D-011

投稿における炎上推定モデル生成のための疑似学習データセット構築の検討
A Pseudo Dataset for Training a Model for Estimating Flames in SNS Posts

吉田 裕輔† 鷹野 孝典‡
Yusuke Yoshida Kosuke Takano

1. はじめに

SNS[1][2][3]の普及による，個人の情報発信機会が増える一方で，SNS に投稿された意見やコメントに対する炎上現象が社会問題となっている．炎上の社会的影響としては，企業であれば，イメージ低下，株価暴落，事業撤退などのリスクがある．個人では，現住所，家族構成，友人関係の特定，さらに進学，内定の取り消しなどのリスクがある．

SNS に投稿された意見やコメントに対する炎上可能性の推定のために，多様な要因に対応した炎上推定モデルの構築が重要である．炎上要因の一つとして，投稿者の考えと世間の考えのズレが大きく異なる乖離度が挙げられるが，乖離度計算には，時系列で展開される意見間の対立度を必要とする．このため，本研究では特に多様な視点の意見集合を対象として，意見間の対立度に着目したニューラルネットワークモデル学習のための，疑似学習データセットの構築手法について提案する．

2. 炎上推定モデル構築のための疑似学習データセット

2.1 炎上推定モデル

炎上検出手法として，特定の炎上要因を絞った手法が提案されている[4][5][6]．しかし，炎上の要因は多岐にわたるため，一つの要因に着目するだけでは炎上推定は困難である．本研究では，炎上要因の乖離度，表現，話題性，話題の伝播速度の 4 つの要因に着目した推定モデル(図 1)の構築を目指している．図 1 において，「乖離度」は SNS 投稿者の考えと世間の考えのズレの大きさ，「表現」は差別表現，中傷表現などの炎上を引き起こす要因となる表現，「話題性」は社会での話題のインパクト，「話題の伝播速度」は SNS での情報の伝わりやすさを表している．

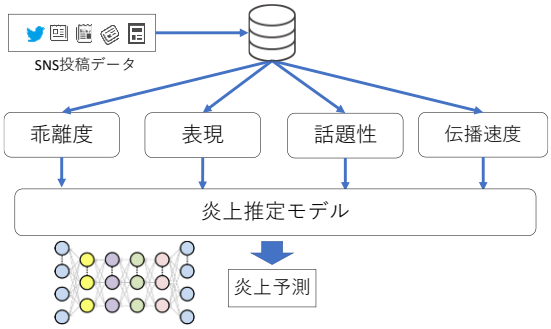

図 1 複数の要因に着目した炎上推定モデル

†神奈川工科大学大学院工学研究科情報専攻，Graduate School, Kanagawa Institute of Technology
‡神奈川工科大学情報学部情報工学科，Department of Information and Computer Sciences, Faculty of Information

2.2 疑似学習データセット

提案する疑似学習データセットは，意見集合，各意見の分類情報，意見の時系列情報，意見間の対立度で構成される(表 1).

表 1 疑似学習データセット

項目	説明
意見集合	SNS に投稿される意見の集合
意見の分類情報	意見を n 個のグループに分けた情報
意見の日時情報	意見が投稿された日時
意見間の対立度	意見間の相違による対立度合い

対立度は，文章の組に対して n 段階整数で定義する．例えば，1～5 の 5 段階として，対立する可能性が高ければ対立度に高い数値を設定する．なお，対立度は意見の乖離度計算に用いられる．乖離度計算は，一定期間の対立度の平均に標準偏差の 3 倍を加えた 3σ 法[7]で乖離度を測る方法などがある． SNS で投稿される意見は，時系列で展開されるため，全ての意見に対して日時情報を付与する．

提案する疑似学習データセットのスキーマを表 2，3 に示す．また，意見データ，対立度データの例を表 4，5 に示す．表 6 に示すように，分類情報の組の対立度から，各分類に属する意見の対立度を付与することができる．

表 2 意見スキーマ

{(ID, INT), (意見, TEXT), (分類情報, INT), (日時, TIMESTAMP)}

表 3 対立度スキーマ

{(分類情報 A, INT), (分類情報 B, INT), (対立度, INT)}

表 4 意見データの例

ID	文章	分類情報	時刻
1	地震が起きたら、自分や家族を守るために、地震が起きたときの対処法を知っておこう。	8	2020-4-10 10:15:32

表 5 対立度データの例

分類情報 A	分類情報 B	対立度
1	8	1

表 6 意見データの組への対立度の付与

文章 A	文章 B	対立度
何の前触れもなく、こんなことが起こるなんて、本当に怖いです。	地震が起きたら、自分や家族を守るために、地震が起きたときの対処法を知っておこう。	1

3. 疑似データセット作成手順

疑似学習データセットは，下記手順で作成される．

Step-1: 文章データの準備
Step-2: 文章ベクトル抽出
Step-3: クラスタリングによる分類情報の付与
Step-4: 対立度の付与
Step-5: 時系列情報の付与

Step-1 での文章データの収集方法として，SNS の投稿データを取得する方法などがある．しかし，SNS で提供される API において投稿データの取得制限がある場合や炎上した投稿は削除される場合も多い．このため，本研究では GPT-3(Generative Pre-trained Transformer 3)[8]などの大規模言語モデルを用いた疑似投稿データの生成を行う．Step-2 では，BERT(Bidirectional Encoder Representations from Transformers)[10]などで学習される埋め込み表現を用いて，Step-1 で収集した文章集合を文書ベクトルとして抽出する．さらに，Step-3 では，Step-2 で抽出した文章ベクトル集合を対象として，k-mean 法[9]などのクラスタリング手法を適用して k 個のクラスへに分類する．クラスタ数はエルボー法[11]などを用いて算出することができる．Step-4 において，各文章に対立度を付与する．まず，2 つのクラスタの組 A と B を抽出し，対立度を n 段階整数で定義する．この対立度の適宜を全クラスタのすべての組を対象として行う．次に，文章集合における各文章間の対立度を，各文章が属するクラスタ間で設定したものと同じ対立度を用いて付与する．最後に Step-5 において，時間順序の整合性を考慮して各文章に字形悦情報を付与する．

4. 実験

3 章の Step-1 から Step-3 の実現可能性について検証する．Step-1 の文章収集のために，GPT-3 モデル“text-davinci-003”を利用し，1,048 件の文章を生成した(表 7)．また，Step-2 に関して 1,048 件の文章から文章ベクトルを生成するために，事前学習済みの BERT モデル“bert-base-japanese-whole-word-masking”[12]を用いた．

Step-3 の文章集合のクラスタリングにおいて，クラスタ数を決めるためにエルボー法を用いた．エルボー法の結果を図 2 に示す．エルボー法の結果より分類数を 10 個に分けた場合の各クラスタの特徴と文章数を表 8 に示す．また，各クラスタにおける文章類似度をコサイン尺度で計算した結果，全クラスタの類似度平均は，0.78～0.92 であった．

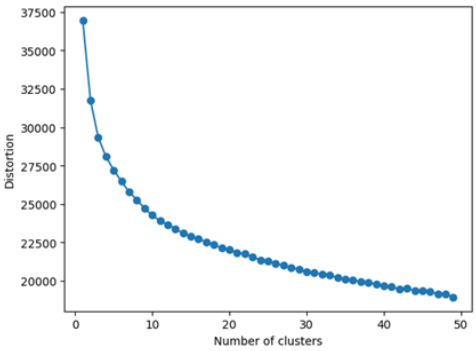

図 2 エルボー法の結果

表 7　生成文章の例

ID	文章
43	地震で残された破壊の数々を見て心が痛みます。
109	地震と津波による破壊は信じられないほどです。家や会社の跡を見ると心が痛みます。

表 8　分類情報の特徴

分類 ID	分類の特徴	文章数
0	被災地や被災者の事を心配する	175
1	恐怖を感じた感想	177
2	地震による被害が想像を超えていた	79
3	被災者にかけるお見舞い	80
4	エクスクラメーションマークを使用	114
5	地震の被害の大きさについて	141
6	被災者への祈り	85
7	心が痛む気持ち	62
8	確認を促進するコメント	86
9	備えや団結する必要性	111

5. 考察とまとめ

表 7，表 8 の結果より，大規模言語モデルを用いた場合でも，ある話題に対して多様な視点の意見集合を生成可能であることの見込みを得ることができた．

今後は，文章データの規模を大きくするとともに，疑似学習データセットを公開できるように整備していく予定である．さらに，本学習データセット適用し，SNS に投稿された意見やコメントに対する炎上可能性の推定のために，意見間の対立度に着目した学習による炎上推定モデルを構築し，実現可能性を検証していく．モデルの学習方法としては，意見の組と対立度を入力とした Transformer 構造を持つニューラルネットワークを用いた学習，あるいは，既存の大規模言語モデルを用いたプロンプト上での学習などを検討している．

参考文献

[1] Twitter, https://twitter.com/.
[2] Facebook, https://www.facebook.com/ .
[3] Instagram, https://www.instagram.com/.
[4] 堺 雄之介, 竹内 幹太, 伊東 栄典, “コメントを利用した炎上動画検出に関する検討”, 情報処理学会研究報告, pp 1-5 (2021).
[5] 三宅 剛史, 松本 和幸, 吉田 稔, 北 研二, “分散表現を用いた有害表現判別に基づく炎上予測”, 人工知能学会第二種研究会資料 (2017)
[6] 川上 幹, 彌冨 仁, “Twitter への投稿テキストによる炎上警告システムの構築”, 日本知能情報ファジィ学会 ファジィ システム シンポジウム, pp 705-708 (2016)
[7] 上本道久, “検出限界と定量下限の考え方”, ぶんせき, 2010 5, pp. 216-221 (2010).
[8] Brown, T. B. et al. “Language Models are Few-Shot Learners” Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 1877-1901 (2020).
[9] 宮本 定明, “クラスター分析入門　ファジィクラスタリングの理論と応用”, 森北出版株式会社 (1999)
[10] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *arXiv preprint arxiv:1810.04805* (2018).
[11] scikit-learn: machine learning in Python, “sklearn.cluster.KMeans”, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html, (最終閲覧日 2023/06/15)
[12] “bert-base-japanese-whole-word-masking”, https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking