OffensEval データセットを用いたツイートの攻撃性の有無およびターゲット推定

花畑圭佑　　青野雅樹
豊橋技術科学大学　情報・知能工学課程
hanahata@kde. cs. tut. ac. jp,　aono@cs. tut. ac. jp

1. はじめに

近年, Facebook や Twitter を代表とする SNS や YouTube のような動画配信サービスの発達により, 攻撃的な書き込みが急増している. 攻撃的な書き込みは利用者の気分を害し, サービスの満足度を下げる要因になるためフィルタリングをする必要がある. SNS への書き込みは年々増加しており, 手動でのフィルタリングは非常に時間がかかる. また, 作業者に精神的苦痛を与える可能性があるため, 自動フィルタリングの需要が高まっている.

以前の研究では様々な攻撃的文書の分類が行われているが, 攻撃性の有無と攻撃対象の双方を対象にした分類タスクは少ない. 本研究ではアンサンプル学習モデルを用いた文書の攻撃性の有無及びターゲットの高精度な推定に取り組んだ.

2. 関連研究

文書の攻撃性を推定する研究は数多く行われている. Facebook と Twitter の書き込みを明らかに攻撃的, 密かに攻撃的, 非攻撃的の 3 クラスに分類する研究[1]や, ヘイトスピーチの識別[2]などがある. また Kaggle では, 攻撃的なコメントを有毒, より有毒, わいせつ, 脅迫, 侮辱, 個人攻撃の 6 クラスに分類するコンペティションが行われた[3].

SemEval2019 の Task6 では, Twitter の書き込みに対して, 攻撃対象に着目した階層的ラベル付けがされたデータセットを用いてコンペティションが行われた. SemEval2020 Task12 においても同様のタスクが開催される予定である.

3. データセット

本研究では, SemEval2019 Task6 OffensEval にて公開されたデータセットである OLID (Offensive Language Identification Dataset) [4]を使用する. OLID はツイートに対して攻撃性に関する階層的なラベリングがされたデータセットである.

データセットの構造について記述する. まずツイートに攻撃性が存在するかどうかについて, 以下のように分類される. 攻撃性があるものは NOT（Not Offensive）クラス, ないものは OFF（Offensive）クラスと定義される. 次に OFF クラスに分類されたツイートについて, 攻撃対象が存在するかどうかで分類される. 攻撃対象が存在するものは TIN（Targeted Insult）クラス, 存在しないものは UNT（Untargeted）クラスと定義される. 最後に TIN クラスに分類されたツイートについて, 攻撃性が何に向けられているかについて分類される. 個人を対象とした攻撃的ツイートは IND（Individual）クラス, 政党や民族といった団体を対象とした攻撃的ツイートは GRP（Group）クラス, IND クラスと GRP クラスのどちらにも当てはまらないものは OTH（Other）クラスと定義される. 図 1 はデータセットの階層構造を表す図である.

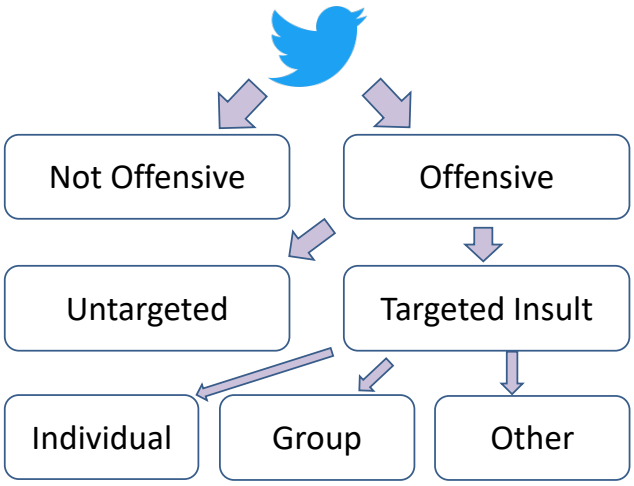

図 1　OLID の階層構造

本研究では上述のデータセットにおける NOT, UNT, IND, GRP, OTH の 5 クラスを分類するモデルを提案する. 各クラスのデータ数は表 1 の通りである. NOT クラスの訓練データが 8840 に対し OTH クラスの訓練データは 395 と, データ数に大きく偏りがあることがわかる.

表 1　データセット総数

	訓練	評価
NOT	8840	620
UNT	524	27
IND	2407	100
GRP	1074	78
OTH	395	35
合計	13240	860

4. 提案手法

以下では我々が提案するツイートの攻撃性およびターゲットを推定するモデルについて説明する.

4.1 ツイートの前処理

全ての入力ツイートに関して前処理を行う. 図2はツイートの一例である.

@USER @USER Go home you're drunk!!! @USER #MAGA #Trump2020 👊🇺🇸👊 URL

図 2 IND クラスのツイート例

ユーザ ID は「USER」,ツイートに添付されていた URL は 「URL」にそれぞれ置き換えられている. ユーザ ID に付随する@, タグ付けに使用される#, 「URL」トークンはツイートを特徴量に変換する際に妨げになると考え, 削除する.「USER」トークンは攻撃対象推定の手がかりになると考え残した. ツイートには絵文字を含むものも多くある. 絵文字に攻撃性や攻撃対象の情報が含まれる可能性がある. emoji ライブラリを使用し絵文字をその絵文字を表すテキストに変換する. 以上の前処理を図 2 のツイートに適用した例が図 3 の通りである.

USER USER Go home you're drunk!!! USER MAGA Trump2020 **punch flag_for_United_States punch**

図 3 前処理を適用したツイート例

4.2 WordNet を用いたデータ拡張

3節にて述べたように, クラス間でデータ数に大きく偏りがある. 訓練データに偏りがあると, データが多いクラスに過剰に分類される, データが少ないクラスに分類されないなどの問題が起こる可能性がある. そこでデータ拡張を行う. 拡張の方針として, 既に存在するツイートの一単語をなるべく意味が変わらないような類義語に変更することで拡張する. 類義語の抽出には WordNet[7]を使用する. WordNet には synset と呼ばれる類義語のグループが定義されている.

まず, ツイートの先頭単語から順に探索し, synsetが存在する単語を発見する. 発見した単語のsynsetから単語をランダムに選択し, その単語を置き換える. synset がある単語が存在しない場合, 文末に「word」トークンを挿入する. 図 3 のツイートに拡張処理を適用した例は以下の通りである.

USER USER **turn** home you're drunk!!! USER MAGA Trump2020 punch flag_for_United_States punch

図 4 元ツイートから拡張されたデータ例

先頭から 3 単語目の「Go」が「turn」に置き換えられていることがわかる.

4.3 BERT を用いた特徴量抽出

文章の攻撃性及び対象の推定においては, 単語毎の攻撃性やその単語がどの単語に係っているかといった情報が重要だと考える. そこで, ツイートを汎用言語モデル BERT[5]を用いて特徴量に変換する.

本研究では事前学習済みモデル「BERT-large Uncased」にツイートを入力して特徴量を得る. 上記モデルは合計 24 の Transformer[6]層で構成される. 層を経るごとに Attention 機構により単語間の関係性が強調された特徴量に変換される.

特徴量抽出の流れについて記述する. まず前処理された入力ツイートを BERT に入力する. 入力単語の最大長を 25 とする. 1 単語は 1024 次元の特徴量に変換され, Transformer 層によって徐々に単語同士の文脈的な特徴が強調されていく. 各層で得られる特徴量は 25×1024 次元となる.

4.4 分類モデル

4.3 節で示した特徴量を以下の 4.4.1 項と 4.4.2 項のモデルに入力し, 双方の出力を用いたアンサンブルモデルを 4.4.3 項で述べる.

4.4.1 多層パーセプトロン(MLP)を用いたモデル

BERT 特徴量を多層パーセプトロンに入力し, 分類モデルを構築する. BERT モデルの, より単語間の関係性が強調されていると考えられる 23 層目の出力を抽出する. 24 層目の出力を使用しない理由は, BERT の事前学習タスクに偏った特徴量が抽出されると考えられるためである. 得られた 25×1024 次元の特徴量の全体の平均をとり, 1024 次元の特徴量を多層パーセプトロンに入力し, 学習する.

4.4.2 Bidirectional-LSTM を用いたモデル

BERT 特徴量を Bidirectional-LSTM に入力し, 分類モデルを構築する. 単語の系列的な特徴を重視

した分類を行うため，単語そのものの埋め込み表現に近い特徴量である， BERT モデルの 1 層目の出力を抽出する．得られた 25×1024 次元の特徴量を系列データとして Bidirectional-LSTM に入力し，学習する．

4.4.3 アンサンブルモデル

ツイートを 4.4.1 項と 4.4.2 項にて構築した二つのモデルに入力し，それぞれクラス確率を得る．得られたクラス確率同士の要素積を取ることでアンサンブルし，高い精度で推定を行う．図 5 に概要を示す．

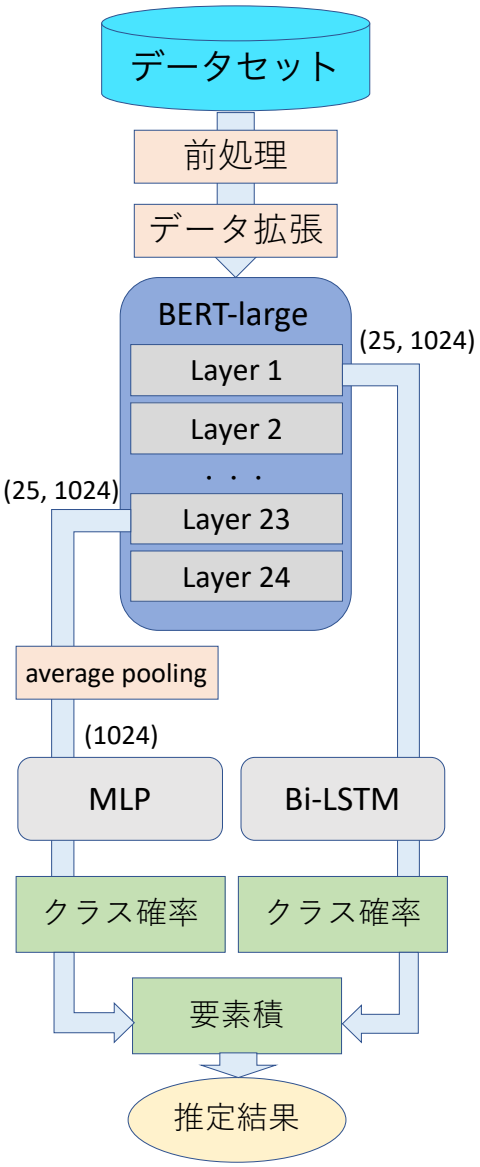

図 5　提案手法の概要図

5.　評価実験

本研究では 3 節で述べた, SemEval2019 Task6 のデータセットを使用した．また, 4.2 節にて述べた手法を訓練データが特に少ないと考えられる UNT, GRP, OTH クラスの訓練データに適用し，データ数を 2 倍に拡張した． データセットの詳細について以下に示す．

表 2　データ拡張処理後のデータセット総数

	訓練(拡張)	評価
NOT	8840	620
UNT	1048 (524)	27
IND	2407	100
GRP	2148 (1074)	78
OTH	790 (395)	35
合計	15233(1993)	860

5.1　評価指標

評価データの数がクラス間でアンバランスであるため，評価指標にはクラスごとに F1 値を計算し全体の平均をとる macro-F1 値を使用する．

5.2　ベースライン

ベースラインとして，ツイートから Bag of Words 特徴量を抽出し，得た素性を多層パーセプトロンに入力し学習する分類モデルを使用する. Bag of Words 特徴量については語彙数を 1000 次元とする．図 6 に概要図を示す．

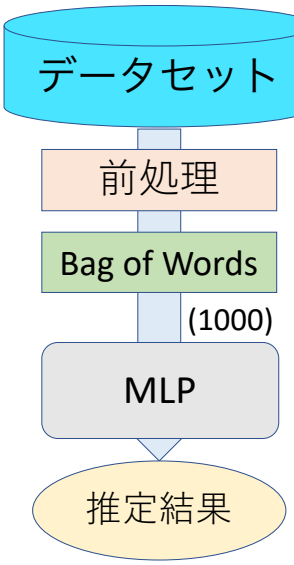

図 6　ベースラインの概要図

5.3　実験結果

ベースライン, MLP モデル, Bi-LSTM モデル，提案手法，それぞれ拡張データを使用して実験を行なった結果を以下に示す．

表 3　各手法の macro-F1 値

手法	macro-F1
Baseline①	0.3459
MLP②	0.3745
Bi-LSTM③	0.3802
MLP+Bi-LSTM④	0.4124
MLP(augmented)⑤	0.4195
Bi-LSTM(augmented)⑥	0.3964
MLP+Bi-LSTM(augmented)⑦	**0.4209**

表 4　各手法のクラス毎の F1 値

	①	②	③	④	⑤	⑥	⑦
NOT	0.82	**0.87**	0.85	0.86	0.86	0.84	0.86
UNT	0.22	0.24	0.33	**0.40**	0.29	0.28	0.24
IND	0.32	**0.48**	0.35	0.38	0.47	0.42	0.46
GRP	0.24	0.23	0.27	0.32	0.31	0.36	**0.41**
OTH	0.06	0.05	0.10	0.09	**0.16**	0.08	0.14

5.4　考察

データ拡張の有効性について考察する. MLP モデル, Bi-LSTM モデル, アンサンブルモデル全てでデータ拡張によって精度が向上している. 特に MLP モデルに関しては, データ拡張を行った UNT, GRP, OTH 全てのクラスに関して精度が向上した. macro-F1 値の向上幅も 3 モデル中大きく, データ拡張が有効であったと言える.

最もよい精度が出た, データ拡張を行ったアンサンブルモデルについて考察する. 他のモデルと比較して, macro-F1 値, GRP クラスに関して最もよい精度が出ている. データ拡張を行わないモデル④では正解できなかったが, データ拡張を行ったモデル⑦では正解することができたテストデータの例を図 7 に示す. 図 7 はアメリカの共産主義者を批判する GRP クラスの攻撃的ツイートである.

#JoinTheFight Join the fight against American Communists... Democrats and Liberals have taken this crap to the edge and there is no coming back. They are openly promoting the Destruction of AMERICA. #JoinTheFightAgainstCommunism URL

図 7　モデル⑦での成功例

一方で, UNT クラスでは精度が他と比べ低くなってしまっている. データ拡張をしていないアンサンブルモデルではよい精度が出ているため, UNT クラスのデータ拡張手法について見直す必要があると考えられる.

OTH クラスについては, データ拡張による精度の向上は確認できたものの, 他クラスと比べ精度が特に低いことは今後の課題と言える. OTH クラスのデータ拡張手法をより工夫するなどが考えられる.

6.　おわりに

本研究では, データ拡張の適用及び BERT の異なる層の特徴量を用いたツイートの攻撃性の有無及びターゲットの推定モデルを提案した. 評価実験ではデータ拡張の有効性, アンサンブルによる一部クラスでの精度の向上が確認できた.

今後の課題として, より効果的なネットワークやアンサンブル手法の検討, 正解率の低い UNT クラス, OTH クラスに対するアプローチの考案等が考えられる.

謝辞

本研究の一部は, 科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した.

参考文献

[1] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC).

[2] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2):223–242.

[3] Kaggle Toxic Comment Classification Challenge https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[4] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of NAACL.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT:Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[7] Princeton University "About WordNet." WordNet. Princeton University. 2010, http://wordnet.princeton.edu