ネガティブ評判情報に特化したコーパスの構築と分析

三澤賢祐† 田内真惟人† Mathieu Domoulin† 中島正成† 水本智也‡
†エン・ジャパン株式会社 ‡東北大学
{kensuke_mitsuzawa, maito_tauchi,
domoulin_mathieu, masanori_nakashima}@en-japan.com
tomoya-m@ecei.tohoku.ac.jp

1 はじめに

テキストマイニングの分野では、Webから獲得した言語資源（Web言語資源）を利用する研究が増加している。特に、文書分類 [1, 9]、評判分類 [11]、意見分析 [6]、著者推定 [3, 4] のタスクではWeb言語資源の利用が活発に行われている。

しかし、Web言語資源を利用するには、実用上の問題がある。第一に、Web言語資源はノイズが多く、研究目的に合わせて、目的のデータだけを抽出する作業を含めた前処理が必要である。第二に、Web言語資源は、投稿者や文書作成者の性別や年齢など、属性データが利用できないことが多い。属性データが利用可能であったとしても、フォーマットがデータごとに異なるなどの問題が存在する。こうした問題に対処するため、Web言語資源を利用する前に、人手で前処理とラベル付けをしなければいけない状況がしばしば発生する [5]。

本論文では、こうしたWeb言語資源に存在する問題に対処したFKCコーパスを紹介する[1]。FKCコーパスは、一般消費者からの意見を収集するサービスである不満買取センター[2] (FKC) で収集された意見より作成されている。FKCでは、一般消費者から商品やサービスに関する意見を収集するため、ユーザーからの意見の「買い取り」[3]を行なっている。[4] また、FKCでは収集した意見を活用し、製品やサービスを改善するため、図1のような分析ダッシュボード[5]や分析レポート[6]の販売を行なっている。

NLPタスクに応用するにあたり、FKCコーパスが他の類似コーパスより優れている点が3点ある。第一に、FKCコーパスには、投稿の属性データと投稿者のプロフィールデータが存在している。第二に、FKCコーパスでの投稿は、「ネガティブな意見」という点ではっきりとしているため、他のWeb言語資源よりもノイズが少ないデータセットと考える。第三に、不満買取センターは実際に運営されているサービスであるため、投稿が増えると共にコーパスサイズも増量可能である。したがって、時系列データを必要とするような研究や、ツールの開発が可能である。以上の3点により、我々はFKCコーパスがNLPタスクに大きく寄与するデータであると考える。

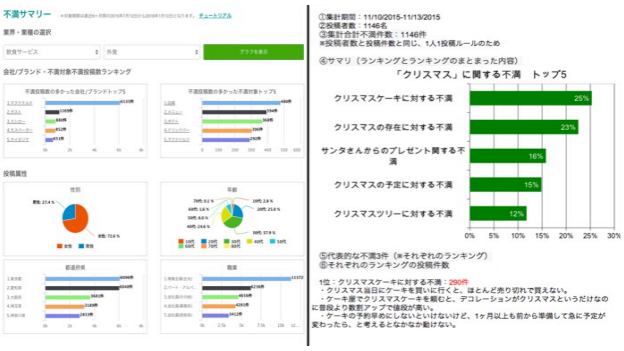

図1: 不満買取センターが提供している分析サービス 左:分析ダッシュボード 右:分析レポート

2 類似データセット

2.1 Twitter

TwitterはNLPタスクで利用されることの多いデータソースである [8, 7]。

しかし、Tweetには属性データが少なく、また投稿者のプロフィールデータも利用できないことが多いため、Twitterから作成したデータセットの利用には前処理、ラベル付けといった手がかかる作業が発生する。一方、FKCコーパスは、Twitterで作成されたコーパスよりもサイズは小さいが、一方で「不満の意見」という意味でデータセットの中身ははっきりとしている。

また、Twitterではハッシュタグが利用可能であるが、一方でハッシュタグが存在しない投稿が多く、ハッシュタグはユーザーが自由に生成できるため、カテゴリ分けに利用するには難がある。一方、FKCコーパスでは、すべての投稿に属性データが付与されているので、カテゴリ分けが可能である。さらに、FKCコーパスにはユーザー属性情報が記録されているので、ユーザー情報の統計を作成し、分析することも可能である。

[1] コーパスは不満買取センターの利用許可の元で利用可能。
[2] http://www.fumankaitori.com
[3] 実際には、金銭的価値のあるポイントをユーザーに付与している。
[4] FKCは2015年3月にサービスを開始し、2015年12月頃には100万件近くの意見が投稿されている。
[5] http://search.fumankaitori.com/
[6] http://corp.fumankaitori.com/service/

2.2 楽天データセット

楽天株式会社は研究向けに数種類のデータセットを公開している[7]。最大のデータセットである楽天市場データセットには、約 150 万件の商品データと約 64 万件のユーザーレビューデータが存在しており、ユーザープロフィールデータや商品レビューの点数といった属性データも充実している。

楽天市場データセットには大量の商品データ、レビューデータが存在する一方で、データのドメインは特定の分野に限られる。例えば、データセットに含まれる商品とレビューはいずれも楽天市場で商品を出品している店舗とその店舗が出品している商品に限られる。一方で、FKC コーパスはドメインの制限なくユーザーからの意見を受け付けている。例えば、FKC コーパスには「人間関係」や「公共交通機関」、「政治・行政」というカテゴリが存在しており、世論調査や家庭環境の調査を実施しようとしている分析者や研究者にとっては有益なデータセットである。

3 不満買取センターとコーパスの属性情報

「不満」という単語はネガティブな感情を表すもので、怒り、悲しみ、残念感、イライラ感など複数の感情を含んでいる。不満買取センターは、人々の製品やサービスへの不満を収集する企業である。

FKC は不満買取センターのユーザー（以下、FKC ユーザー）からこうした意見を収集する一方で、収集した意見の分析を行ない、製造業者・サービス提供者に販売している。したがって、FKC はユーザーと製品・サービスとの提供者の間をつないでいると言え、また時には FKC ユーザーの声が元となり、製品・サービスが改善されることも有り得る。

広くユーザーから意見を収集するため、FKC ではモバイル端末向けのアプリケーション[8]と Web アプリケーションによる簡便な投稿環境を提供している。サービスへの登録は簡単になっており、日本語話者なら誰でもユーザー登録ができるようになっている。FKC ユーザーとして登録をした後は、自由に意見を投稿でき、金銭的価値のあるポイントを受け取ることができる。表 1 に投稿内容とその例を示した。投稿作業も簡単にするために、表 1 のフィールドのうち、記入が必須な項目は「不満内容」のみである。しかし、任意記入の項目を記入すると、ポイントが上昇する仕組みになっている。また、ユーザー属性も記入は任意[9]であるが、この属性情報も記入すると、ポイントが上昇するようになっている。したがって、ポイントが最大になるのは、ユーザー属性がすべて記入されており、また投稿属性もすべて記入されている時である。

3.1 投稿とユーザープロフィールの属性データ

3.1.1 ユーザープロフィール属性

下記の 4 つのユーザー属性情報を記入可能である。ユーザーが記入しない場合は、「unknown」が自動的に記録される。

性別　「男性」か「女性」、「unknown」が記録される。

都道府県　ユーザーが居住している都道府県。47 都道府県のいずれか、もしくは「unknown」が記録される。

誕生年　ユーザーの生年。4 桁の整数が記録される。

職業　ユーザーの職業。12 種類の職種が選択可能になっている。

3.1.2 投稿属性

投稿属性では、「改善提案」、「不満の対象」、「不満の対象の提供者」、「投稿カテゴリ」と「投稿サブカテゴリ」が入力可能になっている。このうち、「投稿カテゴリ」と「投稿サブカテゴリ」のみがカテゴリカルデータになっており、残りは自由記入となっている。

ユーザーは記入した不満内容が該当する項目を、投稿カテゴリには 14 項目から、投稿サブカテゴリは 10-13 項目から選択可能である。表 1 の例では、投稿カテゴリが「公共・環境」、投稿サブカテゴリには「駅・電車」が選択されている。

3.2 アノテーション

FKC では、不特定多数のユーザーがサービスに登録可能で、またいかなる意見も投稿することができるため、サービス運営上、望ましくない投稿がされることもある。こうした望ましくない投稿に対処するために、FKC では、アノテータによる投稿のチェックを行なっている。アノテーションでは、以下の 3 種類の作業を行なっている。

1. 投稿へのチェックラベル付与
2. 間違った投稿属性の修正
3. 「サービス・製品提供者」の正規化

2 と 3 の作業については、ユーザーが選択をしていない場合には「unknown」のままとしている。

[7] http://rit.rakuten.co.jp/opendata.html
[8] iOS 向けと Android 向けにアプリケーションを提供している。
[9] 2015 年 12 月よりユーザー属性の登録は必須となった。

表 1: 不満投稿のフィールドと投稿例

フィールド	データの種類	投稿例
不満内容（必須）	free text	電車が毎日、遅延してばかり
改善提案（任意）	free text	余裕をもったダイヤにした方がいい。
不満の対象（任意）	free text	東京線
サービス・製品提供者 任意）	free text	東京鉄道
カテゴリ（任意）	categorical	公共・環境
サブカテゴリ（任意）	categorical	駅・電車

3.2.1 チェックラベル付け

投稿された不満は、すべて「買取可」か「買取不可」を示すチェックフラグがラベル付けされる。「買取可」投稿とは、ポイントが付与される投稿のことであり、「買取不可」投稿とはサービス運営上で望ましくなく、ポイントが付与されない投稿のことである。サービス運営上で望ましく投稿とは、「重複した投稿」、「日本語として意味をなさない投稿」、「不満ではない投稿」、「名誉棄損、個人情報の記述」などがある。

3.2.2 間違った投稿属性の修正

FKC ユーザーが投稿内容とは違ったカテゴリを選択してしまうケースもしばしば発生する。そういった場合に、アノテータは正しいカテゴリを選択しなおす作業を行なう。例えば表 1 の例では、鉄道に関する不満であるので、「公共・環境」が正しい。しかし、「観光地に行く途中で鉄道に感じた不満」という投稿内容ならばユーザーが「宿泊・観光・レジャー」を選択してしまう可能性がある。こうした場合にアノテータはカテゴリを「公共・環境」に修正する。

3.2.3 「サービス・製品提供者」の正規化

「不満の対象」と「サービス・製品提供者」のフィールドはユーザーが自由記述をできる項目であり、そのために表現揺れがしばしば発生する。この表記揺れに対処するために、アノテータが正規化の作業を行なっている。ただし、正規化を行なっているのは、「サービス・製品提供者」のフィールドのみである。「不満の対象」フィールドは「サービス・製品提供者」よりもさらに多種の表記揺れが発生していると考えられるため、現在は正規化を行なっていない。

3.3 アノテーション手順

FKC ではアノテータとして、日本語を母語とするパートタイマーを雇用している。FKC は日々、不満が投稿されるサービスであるため、買取スピードをできる限り早くしなければならず、1 投稿をチェックするのは 1 人のアノテータのみで、複数人でチェックはしていない。しかし、アノテータ 1 人によるチェックのみでは、チェックの質に差が発生してしまう可能性がある。そこで、アノテータのチェックを再確認するタスクを FKC 社員が行なっている。この社員は FKC の方針とアノテーションの細則も熟知しているため、上記の 2 重チェック体制は有効である。

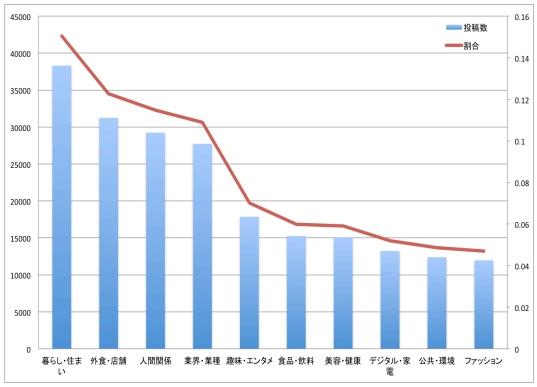

図 2: 投稿属性のトップ 10 項目

4 投稿属性とユーザー属性に注目したコーパス統計

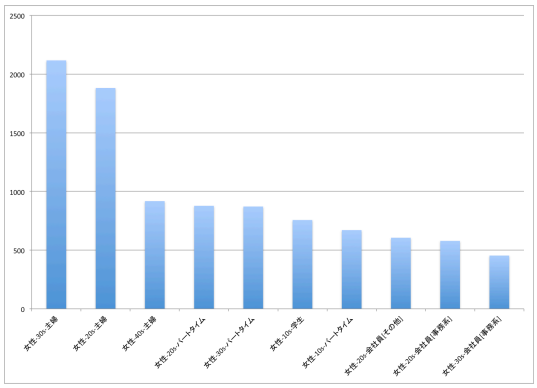

図 3: 性別-年代-職業のユーザー統計のトップ 10 項目

FKC コーパスには 254,683 件の投稿と 25,092 人分のユーザープロフィールが存在している。

3.2.1 で述べた通り、FKC への投稿はアノテータによりチェックされラベルが付与されている。FKC コーパス中の投稿では、買取可の投稿は 241,678 件あり、これは全体の約 95%を占める。一方で、買取不可の投稿は 13,005 件と、全体の 5%程度である。このことから、ほとんどのユーザーは FKC が定めるガイドラインに従って不満を投稿していると言える。

「投稿カテゴリ」と「投稿サブカテゴリ」に関しては、全投稿のうち約 99%が「投稿カテゴリ」を持っており、約 96%が「投稿サブカテゴリ」を持っている。図 2 に投稿カテゴリのトップ 10 項目を示した。トップ 3 項目だけで全体の 39%を占めており、その内容は日常生活に関わる項目である。FKC ユーザーが一般消費者であることを考えると、ユーザーが日常的に体験する事柄に不満を感じ、投稿をしているものと思われる。

ユーザー統計に関しては、約 66%のユーザーがすべてのユーザー属性を記入し、14%がまったく記入をしておらず、残りの約 20%がいずれかのユーザー属性を記入している。すべてのユーザー属性を記入しているユーザーは平均して 11.99 回の投稿をしているのに対し、まったく記入をしていないユーザーは平均 3.51 回の投稿である。このことから、高ポイントを狙うユーザーは、そうでないユーザーに比べてより多くの投稿をしている傾向があり、ユーザーに投稿意欲を働きかける仕組みが動いていると推測できる。

図 3 は、(性別, 年代, 職業) をクロス集計したユーザー属性のトップ 10 を示している。このトップ 10 のユーザー群は全ユーザーのうち 38%を占めており、このユーザー群が投稿している投稿属性を調査すると、大多数が「日用品」、「人間関係」、「食品・外食」であった。図 2 と比較すると、トップ 3 が共通しており、このことから、図 3 に示すトップ 10 ユーザー群が FKC コーパスの代表的な意見発信者であると考えられる。

5 NLP タスクへの応用

FKC コーパスは機械学習を利用した種々の NLP タスクに応用可能である。

FKC コーパスではユーザープロフィール情報が充実しているため、著者推定に応用できる可能性がある。Nguyen らは、ユーザーの年齢を推定するためにブログコーパスで、推定モデルの作成をしており [4]、Mukherjee らはユーザーの性別を推定するため、ブログコーパスを利用したモデル作成をしている [3]。FKC コーパスはユーザープロフィールにユーザーの性別と生年の属性情報を持っているので、この 2 つの手法が応用可能である。また、FKC コーパスでは、ユーザーの居住地、職業、さらに投稿属性として投稿カテゴリと投稿サブカテゴリのデータを持っているので、さらに多くの情報を予測できる著者推定モデルが構築可能である。

別の応用としてドメイン適応が挙げられる。ドメイン適応は、ラベル付きコーパスでモデルを学習し、学習したモデルを使って別のラベルなしコーパスのラベルを予測するタスクである。Dai らは類似したコーパス同士でのドメイン適応 [2] を、Xiao らは非類似コーパス間でのドメイン適応 [10] を提案している。FKC コーパスの投稿属性には、14 項目、投稿サブカテゴリには 10 から 13 の項目が存在している。ドメイン適応の手法を利用すれば、FKC コーパスの投稿属性をラベルデータとして学習した上で、ラベルなしの意見評判データやブログ記事でもカテゴリ分類をすることができるようになる。

6 おわりに

本論文では、FKC コーパスの概要とその NLP タスクへの応用性を紹介した。現時点では、FKC コーパスには 254,683 件の投稿と 25,092 人分のデータのみであるが、FKC は稼働中のサービスであるため日々、投稿件数とユーザー数は増加している。今後は FKC コーパスの件数の拡充に努めると共に、NLP タスクを実際に導入し、その実現性の検証を行なっていく予定である。

参考文献

[1] Daniel Boley, Maria Gini, Robert Gross, Eui-Hong (Sam) Han, Kyle Hastings, George Karypis, Vipin Kumar, Bamshad Mobasher, and Jerome Moore. Partitioning-Based Clustering for Web Document Categorization. *Journal of Decision Support Systems*, 27(3):329–341, 1999.

[2] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering Based Classification for Out-of-domain Documents. In *Proceedings of SIGKDD*, pages 210–219, 2007.

[3] Arjun Mukherjee and Bing Liu. Improving Gender Classification of Blog Authors. In *Proceedings of EMNLP*, pages 207–217, 2010.

[4] Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. Author Age Prediction from Text Using Linear Regression. In *Proceedings of LaTeCH*, pages 115–123, 2011.

[5] Michael G. Noll and Christoph Meinel. Exploring Social Annotations for Web Document Classification. In *Proceedings of SAC*, pages 2315–2320, 2008.

[6] Sylvester Olubolu Orimaye, Saadat M. Alhashmi, and Eu-Gene Siew. Natural Language Opinion Search on Blogs. In *Proceedings of PRICAI*, pages 372–385, 2012.

[7] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC*, pages 17–23, 2010.

[8] Petrovic Sasa, Osborne Miles, and Victor Lavrenko. The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, 2010.

[9] Adam Schenker. *Graph-theoretic Techniques for Web Content Mining*. PhD thesis, University of South Florida, Tampa, FL, USA, 2003. AAI3182715.

[10] Min Xiao, Feipeng Zhao, and Yuhong Guo. Learning Latent Word Representations for Domain Adaptation using Supervised Word Clustering. In *Proceedings of EMNLP*, pages 152–162, 2013.

[11] Zhihua Zhang, Guoshun Wu, and Man Lan. ECNU: Multi-level Sentiment Analysis on Twitter Using Traditional Linguistic Features and Word Embedding Features. In *Proceedings of SemEval*, pages 561–567, 2015.