BERT と能動学習を用いた攻撃的なツイートの分類

及川　正樹†　　井上　潮‡

†東京電機大学　工学研究科　情報通信工学専攻 〒120-8551 東京都足立区千住旭町 5 番
‡東京電機大学　工学部　情報通信工学科 〒120-8551 東京都足立区千住旭町 5 番

E-mail:　†21kmc03@ms.dendai.ac.jp,　‡inoueu@mail.dendai.ac.jp

あらまし　Twitter をはじめとした SNS はコミュニケーションツールとして有用であるが，攻撃的な投稿などの有害な投稿も多く存在する．機械学習によってこのような有害な投稿を分類することで，ユーザを保護することが可能になる．しかし，日本語の学習データがほとんど公開されていないため，分類器の学習に必要なデータが不足するという問題がある．本研究では，攻撃的なツイートか否かの 2 クラス分類を，言語モデルの BERT を用いて行う．この際，学習に必要なデータが不足する状況を踏まえて，学習データを適応的に選択する能動学習を適用する．本研究では，基礎的な手法である Entropy や近年提案された ALPS を含む 10 種類の能動学習手法に関して有効性を検証する．検証の結果，いずれの手法も学習データの選択に有用である事が示された．最も良い手法では，全ての学習データ 8000 件のうちの 1600 件のみの利用で，F 値 0.708 という高い精度が得られた.ランダムに選択した場合は，同程度の精度に達するために 4000 件以上の学習データが必要であり，能動学習により学習データの大幅な削減が可能である事が示された．

キーワード　テキスト分類, Twitter, BERT, 能動学習

1. はじめに

Twitter をはじめとした SNS は，様々なユーザと容易に交流できるなど，コミュニケーションツールとして有用である．一方で，他者を攻撃する投稿をはじめとした，ユーザに害をなす投稿も多く存在する．これらの投稿は，単に閲覧者に不快感を与えるだけでなく，ネット炎上等において当事者を自殺に追いやってしまうという痛ましい事件も引き起こしている．従って，このような有害な投稿を分類し，ユーザを保護する必要がある．

本研究では，有害な投稿の中でも特に重篤な被害を及ぼす攻撃的な投稿に焦点をあて，Twitter におけるこのような投稿である，攻撃的なツイートの分類を行う．分類には自然言語処理モデルの BERT(Bidirectional Encoder Representations from Transformers)[5]を用い，日本語のツイートを対象に，そのツイートが攻撃的か否かの 2 クラス分類を行う．分類器に BERT を用いる理由は，多くの自然言語処理タスクにおいて高い性能を発揮する事が示されているためである[2][3].

BERT は，Attention 機構[14]を持っている点と，大量の文章により事前学習 (pre-training) を行っている点で，従来の RNN(Recurrent Neuarl Network)ベースのモデルと大きく異なる．Attention は並列的に全てのトークンを参照する事が可能であり，高い性能を発揮するだけでなく，計算の効率が良いという利点がある．事前学習は，自己教師あり学習である Masked Language Model(MLM)および Next Sentence Prediction(NSP)を行い，事前に埋め込み表現を学習する．これをもとに，個々のタスクに対しての学習(fine-tuning)を行う事で高い性能を発揮する．

しかし，BERT の fine-tuning を含む分類器の学習において，学習データが不足するという問題がある．特に日本語のデータは，著作権の問題等によりほとんど公開されておらず，既存のデータセットを利用する事が困難である．従って，学習データを用意するための，時間や労力，転じて金銭といったコストが要求される．特に，ラベル付けを人手で行う場合には大きなコストが要求される．これらの問題は，攻撃的なツイートの分類に限らず，多くの自然言語処理タスクに共通する問題である．

攻撃的なツイートの分類においては，次の点からラベル付けのコストが高じると考えられる．

・攻撃的なツイートは，そうでないツイートよりも少ない．
・攻撃的か否かの判断に，複数人によるラベル付けが必要．
・攻撃的なツイートの分類を細分化する必要がある．

このような問題に対処するための手法として，学習データを適応的に選択する能動学習(Active Learning)[13]が存在する．今日までに複数の能動学習手法が提案されおり，学習したモデルの予測結果やデータの分布を利用する事で，単に学習データをランダムに選択する場合よりも学習に有用なデータを選択できる事が実験により示されている[4][6][9][10][11][12][15][16][17][18]．従って，より少ない学習データでモデルがより良い性能を発揮する事が期待でき，学習データを用意するためのコストを削減できる可能性がある．

一方で，より良い性能を発揮する能動学習手法は，データやモデルの組み合わせによって変わる事が指摘されている[6]．深層学習モデルを用いた日本語の文章分類に，能動学習を適用する研究は少なく，日本語の攻撃的なツイートの分類を含む，このような状況における能動学習の有効性は明らかにされていない．

従って本研究では，BERT を用いた日本語の攻撃的なツイートの分類における，能動学習の有効性を検証する．基礎的な手法である Entropy[13]をはじめとして，近年提案された ALPS[15]を含む 10 種類の手法に関して比較を行う．

2. 関連研究

SNS や，インターネット上における有害な文章を分類する研究は，これまで多く行われている．石橋ら[1]は，電子掲示板の一種である 2 ちゃんねるを対象に，SVM を用いた悪口文と非悪口文の分類を行っている．

近年はこのような分類タスクには BERT が用いられることが多い. 松本ら[3]は, 日本語のツイートを対象とし, 相手の感情を逆撫でる内容の, 煽りツイートの分類に BERT を用いている. リプライツイート単体および, そのリプライ元のツイートとの組を対象とする 2 通りの場合において, BERT を用いた分類器を作成している. その他の分類器として非線形 SVM, k-近傍法(KNN), ランダムフォレストを用いて複数の実験を行っているが, BERT が最も良い精度を示している. なお, 煽りと非煽りのデータ数が同数でリプライツイートのみを学習した BERT が, 最も良い精度である F 値は 0.715 を示した事が報告されている. 一方で, ツイート組を学習した場合と比較して, 両者にほとんど差がないことも指摘されている.

本研究で分類を行う攻撃的なツイートの一種として皮肉がある. 皮肉は, 肯定的な表現で否定的な内容を示すことから, 文の極性判定を困難にすることが知られている. そのため, 皮肉を対象として分類を行う研究は盛んに行われている分野である. 諏訪ら[2]は, 日本語のツイートを対象に BERT による皮肉の分類を行っている. ベースラインとなる Bi-LSTM 及び Attention を用いたモデルに対して, BERT が最も良い精度を示したことが報告されている. また, 皮肉の分類にツイート中の絵文字を利用した実験を行っている. 実験では絵文字を用いた場合の方が良い精度を示すことが確認されているが, 誤分類が多い事が指摘されている.

能動学習は, 様々なドメインのデータを対象に, 様々な機械学習モデルを用いて研究が行われている. Ein-Dor ら[6]は, BERT による文章分類に能動学習を適用した大規模な研究を行っている. 複数の英語のデータセットを対象に, 様々な能動学習手法を用いた実験を行い, 能動学習の有効性を検証している. 能動学習によって選択した場合には, ベースラインであるランダムな選択に対して, F 値が 4%から 8%高い事が報告されている. このことから, 能動学習と BERT の組み合わせが有効であり, 特に分類対象のデータ(正例)が不足する状況における有効性を示している.

深層学習モデルへの転換に伴い, 新たな能動学習手法の提案や, 旧来の手法を深層学習モデルに適応させる研究もいくつか行われている. Yuan ら[15]は, BERT および BERT をベースとしたモデルを対象にした能動学習手法を提案している. BERT の事前学習に着目しているという点で他の手法と大きく異なり, 事前にラベル付きデータを必要とせずに能動学習による選択が行えるという利点がある.

3. 前提条件

3.1. 攻撃的なツイートの定義

本研究では, 攻撃的なツイートを次のように定義する.

・皮肉や嫌味を含む, 攻撃となる表現・内容.
・非難や, 過度・過激な批判を行う表現・内容.
・ある事柄に対しての反応として失礼な表現・内容.

上記表現・内容が含まれると, 自然な範囲で解釈できるツイートを攻撃的なツイートと定義する. ただし, 次のいずれかを満たす場合は攻撃的なツイートではないとする.

・自虐など, 対象が自分自身のみの場合.
・明らかに攻撃の意図がないと解釈できる場合.
・他者の発言等の引用部分が攻撃的であるが, ツイートにおいてその引用部分を支持しない場合.

3.2. 使用データ

本研究で用いるツイートの収集は, 2021/9/16 から 2021/11/10 の間で行い, TwitterAPI の StreamingAPI を使用した. 収集したツイートの内, 次の条件を満たすツイートからランダムに選んだ 10000 件のツイートをラベル付けの対象とした.

条件1: 日本語のツイートである.
条件2: 100 件以上リツイートされている.
条件3: 動画, 画像, URL を含まない.
条件4: リプライを 10 件以上取得する事が出来た.
条件5: 特定の文字列を含まない.

なお条件 5 は, キャンペーンへの応募を促すツイートなどのノイズとなるデータを除外するために設けた. 特定の文字列としては, リツイート, いいね, フォローなどの操作を表す単語, プレゼントや抽選などの単語, また, プレゼントを表す絵文字を指定し, これらを含むツイートを除外した.

抽出した 10000 件のツイートに攻撃的なツイートか否かのラベル付けを人手で行った. この作業は第一著者のみが行い, ツイートの文章を読み, 前節にて示した攻撃的なツイートの定義に合致するか否かの判断を行った. ラベル付けの結果として, 攻撃的なツイート 2197 件と攻撃的でないツイート 7803 件を得た.

これらのラベル付けした 10000 件のツイートを, 学習, 検証, テストデータに分割した. 分割に際しては, 層化サンプリングを行い, それぞれ(80/10/10)%となるようにデータセットを構築した. なお, 能動学習を用いた実験を行う場合は, 学習データの一部のみを実際に学習に使用し, 残りの学習データを選択対象のラベルなしデータとして疑似的に扱う.

3.3. BERT に関して

本研究では, 東北大学の乾研究室が公開している, 日本語の事前学習済み BERT[1]を使用する. モデルの種類として, Huggingface で公開されている bert-base-japanese-whole-word-masking を使用する. 事前学習は日本語の Wikipedia 記事で行われており, モデルの構造は標準的な BERT-base に基づいたものである.

ここで, BERT を用いた文章のベクトル化手法を定義する. 入力トークンの系列(文章) xをベクトル化することを$\psi(x)$と表し, [CLS]トークンに対しての最終層の出力を文章ベクトルとする. なお, このベクトルは BERT の出力次元数である 768 次元である.

4. 能動学習の問題設定

能動学習は, 学習データを適応的に選択する手法である. 基本的な問題設定は次の通りである.

手順 1: ラベルなしデータから, モデルの情報等をもとに, ある基準に従って学習に使用するデータを選択する.
手順 2: ラベルを付与する存在のオラクルが, このデータにラベルを付与し, 新たに学習データとして追加する.

オラクルとして最もよく用いられるのは, 人手でのラベル付けである. 本研究においても, 付与されるラベルは全て, 人手でラベル付けを行ったものである.

ここで, 能動学習がうまく機能するかは, 手順 1 において, どのような基準に従って学習データを選択するかが重要であ

[1]https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

る．本章において，能動学習の手順および，比較を行う能動学習の選択基準について述べる．

4.1. 能動学習の手順

ラベルなしデータの扱いに関して，いくつかの手法が存在するが，本研究では，pool-based active learning を適用する．

Pool-based active learning では，少量のラベル付きデータ$\mathcal{L}_0$と，大量のラベルなしデータ$\mathcal{U}_0$をあらかじめ用意する．

具体的な手順として，はじめに$\mathcal{L}_0$でモデルの学習を行い，能動学習の選択基準に基づき$\mathcal{U}_0$の中から追加する学習データ$\mathcal{B}_1$を選択する．

このとき，$\mathcal{B}_1$にラベル付けを行い，$\mathcal{U}_0$から$\mathcal{B}_1$を取り除き$(\mathcal{U}_1 \leftarrow \mathcal{U}_0 \setminus \mathcal{B}_1)$，$\mathcal{L}_0$に$\mathcal{B}_1$を加える$(\mathcal{L}_1 \leftarrow \mathcal{L}_0 \cup \mathcal{B}_1)$．その後，選択したデータを加えた新たな学習データ$\mathcal{L}_1$でモデルを学習し，新たにデータの選択を行う．この手順を繰り返し行い，徐々に学習データを増加させていく．

データの選択をT回行うとすると，t回目$(t \in [1, T])$の繰り返しに関して次のように一般化できる．

手順1: $\mathcal{L}_{t-1}$での学習をもとに$\mathcal{B}_t \subset \mathcal{U}_{t-1}$を選択する．
手順2: $\mathcal{B}_t$にラベル付けを行う．
手順3: $\mathcal{U}_{t-1}$から$\mathcal{B}_t$を取り除く$(\mathcal{U}_t \leftarrow \mathcal{U}_{t-1} \setminus \mathcal{B}_t)$.
手順4: $\mathcal{L}_{t-1}$に$\mathcal{B}_t$を加える$(\mathcal{L}_t \leftarrow \mathcal{L}_{t-1} \cup \mathcal{B}_t)$.
手順5: 新たな学習データ$\mathcal{L}_t$でモデルの学習を行う．

一般に，深層学習モデルは学習に時間がかかる事から，$\mathcal{B}$を複数件選択するバッチ能動学習が用いられる．本研究においても，一度に複数件の選択を行うバッチ能動学習を用いる．

4.2. 能動学習の選択手法

本節では，能動学習手法に関して概説する．

まず，能動学習手法の分類の一つとして，warm-start と cold-start に分ける事ができる．両者の大きな違いは，初期の学習データ$\mathcal{L}_0$を選択可能かどうかという点にある．

warm-start の選択手法は，$\mathcal{L}$でモデル学習したうえで，追加するデータの選択を行う．この手法では，最初の選択において必要になる，初期の学習データ$\mathcal{L}_0$自体は能動学習によって選択できないという欠点がある．

一方で，cold-start の選択手法は，$\mathcal{L}$でモデル学習する必要がないという点で優れている．従ってこの手法では$\mathcal{L}_0$を能動学習によって選択する事が可能である．

4.2.1. Warm-start 能動学習の選択基準

Warm-start 能動学習では以下の選択基準を用いる．

Entropy[13]**:** $\mathcal{L}$を学習したモデルP_θで$x \in \mathcal{U}$の予測を行い，予測の不確実性に基づいて選択を行う．ここでは，不確実性の指標としてエントロピーを用い，エントロピーが大きいデータを選択する．i番目のラベルをy_iとすると選択するデータは次のように表される．

$$\underset{x \in \mathcal{U}}{\operatorname{argmax}} -\sum_i P_\theta(y_i|x) \log P_\theta(y_i|x) \tag{1}$$

他の指標としては，最も高い予測確率が最も低いデータを選択する Least confident や，最も高い予測確率と次に高い予測確率の差が大きいデータを選択する Margin of confidence などが提案されている．ただし，2 クラス分類においては，エントロピーを含めたいずれの指標も等価である．

Bayesian Active Learning by Disagreement (BALD)[9]**:** Entropy と同様に予測の不確実性に基づいて選択を行う．ドロップアウトは通常，モデルの学習時のみに適用するが，これを予測時に適用することで近似的にベイズ推論を行える事が示されている[8]．本手法は，ドロップアウトを適用したうえで複数回行った予測から，不確実性を計算する．不確実性の指標としては，max-entropy を用いる．max-entropy は，得られた複数の予測から，クラスごとに予測確率を平均して計算したエントロピーの大きいデータを選択する．i番目のクラスの予測確率の平均をp_iとすると，式(2)で表される．なお本研究では，予測を行う回数Nを 10 とした．

$$\underset{x \in \mathcal{U}}{\operatorname{argmax}} -\sum_i p_i \log p_i \tag{2}$$

Core-Set[12]**:** k-center 問題の解を利用し，$\mathcal{L}$と$\mathcal{U}$の全てを学習に使用した場合に近くなるデータの選択を行う．本研究では貪欲法を用い，$\mathcal{L}$を学習した BERT の特徴空間における，k-center 問題の解を利用する．すなわち，全てのデータをψによって文章ベクトルに変換した上で，k-center 問題を解く．

Discriminative Active Learning (DAL)[10]**:** $\mathcal{L}$に含まれない特徴を持ったデータを選択する．具体的には，$\mathcal{L}$と$\mathcal{U}$のどちらに属するかを分類するモデル$\widehat{P_\theta}$を学習し，$\mathcal{U}$に属すると予測した確率が高いデータを選択する．$\widehat{P_\theta}$の入力には，$\mathcal{L}$を学習した BERT によって変換された，文章ベクトルを用いる．データが$\mathcal{U}$に属する事を示すラベルをuとすると選択するデータは式(3)で表される．

$$\underset{x \in \mathcal{U}}{\operatorname{argmax}}\, \widehat{P_\theta}\big(y = u|\psi(x)\big) \tag{3}$$

Expected Gradient Length (EGL) [17]**:** 学習したモデルに対して大きな変更を加える事が期待されるデータを選択する．具体的には，$\mathcal{L}$を学習したモデルP_θが，$x \in \mathcal{U}$を学習したときの，勾配の大きさの期待値が高いデータを選択する．あり得るすべてのラベルの，各場合に関して計算を行う事で，期待値を算出する．各ラベルに対するモデルの予測確率と，そのラベルを正解ラベルとして学習した場合に計算される勾配の大きさを乗算し，加算したものが勾配の大きさの期待値である．なお，勾配の大きさには L2 ノルムを用いる．損失関数lの勾配を∇lとすると式(4)で表される．本研究では，出力層の ユニットのみを対象として，勾配長を計算した．

$$\underset{x \in \mathcal{U}}{\operatorname{argmax}} \sum_i P_\theta(y_i|x) \|\nabla l(P_\theta(x), y_i)\| \tag{4}$$

Batch Active Learning by Diverse Gradient Embeddings (BADGE)[4]**:** 勾配ベクトルのクラスタリングによって，モデルの状態と，データの多様性を考慮して選択を行う．$\mathcal{L}$を学習したモデルP_θで$x \in \mathcal{U}$の予測を行い，最も予測確率の高かったクラスを，仮の正解ラベル$\hat{y}$とする．勾配ベクトルg_xは式(5)で表され，P_θにおいて正解ラベルに$\hat{y}$を用いてxを学習したときの，出力層の重みθ_{out}に対する勾配を要素として持つベクトルである．なお，損失関数lはクロスエントロピーである．

$$g_x = \frac{\partial}{\partial \theta_{\text{out}}} l(P_\theta(x), \hat{y}) \tag{5}$$

この勾配ベクトルのクラスタリングを行い，各クラスタの中心に最も近いデータをそれぞれ選択する．クラスタリングの手法としては k-means++を用い，クラスタリングに際して L2 正規化を行う．クラスタ数として，選択するデータの件数を指定する．

Contrastive Active Learning (Contrastive)[11]**:** 特徴空間において，近傍のデータと予測確率が異なるデータを選択する．$x \in \mathcal{U}$に関して，KNN を用いて，学習した特徴空間におけるk個の近傍データ$\{x_1, \cdots, x_k\}$を$\mathcal{L}$から選び出す．その後，学習し

たモデルP_θの予測確率を用いて，x と各近傍データの全ての組み合わせに対して KL ダイバージェンスを計算する．最後に平均を求め，この値の大きいデータを選択する．これを式(6)に示す．なお，選択する近傍のデータ数kは 10 とした．

$$\underset{x\in\mathcal{U}}{\operatorname{argmax}}\frac{1}{k}\sum_{i}^{k}\mathrm{KL}\left(P_\theta(x_i)||P_\theta(x)\right) \tag{6}$$

4.2.2. Cold-start 能動学習の選択基準

Cold-start 能動学習では以下の選択基準を用いる．

Active Learning by Processing Surprisal (ALPS)[15]**:**BERT の事前学習に用いられる MLM の損失を利用して，データの選択を行う．

MLM は，入力トークンのいくつかをランダムにマスクし，そのトークンが何であったかの予測を行うタスクである．出力は，マスクされたトークンごとに，モデルの語彙すべてに対しての確率として得られる．MLM 損失は，マスクされる前の元々のトークンを正解として，クロスエントロピーを用いて計算される．

本手法では，Surprisal Embeddings として，入力トークンの系列長(事前に定めた最大トークン長)の要素を持つベクトルを定義する．このベクトルのi番目の要素は，i番目の入力トークンに対して計算された MLM 損失である．ただし，i番目の入力トークンがマスクされていない場合には，代わりに要素として 0 を持つ．$x\in\mathcal{U}$に対して Surprisal Embeddings を求め，このベクトルのクラスタリングを行う．そして，各クラスタの中心に最も近いデータをそれぞれ選択する．

なお，クラスタリングの手法としては k-means を用い，クラスタリングに際して L2 正規化を行う．クラスタ数として，選択するデータの件数を指定する．

BERT-KM[15]**:** 本手法では，ALPS で用いた Surprisal Embeddings の代わりに，事前学習のみを行った BERT の埋め込み表現を用いる．すなわち，$x\in\mathcal{U}$に対する$\psi(x)$を用いて，クラスタリングを行う．なお，クラスタリングの手順は ALPS と同様である．

また，類似した手法として，FT-BERT-KM が存在する．こちらは，cold-start の手法ではなく，埋め込み表現への変換ψに，$\mathcal{L}$を用いて学習(fine-tuning)を行った BERT を用いる．

5. 実験

本研究では，3 つの実験を行う．なお，いずれの実験も，モデルの評価には F 値を用いる．

実験 1 では，精度の推移および，全ての学習データを使用した場合の精度を確認する事を目的として，少量の学習データから徐々に学習データを増加させる実験を行う．

実験 2 では，warm-start の能動学習手法の有効性を確認する事を目的として，少量の学習データから能動学習を用いて徐々に学習データを増加させる実験を行う．この時，一度に選択するデータの件数$|\mathcal{B}|$を，(50/250/500)件の 3 通りで比較を行う．

実験 3 では，cold-start の能動学習手法の有効性を確認する事を目的として，実験 2 と類似した実験を行う．ただし，$\mathcal{L}_0$の選択に能動学習を用いるという点で異なる．

ここで，学習に関わるパラメータは全て同一のものを使用する．バッチサイズを 10，学習率を 10^{-5} とし，Adam による最適化を行う．学習は 10 エポック行い，検証データに対して最も良い精度を示したエポックのモデルを評価および，能動学習によるデータの選択に使用する．

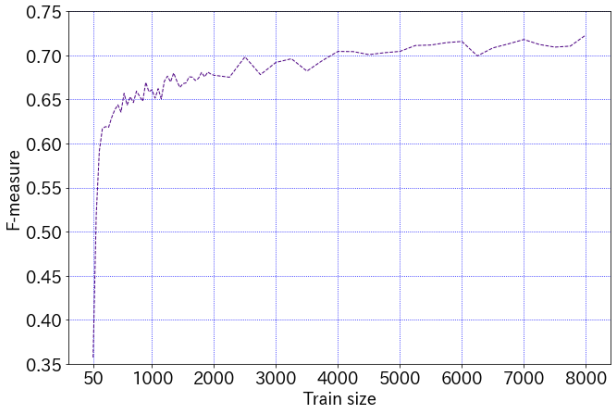

図 1 ランダムな学習データの追加における F 値

ただし，Ein-Dor ら[6]が指摘している通り実際の状況では検証データが十分に使用できない事が考えられるが，本研究でも同様に，結果の安定のために検証データを用いる．

5.1. 実験 1: 全ての学習データの使用

本実験では，精度の推移および，全ての学習データを使用した場合の精度を確認する事を目的とし，50 件の学習データから全ての学習データ 8000 件まで徐々に学習データを増加させる実験を行う．

5.1.1. 実験手順

最初に学習データ 8000 件から，初期の学習データを，層化サンプリングによって 50 件選択する．その後学習データが 2000 件に達するまでは 50 件ずつ，以降は 250 件ずつ，残りの学習データからランダムに学習データを追加する．なお，追加の際には層化サンプリングを用いない．各段階における学習データを用いてモデルの学習と評価を行う．

これを，初期の学習データ$\mathcal{L}_0$を変更し 5 回実験を行う．

5.1.2. 実験結果

5 回の実験の平均の F 値を図 1 に示す．まず初期の学習データ 50 件のみでは，F 値 0.357 と非常に低い精度である事がわかる．一方で，学習データが 200 件に達するまでの追加では急激に精度を上げている．この時，学習データが 200 件という少量のデータのみで，F 値 0.617 という比較的高い精度を示している．その後は，全ての学習データである 8000 件に達するまで，精度の上昇が鈍化しながらも，緩やかに上昇する．

全ての学習データ 8000 件で学習したときの精度は F 値が 0.723 となった．同時に，5 回の実験の平均値のみを考慮すると，この値が最大値となる．この値は，実験 2 と実験 3 において，一部の学習データのみを用いる場合に達成可能な精度の一つの目安となる．

5.2. 実験 2: Warm-start 能動学習の適用

4.2 章にて示した warm-start の手法に関して，4.1 章の手順に従って学習データを追加する実験を行う．同時に，能動学習を用いないランダムな選択と比較を行う．

5.2.1. 実験手順

最初に学習データ 8000 件から，初期の学習データ$\mathcal{L}_0$を，層化サンプリングによって 100 件選択する．その後残りのデータをラベルなしデータ$\mathcal{U}_0$として扱い，4.1 章と同様の手順で学習データの追加を行う．この時，一度に選択するデータの件数$|\mathcal{B}|$を，(50/250/500)件の 3 通りで比較を行う．追加するデータの総数を 1500 件とし，それぞれ$T=\lceil 1500/|\mathcal{B}|\rceil$回の繰り返しを行う．なお，選択したデータ$\mathcal{B}$へのラベル付けに関して

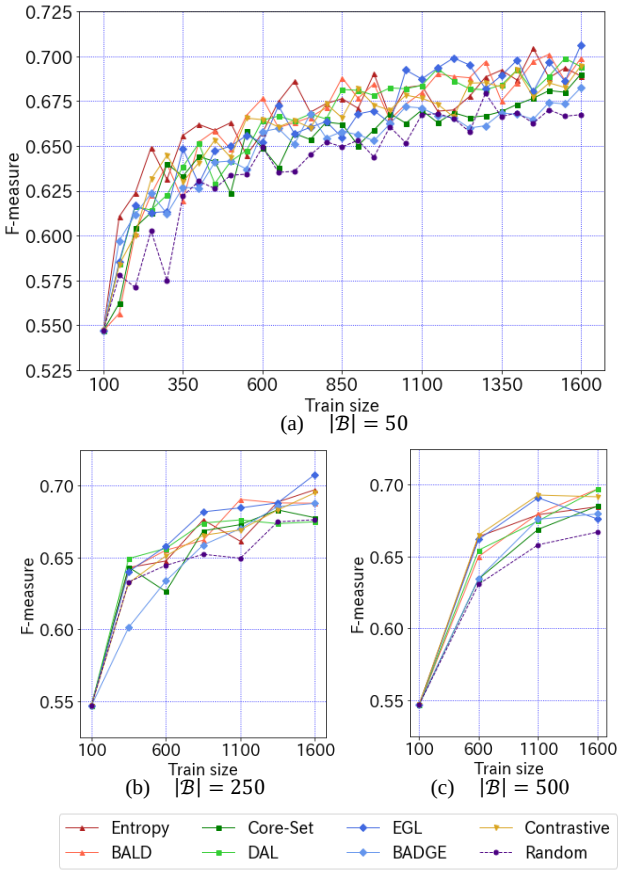

図 2 Warm-start 能動学習による学習データの追加における F 値

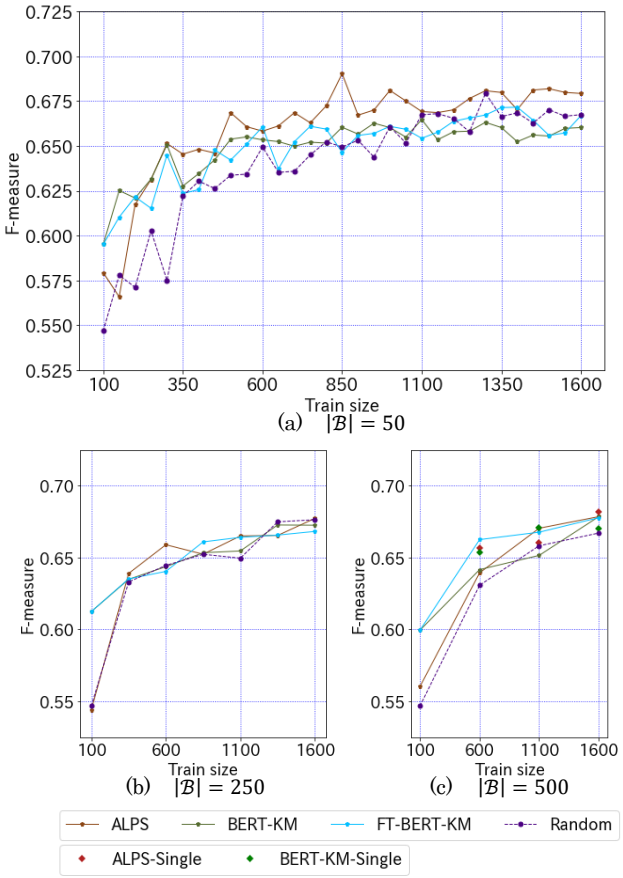

図 3 Cold-start 能動学習による学習データの追加における F 値

は，あらかじめ 3.2 節にて付与したラベルをそのまま用いることとする．

$\mathcal{L}_0$から$\mathcal{L}_T$までの各繰り返し時のデータで学習したモデルに関して F 値による評価を行う．なお，実験 1 と同様に，初期の学習データ$\mathcal{L}_0$を変更し 5 回実験を行う．各能動学習手法に関して独立に実験を行うが，$\mathcal{L}_0$を用いて学習したモデルのみ共通とする．

5.2.2. 実験結果

5 回の実験の平均の F 値を図 2 に示す．一度に選択するデータの件数に関わらず，全体としてはいずれの能動学習手法も，ランダムな選択を上回る結果となった．同時に，それぞれの条件において，各能動学習手法間に優劣がみられた．特に繰り返しの前半では，一部の手法がランダムな選択を下回る場合が確認できた．各手法の詳細な比較は 6 章にて行う．

最終的に最も高い精度を示したのは，$|\mathcal{B}| = 50$の場合は EGL が F 値 0.706，$|\mathcal{B}| = 250$の場合は EGL が F 値 0.708，$|\mathcal{B}| = 500$の場合は BALD が F 値 0.697 となった．なお，ランダムな選択の場合における最終的な F 値は，順に 0.667，0.676，0.667 である．このことから，最も良い手法で F 値が 0.03 から 0.04 程度向上する事が確認できる．

また図 1 からは，一つの目安である F 値 0.7 を恒常的に上回るためには，学習データが 4000 件以上必要である事が確認できている．従って，能動学習を用いる事で，その 4 割の件数である 1600 件の学習データのみでこの精度を達成した事となる．これは，能動学習により，学習データを大幅に削減できる事を示唆している．

なお，全ての学習データ 8000 件で学習したときの精度は F 値が 0.723 であったことから，その差はおよそ 0.02 である．

5.3. 実験 3: Cold-start 能動学習の適用

4.2 章にて示した cold-start の手法である ALPS，BERT-KM，および FT-BERT-KM に関して実験を行い，ランダムな選択と比較を行う．

5.3.1. 実験手順

基本的な手順は実験 2 と同様である．ただし，初期の学習データ$\mathcal{L}_0$の選択に能動学習を用いるという点で異なる．なお，FT-BERT-KM のみ$\mathcal{L}_0$の選択が不可能であるため，$\mathcal{L}_0$の選択に BERT-KM を用いる．その他の手法は，$\mathcal{L}_0$の選択および追加するデータ$\mathcal{B}$の選択を同一の手法によって行う．

実験 2 と同様に，一度に選択するデータの件数$|\mathcal{B}|$を，(50/250/500) 件の 3 通りで比較を行う．また，ALPS と BERT-KM に関して，$\mathcal{L}_0$の件数として(600/1100/1600)件を一括で選択した場合の実験を行う．

$\mathcal{L}_0$から$\mathcal{L}_T$までの各繰り返し時のデータで学習したモデルに関して F 値による評価を行う．実験は，初期の学習データ$\mathcal{L}_0$の選択を含めて 5 回行う．各能動学習手法に関して独立に実験を行うが，BERT-KM と FT-BERT-KM のみ，$\mathcal{L}_0$を用いて学習したモデルを共通とする．なお，比較対象となるランダムな選択は実験 2 の結果をそのまま用いる．

5.3.2. 実験結果

5 回の実験の平均の F 値を図 3 に示す．なお，$\mathcal{L}_0$の件数として(600/1100/1600)件を選択した結果は図 3(c)に示す．

まず，初期の学習データ$\mathcal{L}_0$を学習した結果から，BERT-KM はランダムな選択よりも高い精度を示している事が確認できる．一方で，ALPS はランダムに選択した場合とほとんど変わ

らない精度である. ALPS は BERT-KM より優れている事が報告されており[15], 今回の実験では異なる結果が得られた.

その後の学習データの追加に関しては, $|\mathcal{B}| = 250$の場合を除けば, 能動学習を用いた選択が, ランダムな選択を上回る結果となった. 一方で, 良い結果を示した warm-start の手法と比較すると, 精度の上昇は低い. なお, 各手法の詳細な比較は 6 章にて行う.

また, ALPS と BERT-KM の両手法に関して$\mathcal{L}_0$を(600/1100/1600)件のそれぞれを一括で選択した場合の結果は, 100 件の$\mathcal{L}_0$に対して徐々に学習データを追加した場合と同程度かそれ以上の精度を示す結果となった. これはすでに報告されている結果[15]と同様の結果となった.

6. 分析

5 章にて行った実験 2 および実験 3 の結果に関して分析を行う. まず統計的な検定を行った後, 追加する学習データとして選択したデータに関する分析を行う.

6.1. 統計的な検定

Warm-start および cold-start 能動学習を用いた実験結果を用いて, 能動学習手法間の優劣および, ランダムな選択に対しての優劣を明らかにするための検定を行う.

6.1.1. 検定方法

Zhang ら[16]は, 近年提案された検定手法である, ASO(Almost Stochastic Order)[7]を用いて能動学習を用いた実験に対して検定を行っている. ASO は深層学習モデル同士の比較を行うために提案された検定方法であり, 本研究では, ASO を用いた検定を行う. その他の検定手法としては, Ein-Dor ら[6]がウィルコクソンの符号順位検定を行っているが, 深層学習モデルと能動学習を組み合わせた実験において, 標準的に使用される検定方法は, 現時点では存在しない.

ASO を用いた検定として, 2 つの手法 A, B のどちらが優れているかの検定を, 全ての手法の組み合わせに対して行う. 有意水準$\alpha = 0.05$とし, ASO の評価値である$\epsilon_{\min}$を計算する.

$\epsilon_{\min} < 0.5$の場合には A が B より優れている事を意味し, $\epsilon_{\min} = 0.0$の場合に A が B より統計的に優れている事を意味する. また, $\epsilon_{\min} = 0.5$の場合には A が B の優劣を決定できない事を意味し, $\epsilon_{\min} > 0.5$の場合は A が B より優れていない事を意味する.

A と B に対しての検定と, A と B を入れ替えて行った検定の, 各$\epsilon_{\min}$の和は 1 となる性質を持っている. 検定に際してこの性質を利用し, 片方の検定のみを行った後, A と B の手法を入れ替えて再度検定を行う代わりに$1 - \epsilon_{\min}$を結果として利用する. なお検定には, 平均を計算する前の, 5 回の実験のそれぞれの結果を用いる. また, 検定の多重性を考慮し, ボンフェローニ補正を行う.

6.1.2. 検定結果

検定結果を表 1 に示す. まずランダムな選択に対して, ほとんど全ての場合において, いずれの能動学習手法も統計的に優れているもしくは, 優れているという事が示された. $|\mathcal{B}|$=50 の場合は Entropy が, $|\mathcal{B}|$=250 の場合は EGL が, $|\mathcal{B}|$=500 の場合は Contrastive が, 他の手法と比較して最も優れている結果となった. いずれの場合でも, 次いで BALD, DAL, EGL が優れているという結果となった. 反対に, Core-Set, BADGE および cold-start の手法は, 他の手法と比べると劣るという結果となった. これらは, データの分布に基づく手法

表 1 ASO を用いた検定結果 (列が手法 A, 行が手法 B. $\epsilon_{\min} = 0.0$を**赤い太字**, $\epsilon_{\min} < 0.5$を*赤い斜体*で示す.)

(a) $|\mathcal{B}| = 50$

	RANDOM	Entropy	BALD	Core-Set	DAL	EGL	BADGE	Contrastive	ALPS	BERT-KM	FT-BERT-KM
RANDOM		1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.82	0.96	0.98
Entropy	**0.00**		*0.01*	**0.00**	**0.00**	*0.01*	**0.00**	**0.00**	**0.00**	**0.00**	*0.01*
BALD	**0.00**	0.99		*0.02*	0.67	*0.34*	*0.12*	0.50	*0.19*	*0.23*	*0.26*
Core-Set	*0.01*	1.00	0.98		0.99	0.99	0.81	1.00	*0.42*	0.68	0.79
DAL	**0.00**	1.00	*0.33*	*0.01*		*0.30*	*0.04*	0.52	*0.15*	*0.19*	*0.20*
EGL	**0.00**	0.99	0.66	*0.01*	0.70		*0.09*	*0.47*	*0.17*	*0.18*	*0.23*
BADGE	**0.00**	1.00	0.88	*0.19*	0.96	0.91		0.98	*0.37*	0.62	0.77
Contrastive	**0.00**	1.00	0.50	**0.00**	*0.48*	0.53	*0.02*		*0.06*	*0.13*	*0.16*
ALPS	*0.18*	1.00	0.81	0.58	0.85	0.83	0.63	0.94		0.58	0.61
BERT-KM	*0.04*	1.00	0.77	*0.32*	0.81	0.82	*0.38*	0.87	*0.42*		*0.26*
FT-BERT-KM	*0.02*	0.99	0.74	*0.21*	0.80	0.77	*0.23*	0.84	*0.39*	0.74	

(b) $|\mathcal{B}| = 250$

	RANDOM	Entropy	BALD	Core-Set	DAL	EGL	BADGE	Contrastive	ALPS	BERT-KM	FT-BERT-KM
RANDOM		1.00	0.99	0.55	0.92	0.98	*0.44*	0.98	*0.17*	0.98	0.94
Entropy	**0.00**		0.76	*0.05*	*0.31*	0.82	**0.00**	*0.07*	*0.01*	*0.04*	**0.00**
BALD	*0.01*	*0.24*		**0.00**	*0.21*	0.87	**0.00**	*0.03*	*0.01*	*0.01*	*0.02*
Core-Set	*0.45*	0.95	1.00		0.94	1.00	*0.08*	0.90	0.65	*0.47*	0.52
DAL	*0.08*	0.69	0.79	0.06		0.81	*0.04*	0.50	*0.17*	*0.01*	**0.00**
EGL	*0.02*	*0.18*	*0.13*	**0.00**	*0.19*		**0.00**	*0.01*	*0.03*	*0.03*	*0.04*
BADGE	0.56	1.00	1.00	0.92	0.96	1.00		0.99	0.79	0.57	0.60
Contrastive	*0.02*	0.93	0.97	*0.10*	0.50	0.99	*0.01*		*0.04*	*0.06*	*0.05*
ALPS	0.83	0.99	0.99	*0.35*	0.83	0.97	*0.21*	0.96		0.96	0.96
BERT-KM	*0.02*	0.96	0.99	0.53	0.99	0.97	*0.43*	0.94	*0.04*		*0.34*
FT-BERT-KM	*0.06*	1.00	0.98	*0.48*	1.00	0.96	*0.40*	0.95	*0.04*	0.66	

(c) $|\mathcal{B}| = 500$

	RANDOM	Entropy	BALD	Core-Set	DAL	EGL	BADGE	Contrastive	ALPS	BERT-KM	FT-BERT-KM
RANDOM		1.00	1.00	0.72	1.00	1.00	0.97	1.00	0.94	0.99	1.00
Entropy	**0.00**		*0.38*	**0.00**	0.61	0.56	**0.00**	0.99	**0.00**	**0.00**	**0.00**
BALD	**0.00**	0.62		**0.00**	*0.42*	0.68	**0.00**	1.00	**0.00**	**0.00**	*0.16*
Core-Set	*0.28*	1.00	1.00		1.00	1.00	0.90	1.00	0.67	0.65	0.95
DAL	**0.00**	*0.39*	0.58	**0.00**		0.68	**0.00**	1.00	**0.00**	**0.00**	*0.13*
EGL	**0.00**	*0.44*	*0.32*	**0.00**	*0.32*		**0.00**	1.00	**0.00**	**0.00**	*0.10*
BADGE	*0.03*	1.00	1.00	*0.10*	1.00	1.00		1.00	*0.25*	*0.02*	0.90
Contrastive	**0.00**	*0.01*	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**		**0.00**	**0.00**	**0.00**
ALPS	*0.06*	1.00	1.00	*0.33*	1.00	1.00	0.75	1.00		0.90	0.99
BERT-KM	*0.01*	1.00	1.00	*0.35*	1.00	1.00	0.98	1.00	*0.10*		1.00
FT-BERT-KM	**0.00**	1.00	0.84	*0.05*	0.87	0.90	*0.10*	1.00	*0.01*	**0.00**	

(diversity-sampling)の一種であり, データの分布が選択に直接的な影響を与えるという点で共通している.

バッチ能動学習では, Entropy や BALD のようなデータの不確実性に基づいた手法(uncertainty-sampling)は, 類似したデータが選択される可能性がある. そのため, データの不確実性のみでなく, 多様性を考慮する必要が示されている[11]. 一方で今回の結果では, Contrastive を除くと多様性を考慮しない手法が優れている結果となった.

ただし, 時系列的な変化に対しては本検定の範囲ではない. 例えば, 図 1 の$|\mathcal{B}|$=50 の場合, 前半は Entropy が高い精度を示しているが, 後半は EGL が高い精度を示している. このような特徴を評価するためには, 他の分析が必要である.

6.2. 選択したデータに関する分析

選択したデータに関して, 4 つの指標を用いて分析を行う.

6.2.1. 分析の指標

Diversity in input space (DIV-I): 入力空間における, 選択したデータ$\mathcal{B}$の多様性を示す指標である. ここでは, $\mathcal{B}$と$\mathcal{U}$における入力トークンの集合である$\mathcal{V}_{\mathcal{B}}$と$\mathcal{V}_{\mathcal{U}}$を用いて, Jaccard 係数を計算する. $\mathcal{B}$が多様なトークン(語彙)を含んでいるほど高い数値を示す.

$$J(\mathcal{V}_{\mathcal{B}}, \mathcal{V}_{\mathcal{U}}) = \frac{|\mathcal{V}_{\mathcal{B}} \cap \mathcal{V}_{\mathcal{U}}|}{|\mathcal{V}_{\mathcal{B}} \cup \mathcal{V}_{\mathcal{U}}|} \tag{7}$$

Diversity in feature space (DIV-F): モデルが学習した特徴空間$\hat{\mathcal{X}}$における, 選択したデータ$\mathcal{B}$の多様性を示す指標である. 各$x_i \in \mathcal{U}$に関して, 最も近い$x_j \in \mathcal{B}$とのユークリッド距離$d(x_i, x_j)$の平均の逆数によって計算される. $\mathcal{B}$が多様なデータを選択するほど高い数値を示す. なお, 特徴空間$\hat{\mathcal{X}}$への変換としてψを用いる.

$$D(\mathcal{B}) = \left(\frac{1}{|\mathcal{U}|} \sum_{x_i \in \mathcal{U}} \min_{x_j \in \mathcal{B}} d(x_i, x_j) \right)^{-1} \tag{8}$$

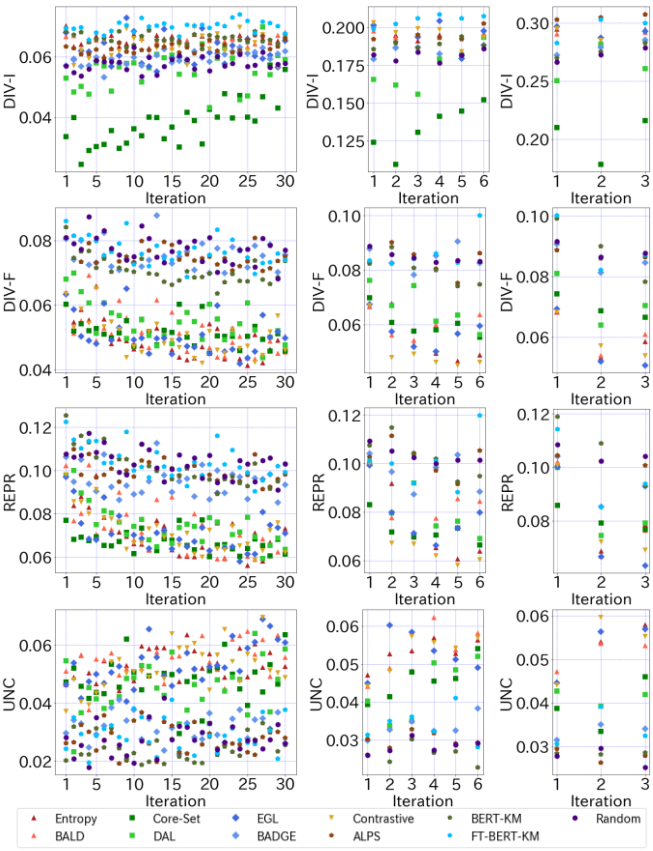

図 4　各分析指標の結果
（上から DIV-I, DIV-F, REPR, UNC　左から$|\mathcal{B}| = 50, 250, 500$）

Representativeness (REPR): 特に不確実性に基づいた能動学習手法では，外れ値を選択する傾向がある事が報告されている[6]．本指標は，モデルが学習した特徴空間$\hat{X}$における，選択したデータ$\mathcal{B}$における外れ値の選択を定量化した指標である．Zhu らによって提案された KNN-Density[18]に基づき，$x_j \in \mathcal{B}$におけるk個の近傍データ$\{x_{i1}, \cdots, x_{ik}\}$とのユークリッド距離の平均の逆数によって計算され，近傍のデータとの距離が近いほど高い数値を示す．本研究では，近傍データ数kを 10 とする．

$$R(\mathcal{B}) = \left(\frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \frac{1}{k} \sum_{j}^{k} d(x_i, x_{ij}) \right)^{-1} \tag{9}$$

Uncertainty (UNC): 選択したデータ$\mathcal{B}$が，モデルにおいて不確実なデータであるかを示す指標である．全ての学習データ$\mathcal{L} \cup \mathcal{U}$で学習したモデル$P_{\theta'}$用いてエントロピーを計算する．不確実なデータが選択されるほど高い数値を示す．

$$E(\mathcal{B}) = -\frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \sum_{i} P_{\theta'}(y_i|x) \log P_{\theta'}(y_i|x) \tag{10}$$

6.2.2. 分析結果

分析結果を図 4 に示す．いずれの指標も，$|\mathcal{B}|$の大きさによらず似たような傾向を示した．

DIV-I に関しては，ALPS が最も高く，Core-Set が著しく低い結果となった．次いで，DAL，BADGE およびランダムな選択が低い結果となった．Warm-start の手法に限っては，特に精度の良かった手法は高い値を示した．DIV-F に関しては，DIV-I と事なる特徴を示し，cold-start の手法および BADGE とランダムな選択が高い結果となった．隔たりはあるが，次点で DAL および Core-Set が高い．

REPR に関しては，DIV-F と似た傾向を示したが，Core-Set は比較的低い結果となった．UNC は不確実性に基づいた手法である Entropy, BALD, Contrastive が特に高く，次点で EGL も高い．良い精度を示した手法は UNC が高い傾向にあるが，いずれの指標もモデルの精度の良し悪しとの明確な関係は確認できなかった．

それぞれの手法に着目すると，精度の良かった手法は DIV-F および REPR が低く，多様なデータの選択が行えておらず外れ値も選択している事が示された．また，特徴空間において多様なデータを選択する Core-Set が，入力空間においては多様なデータを選択できていない事が明らかになった．cold-start の手法を含むデータの分布に基づいた手法に関して，その多くは UNC が低く，それ以外の指標が高いという傾向を示した．

7. 考察

まず，本研究の結果から，BERT を用いた攻撃的なツイートの分類における，優れた能動学習手法について考察する．

本研究では，$|\mathcal{B}|$の大きさによらず，不確実性に基づく手法である Entropy, BALD および EGL が特に良い精度を示した．$|\mathcal{B}|$が比較的大きい値である 500 件の時には，不確実性と多様性に基づいた手法である Contrastive が特に良い精度を示した．このような結果が得られた要因として，選択対象のラベルなしデータ$\mathcal{U}$が比較的小さかったために，$\mathcal{U}$の分布が疎であり，似たような特徴を持つデータが少なかった事が考えられる．故に，バッチ能動学習における不確実性に基づいた手法の，似たようなデータを選択してしまうという欠点が現れなかった事が考えられる．

この点を考慮しなければ，総合的には BERT を用いた攻撃的なツイートの分類には，Entropy が最も有用な warm-start の手法であると考えられる．表 1 の結果から，$|\mathcal{B}|$が 50 件の場合は最も優れた手法であり，250 件および 500 件の場合も一部の手法には劣るものの，優れた結果を示している．また，単に精度の向上という点で優れているだけでなく，選択の基準が単純であるという点でも優れている．そのため，実装が容易であり，また実行時間の点でも優れる．本研究では，実行時間の計測は行っていないが，Ein-Dor ら[6]の実験では能動学習の選択手法としては最も早い実行時間である事が示されている．故に，本研究では，Entropy が総合的には最も有用な warm-start の手法であると考える．

ただし，能動学習を適用する上での事前の要求によっては，他の手法を用いた方が有効である事が考えられる．図 1 の結果から，最終的に最も良い精度を示した手法は，$|\mathcal{B}|$が 50 件および 250 件の場合は EGL, 500 件の場合は BALD である．また，総合的な実行時間を踏まえると，ALPS や BERT-KM で大量の$\mathcal{L}_0$を一度に選択した方がより早いと考えられる.従って，何件のラベル付けを行うかや，どれだけの時間をかけるか，具体的に目標となる精度はどの程度か等の条件に応じて適切な能動学習手法を選択するべきであると考えられる．

また，本研究では実験を行っていないが，複数の手法を組み合わせることも考えられる．特に，cold-start の手法によって$\mathcal{L}_0$を選択した後に，warm-start の手法によって以降のデータを追加するような事が考えられる．

最後に，cold-start の手法が際立って良い精度示すことがなかった点について考察する．

まず, ALPS と BERT-KM の両手法に関わる要因として, クラスタリングが有効ではなかった事が考えらえる. Warm-start の手法である Core-set および BADGE の精度が低いことから, 本研究で用いたデータに対しては, データの分布に基づき, 多様性を獲得するという方向でのデータの選択が有効でなかった可能性がある.

次に, ALPS に関しての要因として, ツイートの文章が短い事が考えられる. 全ての学習データのうち, およそ 6 割のツイートが 50 トークン以下であり, およそ 3 割のツイートは 20 トークン以下である. ALPS において使用する Surprisal Embeddings は入力トークンの系列長の要素を持ったベクトルだが, トークン長が短いデータが多かったために疎なベクトルとなってしまい, 十分な特徴を持っていなかった事が考えられる. この点は, マスクする単語の割合を変更するなど, 異なるパラメータを用いた実験により, 明らかになる可能性がある.

最後に BERT-KM に関しての要因として, 事前学習に使用したデータのドメインと fine-tuning を行うデータのドメインが大きく異なるという点が考えられる. ツイートは, Wikipedia の記事と比較すると口語的な文章や SNS 特有の表現などの異なる特徴を持っている. 従って, 事前学習のみを行った BERT では, より有用な埋め込み表現へと変換できなかった可能性が考えられる. 一方で, fine-tuning 後の BERT を利用する FT-BERT-KM も BERT-KM と同程度の精度を示すことが多かった. この点については, 他の事前学習済みモデルを用いて実験することで明らかになる可能性がある.

8. まとめ

本研究では, BERT を用いた日本語の攻撃的なツイートの分類における, 能動学習の各手法の有効性を検証した. ほぼすべての場合において, 検証に使用した能動学習手法は, ランダムに選択するよりも優れているという事が示された.

能動学習手法同士で比較をすると, $|\mathcal{B}|$=50 の場合は Entropy が, $|\mathcal{B}|$=250 の場合は EGL が, $|\mathcal{B}|$=500 の場合は Contrastive が, 最も優れている結果となった. また, BALD や DAL も比較的優れている結果となった. cold-start の能動学習手法は BERT-KM が初期の学習データ$\mathcal{L}_0$の選択に有効であったが, 以降のデータの追加に関しては warm-start の能動学習に劣る結果となった. ただし, 一度に大量の$\mathcal{L}_0$を選択した場合でもランダムな選択より優れている事が示され, 総合的な実行時間の点では優れている.

今後の課題として, 各能動学習手法が優れた結果を示した原因あるいは, 優れた結果を示さなかった原因を明らかにする必要がある. 特に, 追加で行った分析結果から, 優れた結果を示した手法は, 攻撃的なツイートを本来の割合よりも多く選択している事が判明した. 従って, 学習データラベルの割合と分類精度の関係に焦点を当てた分析が必要である.

ラベル付けコストの観点から実験が困難だが,ラベルなしデータ$\mathcal{U}$のサイズが大きくなった場合の各能動学習手法の振る舞いを明らかにする必要がある. 特に, 事前に予測するという点は, 能動学習における重要な課題である. そのための手段として, $\mathcal{U}$そのものを対象として分析する事や, ラベル付けを必要とせずに計算可能な UNC 以外の指標を, 最初に選択したデータ$\mathcal{B}_1$に対して計算する事などが考えられる.

また, 他のドメインのデータを用いた場合に, 本研究の結果とどの程度一致するのかは, 重要な問題である. 特に, 類似したドメインである, ツイートを対象とした分類や, ツイート以外の攻撃的な文章の分類に関しての, 能動学習の有効性を明らかにする必要がある.

その他の課題としては, 近年提案された他の能動学習手法の有効性についても検証する必要がある. 同時に, 7 章にて考察を行った点についても明らかにする必要がある.

参 考 文 献

[1] 石坂達也, 山本和英, “Web 上の誹謗中傷を表す文の自動検出”. 言語処理学会 第 17 回年次大会, E1- 6, 2011.

[2] 諏訪光輔, 張建偉, “BERT 及び絵文字を利用した日本語文における皮肉の検出”. DEIM2021, H31-4, 2021.

[3] 松本典久, 上野史, 太田学, “BERT を利用した煽りツイート検出の一手法”. DEIM2021, I14-2, 2021.

[4] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, “Deep batch active learning by diverse, uncertain gradient lower bounds”. In Proceedings of the International Conference on Learning Representations, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding”, NAACL-HLT, pp.4171-4186, 2019.

[6] L. Ein-Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. C. M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim, “Active learning for BERT: an empirical study”. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp.7949-7962, 2020.

[7] R. Dror, S. Shlomov, and R. Reichart, “Deep dominance - how to properly compare deep neural models”. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp.2773-2785, 2019.

[8] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: representing model uncertainty in deep learning”. In international conference on machine learning, pp.1050–1059, 2016.

[9] Y. Gal, R. Islam, and Z. Ghahramani, “Deep bayesian active learning with image data”. In Proceedings of the International Conference on Machine Learning, volume 70, pp.1183-1192, 2017.

[10] D. Gissin and S. S.-Shwartz, “Discriminative active learning”. arXiv:1907.06347v1, 2019.

[11] K. Margatinay, G. Vernikosz, L. Barraulty, and N. Aletrasy, “Active learning by acquiring contrastive examples”. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp.650-663, 2021.

[12] O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set Approach”. In Proceedings of the International Conference on Learning Representations, 2017.

[13] B. Settles, M. Lease, and B. C. Wallace, “Active learning literature survey”. Computer Sciences Technical Report 1648 University of Wisconsin–Madison, 2010.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need”. In Advances in Neural Information Processing Systems, 5998–6008, 2018.

[15] M. Yuan, H.-T. Lin, and J. B.-Graber, “Cold-start Active Learning through Self-supervised Language Modeling”. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp.7935-7948, 2020.

[16] M. Zhang and B. Plank, “Cartography active learning”. In Findings of the Association for Computational Linguistics: EMNLP 2021, pp.395–406, 2021.

[17] Y. Zhang, “Active discriminative text representation learning”. 31st AAAI Conference on Artificial Intelligence, 2016.

[18] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, “Active learning with sampling by uncertainty and density for word sense disambiguation and text classification”. In Proceedings of the 22nd International Conference on Computational Linguistics, pp.1137-1144, 2008.