不満調査データの基礎的分析

鵜飼 佑奈　　久野 雅樹
電気通信大学 情報理工学研究科
u1830015@edu.cc.uec.ac.jp

1 はじめに

現代ではインターネットの更なる普及により，オンライン上で誰でも手軽に意見を述べることが可能となった．同時に，そのような意見をテキストデータとして収集したものを，自然言語処理や機械学習といった分野の研究において扱う機会も増えている．国内のデータでは，例えば通販サイトのレビューデータを用いた研究が多数行われており，ユーザー属性とレビュー内容との関連を調べたり，レビュー本文の解析結果を用いた応用システムの構築などが試みられている．

本研究では，人・モノ・サービスなど様々な対象への「不満」を集めたデータである「不満調査データセット」を用いる。このデータセットはレビューデータのような肯定的，否定的どちらのコメントも含まれるものとは異なり，基本的にはネガティブな内容のコメントのみで構成されたデータとなる。「趣味・エンタメ」や「食品・飲料」などといった，不満の内容に応じてユーザーが選択する「カテゴリ」を限定して分析することにも有用であるが,「全体として，どのような不満があるのか」を横断的に調べることも可能であり，他のコーパスと併せて用いることも有効である．

本稿では，不満内容データを用いた応用研究の前段階として，データの内容の基本的な集計に関して取り扱っていく．

2 関連研究

ここでは，不満調査データセットを用いた関連研究を紹介する．

長谷川らは，属性ごとにグループ化されたユーザーと特定の不満対象との関係を可視化するシステムを提案した [1]．結果として，不満の対象となるもの同士の類似度が高いものは，ユーザー属性の類似度も高くなるということが示された．

澤田らは，通販サイトのレビューデータと不満調査データの 2 つを用いて，商品のもつ肯定的特性の抽出が可能となるかの検証を行った [2]．通販サイトにおいて類似商品どうしの違いを調べる際に，レビューデータのみでは共通する特徴語が多くなり，得られる情報が似通ってしまうという問題がある．この研究では「不満内容データにおいて出現頻度が少ない単語を，商品が持つ肯定的な特徴とする手法」を提案し，先述の問題の解決が可能であることを示した．

3 概要

3.1 使用するデータ

分析には，株式会社 Insight Tech が公開している「不満調査データセット」を用いる．データセットの中から，2015 年 3 月 18 日から 2017 年 3 月 12 日の間に投稿された不満に関する情報をまとめたもの（以降.「不満内容データ」とする）と，不満を投稿したユーザーの属性情報をまとめたもの（以降,「ユーザー属性データ」とする）の 2 つを，本研究で取り扱う．それぞれのデータにおいて，分析の中心となり得る項目を表 1，表 2 に示す．

表 1: 不満内容データにおける項目の一例

項目名	内容
category	カテゴリ
sub_category	サブカテゴリ
company	不満の対象の企業名 (自由記述)
text	不満本文
created_at	投稿日時

不満内容データのカテゴリには,「その他」を含め 20 種類が用意されている．また，2 つのデータに共通して user_id という項目が付加されている．これはユー

表 2: ユーザー属性データにおける項目の一例

項目名	内容
gender	性別
birth	生まれ年 (西暦)
job	職業
prefecture	出身地
marital_status	結婚歴 (未婚もしくは既婚)

ザーを区別するための ID を表す項目であり，2 つのデータはこれをもとに紐づけることができる．

3.2 集計・分析について

初めに，不満内容データ，ユーザー属性データの概要を把握するための基本的な集計を行う．次に，テキストの文字数や単語の出現度といった，不満本文に関する分析を行う．最後に，不満内容データとユーザー属性データの紐づけを行い，不満カテゴリと属性の関係，投稿回数と属性の関係について分析を試みた．

分析に使用している端末のメモリの関係で，今回は生の不満内容データの分割によって作成したファイルのうちの 1 つを用いている．

4 結果と考察

4.1 データの集計

不満内容データには，およそ 525 万件の投稿が集約されている．カテゴリごとの投稿数について，表 3 に結果を示す．

「外食・店舗」や「暮らし・住まい」といった，性別や年齢などに関係なく誰もがカ関わるカテゴリの投稿数が，上位になったと言える．

次に，ユーザー属性情報の中の「生まれ年」「職業」のそれぞれについて，男女の人数の集計を行った結果を示す．本研究においては生まれ年に範囲を設定し，おおよその年齢層で考えることとした．データには性別が無回答のもの（約 3 万件）も含まれているが，ここでは無視している．

まず，生まれ年ごとの男女の人数を表 4 に示す.「30 代」「40 代」について，女性の人数が男性のおよそ 3 倍となっている．

次に，職業別の男女の人数を表 5 に示す．

表 3: カテゴリごとの投稿数

カテゴリ	投稿数
外食・店舗	665385
暮らし・住まい	631554
趣味・エンタメ	586370
業界・業種	425365
食品・飲料	415076
公共・環境	267259
人間関係	260003
デジタル・家電	232585
ファッション	219849
美容・健康	178993
医療・福祉	154190
政治・行政	140039
仕事	105460
教育	97115
アウトドア・スポーツ	90748
宿泊・観光・レジャー	82358
ペット	43852
国際・文化	29709
自動車	20045
その他	600460

表 4: 生まれ年

生まれ年 (およその年齢層)	男性	女性
1996 年以降 (10 代)	5523	7668
1986〜1995 年 (20 代)	29669	28221
1976〜1985 年 (30 代)	8592	27156
1966〜1975 年 (40 代)	5879	16948
1956〜1965 年 (50 代)	2637	6516
1955 年以前 (60〜70 代)	1346	1357

表 5: 職業別

職業	男性	女性
会社員 (技術系)	6167	4067
会社員 (事務系)	3062	9177
会社員 (その他)	6478	8687
公務員	20832	1078
経営者・役員	1108	575
パート・アルバイト	2019	19148
自営業	2741	2789
自由業	1245	1722
専業主婦 (主夫)	401	28227
学生	7858	11119
無職	1656	1953
その他	1554	3561

表 6: カテゴリごとの不満本文の文字数

カテゴリ	文字数 (平均値)
政治・行政	82.84
教育	74.98
業界・業種	73.69
国際・文化	73.56
医療・福祉	70.87
人間関係	69.55
宿泊・観光・レジャー	67.67
仕事	67.32
外食・店舗	66.97
公共・環境	66.72
ペット	63.30
デジタル・家電	61.01
趣味・エンタメ	60.37
アウトドア・スポーツ	58.14
美容・健康	56.46
ファッション	56.14
暮らし・住まい	55.28
食品・飲料	53.88
その他	53.36

「パート・アルバイト」「専業主婦（主夫）」の女性が目立って多く,「会社員（事務系）」においても女性が男性の 3 倍ほどとなっている．その他は男女間で大きな差は見られないが，世間一般で男性の就業者が多いとされている職業については，男性のユーザー数が上回っている．

4.2 不満本文に関する分析

カテゴリごとの不満本文の文字数について，分析の結果を表 6 に示す．

いずれのカテゴリも，10 字程度の短いものから 200 字弱のものが中心となっていた．全体としてはカテゴリによって大きな文字数の差は見られなかったが,「政治･行政」カテゴリについて，わずかに値が大きくなっている．製品やサービスへの不満を端的に書いて終わりとすることもできる他のカテゴリに比べ，ユーザーによる専門的な見解が含まれる可能性の高いカテゴリにおいて，文字数がやや多くなる傾向があったと考える．また，企業からのものと思われる長文のメールの本文をそのまま貼り付けて投稿するユーザーも見られた．今後，形態素解析を行っていく際に支障をきたす可能性が高いので，処理をすべき問題だと考える．

表 7: カテゴリごとの頻出単語

カテゴリ	1 位	2 位	3 位	4 位	5 位
政治	人	ない	税金	ほしい	いい
教育	子供	人	ない	学校	先生
業界	不満	ない	ポイント	人	ほしい
国際	人	日本	ない	国	いい
医療	病院	薬	人	ない	ほしい
人間関係	人	私	自分	ない	子供
宿泊	人	ほしい	高い	ホテル	欲しい
仕事	仕事	人	会社	時間	自分
外食	ない	人	店	ほしい	店員
公共	人	ほしい	車	ない	電車
ペット	犬	猫	ペット	人	ない
デジタル	ほしい	ない	時	欲しい	いい
趣味	人	ない	ほしい	番組	いい
アウトドア	車	人	ない	ほしい	こと
美容	ない	美容	いい	ほしい	人
ファッション	服	サイズ	ない	ほしい	もの
暮らし	ない	ほしい	時	いい	もの
食品	ほしい	ない	味	美味しい	もの
その他	人	ない	ほしい	いい	何

次にカテゴリごとの頻出単語について，上位 5 つを表 7 に示す．なお品詞は名詞と形容詞に限定し，いずれのカテゴリでも上位となった ”の”, ”こと”, ”よう”, ”ん” は除いた結果となる．また，カテゴリ名が長いものは省略して記載している．

”ほしい” が上位となっているカテゴリが多いことから，不満だけでなく要望も述べる傾向があるとわかる．また ”人” も頻出となっており，カテゴリに関わらず，人に関する何かしらの記述がされる傾向があることもわかる．結果から除いた 4 つの単語を含めると，おおよそ上位 10 個の単語に関しては，カテゴリ間で大きな違いはなかった．それ以降の順位の単語で，カテゴリごとの特徴が出やすくなっているように見られた．

4.3 不満とユーザー属性を紐づけた分析

集計の結果，この分析に用いる不満内容データの分割ファイルからは，668 人のユーザー（use_id のみが付与され，その他の属性情報が未登録となっているユーザーも含む）が紐づけられていることがわかった．このうち男性は 107 人，女性は 536 人であった．

まず，男女別のカテゴリごとの投稿数を調べた．表 8 に示す．

この分析におけるユーザーの男女比がおよそ 1:5 と

表 8: 男女別のカテゴリごとの投稿数

カテゴリ	男性	女性
外食・店舗	521	3240
暮らし・住まい	506	2594
趣味・エンタメ	1207	2628
業界・業種	835	2295
食品・飲料	440	1539
公共・環境	298	959
人間関係	148	1369
デジタル・家電	318	794
ファッション	113	968
美容・健康	71	649
医療・福祉	115	635
政治・行政	428	714
仕事	180	342
教育	150	546
アウトドア・スポーツ	435	391
宿泊・観光・レジャー	116	324
ペット	35	191
国際・文化	67	153
自動車	94	67
その他	478	1681

なっていることを踏まえると，男性による投稿数が女性を上回る「アウトドア・スポーツ」と「自動車」は特徴のあるカテゴリと言える．それに対し「ファッション」「美容・健康」は，1:5 という男女比を鑑みても女性の投稿が多くなっており，世間一般で男性・女性それぞれが関心を持ちやすいとされるジャンルにおいて，相関的に不満も増えるということがわかる．「人間関係」についても，女性のほうが気にしやすいという風潮を表した結果と言える．

次に，「職業×性別」で区分した属性ごとに，ユーザー 1 人あたりの不満投稿回数を調べた．表 9 に結果を示す．値は平均値である．値に続く括弧の中は，その区分に属するユーザーの人数を表している．

投稿数にばらつきがあるように見えるが，ユーザー数が少ない区分もあり，この結果だけで深い考察をすることは難しい．少し言い回しを変えただけとなるほぼ同じ内容の不満を何回も投稿するユーザーが存在するので，本文を確認し，何らかの処理を施す必要があると考える．

表 9: ユーザー 1 人あたりの投稿回数の平均

職業	男性 (人数)	女性 (人数)
会社員 (技術系)	59.1 (24)	25.6(19)
会社員 (事務系)	51.4(14)	17.9(53)
会社員 (その他)	48.4(20)	44.3(35)
公務員	NaN(0)	15.3(3)
経営者・役員	NaN(0)	14.0(2)
パート・アルバイト	47.0(8)	41.9(104)
自営業	201.7(9)	28.2(13)
自由業	26.9(10)	15.1(7)
専業主婦 (主夫)	1.7(3)	45.9(251)
学生	6.8(6)	19.7(15)
無職	46.0(6)	83.3(11)
その他	313.6(7)	62.3(23)

5 おわりに

今回の集計・分析は，今後の研究に対する事前準備としての位置づけで行った．集計の結果，登録しているユーザー層に偏りがあること，性別や趣味などに左右されないカテゴリへの不満が多くなりやすいことが確認できた．不満本文に関する分析の結果，ユーザーによる専門的な意見が含まれると考えられるカテゴリに関して，文字数が多くなる傾向があると分かった．またカテゴリごとの頻出単語については，カテゴリ間で目立った違いは見受けられないものの，全体として要望を述べるような形で不満を投稿するユーザーが多く，カテゴリを問わず人に関する記述がされやすいと分かった．不満内容データとユーザー属性データの紐づけによる分析では，ユーザーの性別により投稿する不満カテゴリの傾向が一部異なることが見られた．

今後の展開としては，まずは引き続き紐づけの分析について，不満本文とユーザー属性の関連を調査する．今回の分析とあわせて見られた傾向をもとに，応用システムの提案をしていきたい．

参考文献

[1] 長谷川 徹, 北山 大輔, 不満調査データセットを用いた不満グループの可視化 DEIM Forum, P7-1, 2017.

[2] 澤田 悠治, 北山 大輔, 角谷 和俊 ユーザーの不満情報を用いたアイテムの肯定的特性の抽出, DEIM Forum, P8-5, 2018.