析タスクにおける BERT を用いたデータ蒸留によるラベル偏りの改善

山城颯太[1] 山下達雄[1,2]
[1]ヤフー株式会社 [2]Yahoo！JAPAN 研究所
{soyamash, tayamash}@yahoo-corp.jp

概要

Yahoo!リアルタイム検索[i]ではユーザによって検索されたツイート群に対して感情の極性判定を行い，その統計結果を示す感情分析機能を提供している．本研究ではこの感情分析機能のシステム刷新に伴って発生したアノテーション済み学習データのラベル偏りの問題と，これに対して行った BERT を用いたデータ蒸留によるラベル偏り改善の取り組みについて報告する．

1 はじめに

ユーザによって生成されたコンテンツ（UGC）に対して感情の極性を判定することによって，すでに提供されているサービスの内容に新たな付加価値を提示し，よりユーザにサービスへの興味を持ってもらうことができると考えられる．

その具体的な事例の一つとして，Yahoo!リアルタイム検索がある．本稿ではこの感情分析機能の刷新に伴って発生したアノテーション済み学習データのラベル偏りの問題と，それに対して行った BERT を用いたデータ蒸留の取り組みについて報告する．

1.1 リアルタイム検索と感情分析

Yahoo!リアルタイム検索はインターネットで広く使われている代表的な SNS である「Twitter」のツイートを対象として始まった検索サービスである．2011 年 6 月にサービスが開始され，現在も継続して運用されている．

検索以外に，その時々のバズトピックとそれに対する感情分析情報を提供する機能がある．バズトピックは，ツイートに含まれる，一定時間内に急激に頻度が増大したキーワード（トレンドワード）をベースに自動的にまとめられる．バズトピックに含まれるツイートの感情ラベルを集計した「感情の割合」も提示される．感情ラベルは検索対象となるツイートすべてに対し，感情分析機能により自動付与される（個別ツイートの感情ラベルは現在ユーザには提示されない）．

本稿では，置き換え対象（比較対象）である感情分析機能を「旧モデル」と呼ぶ．旧モデルは文献[1]のものが基本となっているが，正解ラベル付きツイートの使用量や細かなロジックなど若干の変更はある．

2 データセット

本研究で使用した正解ラベル付きツイートは，リアルタイム検索でトレンドワードとして選ばれた検索語を含むツイートから抽出し，それらに感情ラベル（ポジティブ・ネガティブ・ニュートラル）を 3 人のアノテータが付与したものである．作成時期とラベル決定方法の異なる 2 つのデータセットがある．

- データセット 1: 約 5 万件．ツイートの時期は 2013 年ごろ．アノテータ 3 人の付与したラベルが一致したツイートのみを用いる．主に学習データとして用いた．
- データセット 2: 約 800 件．ツイートの時期は 2020 年前半．主に評価データとして用いた．

データセット 2 には約 50 のトレンドワードが含まれる．それぞれのトレンドワード（トピック）に対してアノテータがそのトレンドワードを含むツイートを見て感情ラベルを付与している．

[i] https://search.yahoo.co.jp/realtime

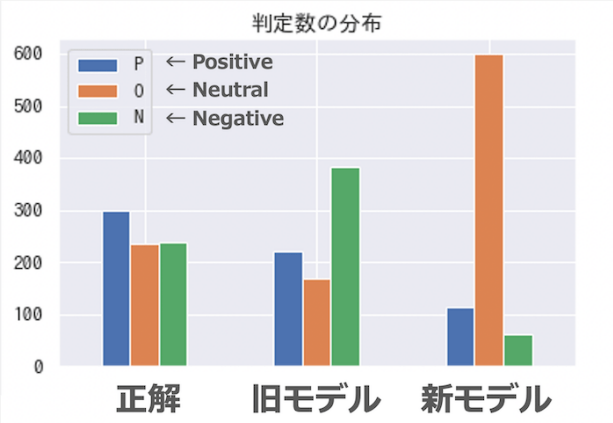

図 1 モデルごとの出力ラベル比率

3 ラベル偏りの問題

データセット 1 を学習データとし，ポジティブ・ネガティブ・ニュートラルを出力する多クラス分類モデルを作成した．これを「新モデル」と呼ぶ．ここでは推論速度とメモリ使用量の観点からシンプルな形態素ベースの線形分類器を採用した．作成されたモデルでデータセット 2 の各ツイートにラベル付与し，精度評価を行ったところ，旧モデルと同等の精度を達成した．しかしその一方で，Recall 不足が確認された．この原因についての考察を以下に述べる．

3.1 モデルごとのラベル出力の偏り

図 1 にデータセット 2 のツイート群に対する旧モデルと新モデルの出力ラベルの比率を示す．一番左の棒群は人手で付けられたアノテーションラベル（正解）であり，ここではポジティブ・ニュートラル・ネガティブの比率が大体等しくなるようツイートが収集されたことがわかる．

中央の棒群は，同じツイート集合に対する旧モデルの出力ラベルの比率を示している．人手ラベルの分布と比べて少々ネガティブの出力数が多いが，比較的人手ラベルと似たラベル比率である．ただし，ここでは人手ラベルとの一致（正誤）は見ず，あくまで出力されたラベルの比率だけに着目していることに注意されたい．

一番右の棒群が新モデルの出力ラベル比率である．一目でわかるように，ニュートラルの出力数がずば抜けており，ツイート全体の約 75%を占めている（Recall 不足）．これは明らかに均等な出力とは言えない．

3.2 学習データのラベルの偏り

3.1 節で述べた問題が起きた理由を我々は精査し，その原因が今回使用した学習データにあるという仮説を立てた．

データセット 1 の約 5 万件のツイートに付けられた人手ラベルを改めて観察したところ，ポジティブ・ニュートラル・ネガティブの比率が大体 1:5:1 になることがわかった．つまりモデルの側からしてみれば，与えられたツイートに対して無条件にニュートラルさえ出力しておけば，ツイート内容を考慮せずとも 5/7（約 71%）の確率で正解できてしまうということになる．そのためモデルは自身の Accuracy を上げるために，出力すべきラベルの確信が持てないツイートに対してはニュートラルを比較的多く出力するようになる．実際，改めて図 1 を見直すと，新モデルの出力比率は 1:5:1 に近そうに見える．この偏りは，学習データ中のラベル出現比率を過学習してしまった結果だと言える．

3.3 アノテーションの難しさ

ではそもそもなぜ，今回の学習データはそのような偏りの大きいラベル比率（imbalanced data） になってしまっているのか．もちろん，Twitter 中の真のラベル比率自体が偏りを抱えていることも原因の一つであると考えられる．しかし今回使用している学習データは真のラベル比率以上の偏りを抱えているように見受けられた．そしてその大元をたどれば，ツイートの印象を推定するというタスクそのものの難しさに原因があると言える．

たとえば，「今日のお昼はカレーだった。大盛り無料で嬉しかった。だけど最後まで食べきれず残念……」というツイートのラベルは何かを考える．一文目だけなら出来事を報告しているだけなのでニュートラル．二文目も含めて読むなら「嬉しかった」と言明しているのでポジティブ．しかし三文目まで着目すると「残念」と言明しているのでネガティブ．このように，一つのツイート中にさまざまな印象が入り乱れている．ここに挙げたツイートは作例だが，これと同様に Twitter には皮肉や自虐など判定の難しいツイートが多く含まれている．また絵文字や感嘆符など，その包含によってツイートの印象が大きく変わる文字も多く出現する．

この難しさは人手で正解ラベルを付ける場面にも直接影響している．今回は学習データとしてデータセット 1 の約 5 万件のツイートを用いていたが，これはもともと約 25 万件のツイートそれぞれに 3 人の人間によって正解ラベルが付けられたのち，3 人とも同じラベルを選んだツイートのみを集めたものである．つまり残りの約 20 万ツイートについては，3 人の付けたラベルが完全一致しなかったということである．

その結果，集められた約 5 万件のツイートについては,『複数人の人手ラベルが一致しやすいデータ』というサンプル選択の偏り（sample selection bias）が掛かっていると考えられる．その偏りが，このデータを用いて学習させたモデルの予測結果において，出力ラベルの極端な比率として現れたのだと考えられる．

4 対処手法

3 節で述べた問題に対して，我々は BERT を用いたデータ蒸留によるオーバーサンプリングを行うことで対処した．BERT とは Transformer ベースの事前学習済み言語モデルであり，目的タスクに応じて追加で fine-tuning を行う手法である[2]．これを用いた我々の対処手法の詳細を以下に報告する．

4.1 オーバーサンプリング

偏りの大きいラベル比率（imbalanced data）に対するシンプルな手法としてはオーバーサンプリングが考えられる．これは比較的量が少ないラベル（今回の例で言えばポジティブとネガティブ）のデータを改めて収集・追加することで，学習データ中のラベル比率を調整する手法である．

しかし，ここでデータを追加するということはすなわち，新たに集めたツイートに対して追加で人手ラベルを付与する必要があるということである．数万件のラベル付けとなればその人手コストは無視できないほど甚大なものになる．

4.2 BERT を用いたデータ蒸留

そこで我々は，BERT を使用することで，このラベル付け作業を自動化した．我々が BERT を用いた手順は以下のようになる．

（1）まず，手元にある約 5 万件のデータを用いて BERT を学習させる．（2）次に，新たに収集したツイート約 200 万件に対して BERT による推論を行い，

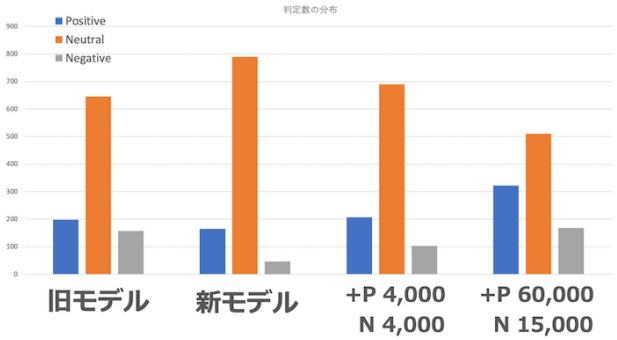

図 2 データ追加後の出力ラベル比率

ここで付与されたラベルを擬似的な正解ラベルと見なす．（3）それから，BERT によって付けられたラベルがポジティブ・ネガティブであるツイートをそれぞれ数千〜数万件ずつ集め，もともとの約 5 万件の学習データに追加する．（4）最後に，この新たな学習データを用いて我々のモデルを学習させる．

この一連の手順（この手法はデータ蒸留[3]と呼ばれている）を踏むことで，我々の新モデルはもともとの学習データに含まれていたラベル偏りの影響を大幅に軽減することができた．

5 比較評価

5.1 出力ラベルの比率の改善

実際にどれほど出力ラベルの比率が改善されたかを図 2 に示す．

一番左の棒群は，ランダムにサンプリングした 1,000 ツイートに対する置き換え対象のモデル（旧モデル）の出力ラベルの比率を示している．左から二番目の棒群は，同じツイート集合に対する何も手を加えていない今回新たに作成したモデル（新モデル）の出力ラベルの比率を示している．右から二番目の棒群は新モデルの学習データにポジティブと判定されたツイートを 4,000 件，ネガティブと判定されたツイートを 4,000 件追加した時の出力ラベルの比率を示している．一番右の棒群は新モデルの学習データにポジティブと判定されたツイートを 60,000 件，ネガティブと判定されたツイートを 15,000 件追加した時の出力ラベルの比率を示している．学習データにおけるポジティブ・ネガティブツイートの量を増やせば増やすほど，よりニュートラルの出力が減り，代わりにポジティブ・ネガティブの出力が増えていることがわかる．この出力ラベル比率の改善は最終的な精度の改善に直結するものと考えられる．

5.2 自動評価の限界

では実際のところ，どれほどの擬似ラベル付きデータを追加すればいいのか．一つの案として，学習データの一部をあらかじめテストデータとして確保しておき，このデータに対する自動評価スコアが最大になるような追加量を求めることが考えられる．しかし，おそらくこの部分データ自体もラベル偏りを含んでいるため，必ずしも正確な評価ができるとは限らない．理想的には，Twitter 中の真のラベル比率をどうにか観測し，これと一致するように訓練データの比率を調整したい．しかし，そもそもツイートに対して付けられる人手ラベル自体がほとんど一致しないため，真のラベル比率はどうしようとも正確に観測できるものではないと考えられる．

5.3 人手モデル選択・トピック単位評価

そこで我々は人手での評価に基づいてこのデータ追加量を決めた．具体的には 1,000〜5,000 件刻みでデータ量を変化させたモデルを約 60 種類用意した．このうち最も人間の感覚と近い出力ができていそうなモデルを 5 種類選出し，これらを用いて最終的な決定のためにトピック単位の比較評価を行った．

各モデル出力の評価にはデータセット 2 のトピック単位の正解ラベルを用いた．これはサービスの利用実態に近い形にするためである．ツイートごとに自動付与された感情ラベルをトピックごとに集計し，一番頻度が高いものをそのトピックの感情ラベルとし，これをトピックの正解ラベルと比較することで精度を出した．表 1 に今回のいくつかのバリエーションと旧モデルによるトピック単位の Accuracy を挙げる．最終的には，ポジティブ（P）60,000 件，ニュートラル（O）2,000 件，ネガティブ（N）45,000 件のデータ追加を行った新モデルを採用した．

表 1 トピック単位の評価結果

モデル + 追加データ内訳	Acc.
旧モデル	0.683
新モデル（追加データなし）	0.244
+ P: 4,000, N: 4,000	0.512
+ P: 60,000, N: 15,000	0.878
+ P: 60,000, O: 2,000, N: 20,000	0.756
+ P: 60,000, O: 2,000, N: 30,000	0.805
+ P: 60,000, O: 2,000, N: 45,000	**0.902**

6 おわりに

本稿では Yahoo!リアルタイム検索の一機能である感情分析機能の刷新に伴って発生したアノテーション済み学習データのラベル偏りの問題と，それに対して行ったオーバーサンプリングの取り組み，その評価に際して行った人手でのラベル比率の調整，トピック単位評価について報告した．

本稿で示された手法は，BERT を教師モデルとするデータ蒸留によって，従来のオーバーサンプリングに必要とされる人手アノテーションのコストを大幅に下げ，かつ柔軟なラベル比率の調整を可能としたものである．

今回報告した範囲では，あくまで簡易な比較評価のみを行っており，より高性能なモデルとの比較や詳細な評価・分析はまだ十分に行えていない．しかし，実際にサービスに提供した範囲では，旧モデルと比べて遜色ない Precision・Recall・速度を示しており，もとの目的であったモデル刷新の観点では十分な性能を示している．

参考文献

1. ヤフージャパンのリアルタイム検索における感情分析. **野畑 周, 内藤 弘朗, 清水 徹.** 第 5 回テキストマイニング・シンポジウム, 2014.
2. **Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
3. **Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He.** Data Distillation: Towards Omni-Supervised Learning. In *Proceedings of CVPR*, 2018, pp. 4119-4128.