対象を考慮した日本語「愚痴」データセットの構築

伊藤和浩　村山太一　矢田竣太郎　若宮翔子　荒牧英治
奈良先端科学技術大学院大学
{ito.kazuhiro.ih4,murayama.taichi.mk1,s-yada,wakamiya,aramaki}@is.naist.jp

概要

愚痴（complaints）は，現実が自分の期待に反しているときに行われる基本的な言語行為である．愚痴に関連するデータセットはいくつか構築されているが，愚痴が向けられている対象についてのラベルを含むデータセットはまだ存在しない．本研究では，Twitter から収集したテキストに対して対象ラベルの付与を含むアノテーションを施した愚痴データセットを構築し，分析した．このデータセットを用いて愚痴テキストの分類の実験を行ったところ，テキストが愚痴かどうかを識別する二値分類タスクで Accuracy 90.9，愚痴の対象ラベルを当てるマルチクラス分類タスクで Accuracy 65.4 を達成した．

1　はじめに

愚痴（complaints）は人間の基本的な言語行為であり [1]，語用論においては「ある状態，製品，組織，または出来事に対して，現実と期待との間の否定的な不一致を表現するために用いられる基本的な発話行為」と定義されている [2]．愚痴を分析することは，人間の基本的な振る舞いの理解に寄与するだけでなく，企業の商品やサービスの改善に役立つなど様々な利点がある．そのため，心理学 [3] や言語学 [4]，マーケティング [5] などの分野で，愚痴について多くの研究が行われてきた．

近年，自然言語処理分野においても愚痴に関連する研究が行われている．例えば，Twitter の企業アカウントへ寄せられた投稿にサービスカテゴリ（食品，車，電気製品など）のラベルを付与した complaints データセットが構築されている [6]．また，Trosborg による区分け [7] に基づき complaints の深刻度を 4 つのラベル (No explicit reproach, Disapproval, Accusation, Blame) で付与した不満データセットも構築されている [8]．他にも，商品・サービスに対する不満を収集する不満買取センターのデータに基づくコーパスが構築されている [9]．不満の対象がラベルには含まれるが，データの性質上，対象は商品・サービスに限定されている．以上のように，愚痴に関連するテキストを収集したデータセットやコーパスはいくつか構築されているものの，一般的な愚痴の対象ラベルが付与されたデータセット構築は行われていない．

愚痴の話題や深刻度を識別すること（「何を」についての情報）が愚痴の性質の解明に役立つのと同様に，愚痴が向けられている対象を明らかにすること（「誰に」についての情報）も愚痴の分析に有用であると考え，本研究では，愚痴の対象を付与したデータセットを構築する．具体的には，Twitter のツイートテキストに，愚痴か否かのラベルと，愚痴の対象を表す 5 つのラベルを付与した愚痴データセットを構築・分析する．さらに，構築した愚痴データセットを用いて愚痴テキストの分類の実験を行う．

2　データセット

2.1　データ収集

本研究では，Twitter の日本語ツイートを用いてデータセットを構築した．Twitter ではユーザが愚痴を吐く行動がよく見られる．さらに，愚痴と関連する abuse [10] や hate speech [11]，offensive language [12] や complaints [6] などのデータセット構築にも Twitter のツイートが利用されている．

まず，TwitterAPI を用いて，2006 年 3 月 26 日〜2021 年 9 月 30 日までの「#愚痴」を含む 64,313 件のツイートを収集した．次に，URL を含むものやテキストが重複するもの，リツイート，そして bot によるツイートを除外した．なお，bot によるツイートの判定は，投稿時に使用されたアプリケーションにより行った．具体的には，Twitter for iPad, Twitter for iPhone, Twitter Web App, Twitter Web Client, Keitai Web 以外からのツイートは bot によるツイートである可能性が高いとみなし，全て除外した．その後，残りのツイートテキスト（以降，テキスト）から全ての

表 1　対象ラベルごとのテキスト例

ラベル	テキスト
SELF-SPECIFIC	しかしたぶん全部顔とか行動に出ちゃってるから最低なのは自分なんだよね 向こうには落ち度はないし 勝手に苛ついてるだけだしね
SELF-COLLECTIVE	海外の方でさえ、ありがとうございますってしっかり言ってくれるっつーのに、日本人ときたら…会釈すらしない人ざらだよね
OTHER-SPECIFIC	母親が私の話無視するから私も母親の話無視したらめっちゃ怒られたんだけど。自分はよくて相手はダメなんか？
OTHER-COLLECTIVE	電車で人が降りてる時くらいみんなスマホ見るの止めようぜ　降りる側からしたら普通に邪魔だわ
NON-HUMAN	明日大雪とかやんなっちゃうよ　こんな日に限って仕事忙しいし。最低だわ。憂鬱で仕方ない。リスパダール飲もう。

ハッシュタグを削除し，30 文字以下のツイートを除外した．最後に，投稿月ごとの層別サンプリングを行い，7,573 件のテキストを取得した．

2.2　アノテーション

7,573 件のテキストに対して愚痴ラベルと対象ラベルのアノテーションを行った．本研究で作成したアノテーションガイドラインに従って，3 名のアノテーターが，それぞれ 2,524 件または 2,525 件のテキストに対して，以下の手順でラベルを付与した．

First Stage　テキストが愚痴であるか，そうでないかを識別する．愚痴である場合は POSITIVE ラベル，そうでない場合は NEGATIVE ラベルを付与する．

Second Stage　POSITIVE ラベルが付与されたテキストについて，愚痴の対象を分類する．発信者を含む場合を SELF，発信者を含まない場合を OTHER，特定の人（1 人〜2，3 人）に向けられている場合は SPECIFIC，集団の人々に向けられている場合は COLLECTIVE とし，これらの組み合わせで定義された 4 つの分類ラベルを付与する．これら 4 つのラベルに加え，対象が人間以外の場合は NON-HUMAN ラベルを付与する．さらに，対象が一意に定まらない場合や不明瞭な場合は NEGATIVE ラベルを付与した．

- SELF-SPECIFIC: 発信者自身
- SELF-COLLECTIVE: 発信者を含む集団
- OTHER-SPECIFIC: 特定の他者
- OTHER-COLLECTIVE: 発信者を含まない集団
- NON-HUMAN: 人間以外（動物，天候，概念など）

アノテーションの結果，7,573 件のテキストのうち，6,418 件には POSITIVE，残りの 1,155 件には NEGATIVE のラベルが付与された．また，愚痴の対象ラベルごとのデータ数は，SELF-SPECIFIC が 426 件，SELF-COLLECTIVE が 42 件，OTHER-SPECIFIC が 3,866 件，OTHER-COLLECTIVE が 648 件，NON-HUMAN が 1,436 件であった．SELF-COLLECTIVE のテキスト数が他のラベルと比べて少ない理由として，自分を含む集団が愚痴の対象になる場面では，自分（SELF-SPECIFIC）か他人（OTHER-SPECIFIC または OTHER-COLLECTIVE）のいずれかのみを対象として愚痴が表出される傾向がある可能性が考えられる．対象ラベルごとの例を表 1 に示す．

表 2　ラベルごとの文字数の統計量

ラベル	平均値	中央値	標準偏差
POSITIVE	82.0	81.0	32.8
SELF-SPECIFIC	75.8	72.5	32.2
SELF-COLLECTIVE	87.2	86.0	30.6
OTHER-SPECIFIC	83.2	83.0	32.4
OTHER-COLLECTIVE	87.8	89.0	32.5
NON-HUMAN	77.8	74.0	33.8
NEGATIVE	68.3	62.0	32.1

3 名のアノテーション結果のサンプルと，異なる 1 名のアノテーターによる結果との一致率を Cohen's Kappa により求めた．その結果，二値分類（POSITIVE または NEGATIVE）では 0.798，6 ラベル分類（愚痴の対象 5 ラベルまたは NEGATIVE）では 0.726 となり，十分に信頼できる値となった．

2.3　データセット分析

2.3.1　文字数

ラベルごとの文字数の統計量を表 2 に示す．データセット全体の文字数の平均は 79.9 字，中央値は 78.0 字であった．最も文字数が少なかったラベルは SELF-SPECIFIC（平均 75.8 字，中央値 72.5 字），最も多かったラベルは OTHER-COLLECTIVE（平均 87.8 字，中央値 89.0 字）であった．自身に関する愚痴は状況説明が比較的少ない一方，他の集団に関する愚痴は説明的になる傾向が一因であると推測される．

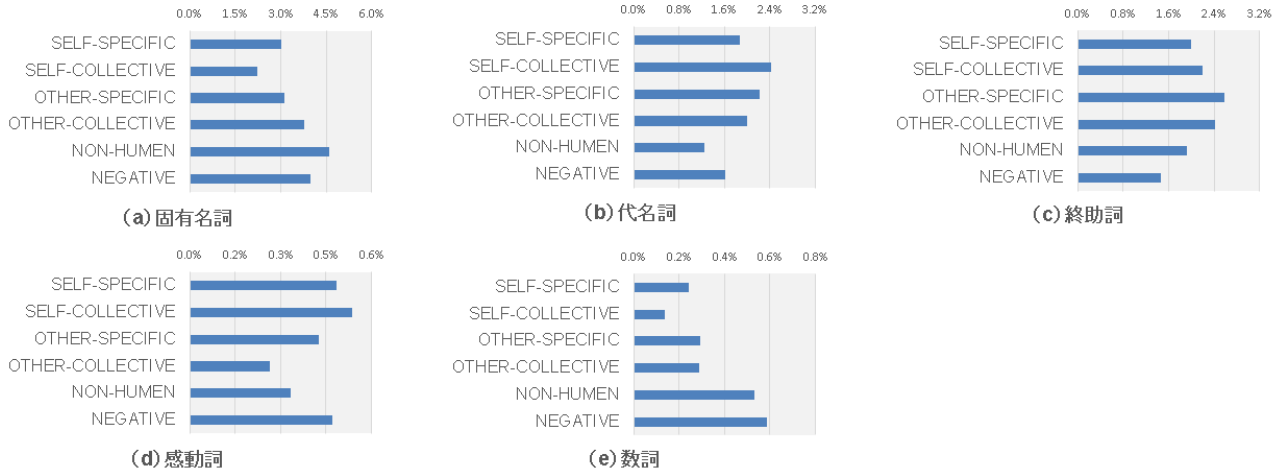

図 1　ラベルごとに集計した，総語数に対する品詞ごとの割合

さらに，POSITIVE の方が NEGATIVE に比べて文字数が多かった（POSITIVE は平均 82.0 字，NEGATIVE は平均 68.3 字）．これは対象が不明瞭なものは NEGATIVE とするアノテーションルールにより，説明がほとんどない短い投稿が NEGATIVE に振り分けられたためであると推測される．

2.3.2　品詞

ラベルごとの品詞の傾向を分析するために，総語数に対する品詞ごとの割合を求めた．形態素解析には MeCab[1)] を用いた．図 1 に集計結果を示す．特徴的な傾向が見られた結果について議論する．

固有名詞の割合（図 1a）は，NON-HUMAN や NEGATIVE で相対的に高い一方，SELF-SPECIFIC, SELF-COLLECTIVE や OTHER-SPECIFIC で低い．自分や特定の他者に関係する愚痴は，対象を固有名詞で明言することを避け，隠す傾向があると推測できる．固有名詞（太字）を含む NON-HUMAN の例を以下に示す．

- **アクションテレビ**って**スカパー**のチャンネル，日本語吹き替えのを流せや！
- 皆さんも知っていると思いますが，**大阪府**の高校の臨時休校延長決まった～今年就活やのに,,, 動き出すんが遅くなるやん

代名詞の割合（図 1b）は，SELF-COLLECTIVE や OTHER-SPECIFIC で高く，NON-HUMAN で低い．対象が人ではない NON-HUMAN の場合，代名詞を用いずに対象を明示する傾向が読み取れる．代名詞（太字）を含む OTHER-SPECIFIC の例を以下に示す．

- いち社会人としてあのメール文は間違ってると思うんだけど．**お前**て w 来いて w そんなん言われてわかりましたってなるかよ．**テメェ**が来い

終助詞（～ね，～よ，～かしら，～もん，など）の割合（図 1c）は，OTHER-SPECIFIC と OTHER-COLLECTIVE で高く，NEGATIVE では低い．POSITIVE 全体では 2.38 %を占めるのに対し，NEGATIVE においては 1.45 %に留まることから，終助詞は愚痴であることを示す重要な特徴量になると考えられる．終助詞（太字）を含む OTHER-SPECIFIC の例を以下に示す．

- わたしが居ないとミルクしまってある場所すらわかんないのか**よ**
- 病んでる自慢してかまってもらえて可愛い女の子はいいです**ねー**．ブスはかまってもらえないのです**よー**

感動詞（すみません，あーあ，ww，など）の割合（図 1d）は，SELF-SPECIFIC, SELF-COLLECTIVE, OTHER-SPECIFIC で高く，OTHER-COLLECTIVE, NON-HUMAN で低い．自身と直結する対象の場合はより主観的な表現になり，感動詞も増加する傾向があると考えられる．感動詞（太字）を含む OTHER-SPECIFIC の例を以下に示す．

- **ほんと**さ，意味わかんないよね携帯いじるなら家でもいいのになんでうちの家来る必要あるの？言っとくけどそんな暇じゃないよ我
- しつこい．**あぁ**しつこい．わたしも断り方わるいんだろな．**あぁ**しつこい…

数詞の割合（図 1e）は，感動詞と逆の傾向が見ら

1）https://taku910.github.io/mecab/

れ，NON-HUMAN で比較的高く，SELF-SPECIFIC や SELF-COLLECTIVE で低い．感動詞と関連して，自身と対象との心理的な距離が大きい場合，客観的な事実を含む愚痴になりやすいことが示唆される．数詞（太字）を含む NON-HUMAN の例を以下に示す．

- カップ式自販機で、**10** 円つり銭切れが表示されていたので、**130** 円投入し、**80** 円のコーヒー買ったらお釣り **50** 円全部 **10** 円玉で出やがった事

3 実験

構築した愚痴データセットを用いて，愚痴か否かを識別する二値分類タスク（アノテーションの First Stage に対応）と，愚痴の対象ラベルを予測するマルチクラス分類タスク（アノテーションの Second Stage に対応）を実施した．

3.1 設定

データセットの 70%を訓練データ，15%を検証データ，15%をテストデータに分割した．機械学習モデルには，Long-Short Term Memory (LSTM) [13] と Bidirectional Encoder Representations from Transformers (BERT) [14] の 2 種類を用いた．なお，BERT モデルは，東北大学が公開している日本語版 Wikipedia で事前学習を行ったモデル[2)]をファインチューニングしたものを採用した．検証データでの結果に基づき，各モデルに対して下記のパラメータを採用した．LSTM モデルについて，単語埋め込みの次元数は 10，隠れ層のレイヤー数は 128，損失関数はクロスエントロピー，最適化手法は Stochastic Gradient Descent (SGD)，学習率は 0.01，エポック数は 100 とした．BERT モデルについて，ツイートごとの最大トークン数は 128，バッチ数は 32，最適化手法は Adam，学習率は 1e-5，エポック数は 10 とした．

3.2 結果と考察

3.2.1 二値分類タスク

二値分類タスクの結果は，LSTM モデルでは Accuracy 85.7 %，F 値 91.7 %，AUC 69.8 %となり，BERT モデルでは Accuracy 90.9 %，F 値 94.8 %，AUC 73.0 %となった．スコアの高かった BERT モデルの混同行列のテストデータ件数に対する割合はそれぞれ True Positive 84.0 %，False Positive 7.3 %，False Negative 1.8 %，True Negative 6.9 %となり，エラーの中では False Positive が比較的多い結果となった．

2) https://github.com/cl-tohoku/bert-japanese

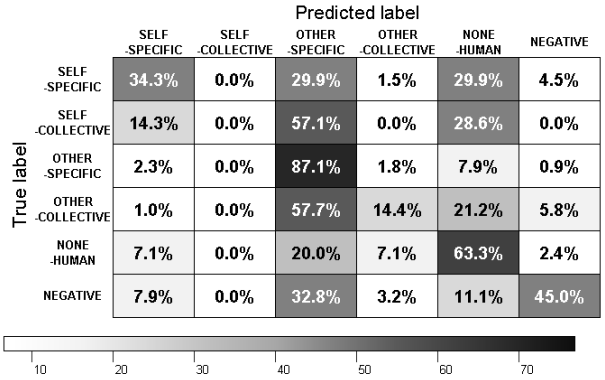

図 2 マルチクラス分類タスクにおける BERT モデルの結果

3.2.2 マルチクラス分類タスク

マルチクラス分類タスクの結果，LSTM モデルの Accuracy は 47.6 %，BERT モデルの Accuracy は 65.4 %となった．スコアの高かった BERT モデルについてラベル同士の識別結果を図 2 に示す．SELF-COLLECTIVE に分類されたツイートは存在しない一方で，OHTER-SPECIFIC や NON-HUMAN に多くのツイートが分類された．これはデータセットの中のラベルごとのツイート数の偏りが反映されていると考えられる．ラベルごとの偏りに関する改善案としてデータ数の少ない SELF-COLLECTIVE のラベルを SELF-SPECIFIC と統合することや，層別サンプリングしたものをモデルの学習に利用することなどが挙げられる．

4 おわりに

本研究では，愚痴の対象ラベルが付与されたデータセットを構築し，内容の分析および機械学習モデルによる分類の実験を行った．実験の結果，BERT モデルにおいて，愚痴か否かを識別する二値分類タスクでは Accuracy 90.9，愚痴の対象を識別するマルチクラス分類タスクでは Accuracy 65.4 であった．

今後の展開として，ソーシャルメディア上の特定の集団内での愚痴の量を測定し，異なる集団や時期ごとの比較分析を想定している．また，ソーシャルメディアに限らず，職場の日報テキストなどの材料から愚痴やその対象を特定し，ウェルビーイングとの関連を調査する研究への展開も検討している．

謝辞

本研究の一部は，JST，AIP-PRISM，JPMJMI21J2の支援を受けたものである．

参考文献

[1] John Langshaw Austin. **How to do Things with Words**. Oxford University Press, 1975.

[2] Elite Olshtain and Liora Weinbach. 10. complaints: A study of speech act behavior among native and non-native speakers of hebrew. In **The Pragmatic Perspective: Selected papers from the 1985 International Pragmatics Conference**.

[3] Robin M. Kowalski. Complaints and complaining: functions, antecedents, and consequences. **Psychological bulletin**, Vol. 119 2, pp. 179–96, 1996.

[4] Camilla Vásquez. Complaints online: The case of tripadvisor. **Journal of Pragmatics**, Vol. 43, No. 6, pp. 1707–1717, 2011. Postcolonial pragmatics.

[5] Chul min Kim, Shinhong Kim, Subin Im, and Changhoon Shin. The effect of attitude and perception on consumer complaint intentions. **Journal of Consumer Marketing**, Vol. 20, pp. 352–371, 2003.

[6] Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. Automatically identifying complaints in social media. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5008–5019, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Anna Trosborg. **Interlanguage Pragmatics: Requests, Complaints, and Apologies**. De Gruyter Mouton, 2011.

[8] Mali Jin and Nikolaos Aletras. Modeling the severity of complaints in social media. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2264–2274, Online, June 2021. Association for Computational Linguistics.

[9] Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, and Tomoya Mizumoto. Fkc corpus : a japanese corpus from new opinion survey service. In **In proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results**, pp. 11–18, Portorož, Slovenia, 2016.

[10] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In **Proceedings of the NAACL Student Research Workshop**, pp. 88–93, San Diego, California, June 2016. Association for Computational Linguistics.

[11] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In **Proceedings of the 13th International Workshop on Semantic Evaluation**, pp. 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[12] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In **Proceedings of the 13th International Workshop on Semantic Evaluation**, pp. 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

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