意見投稿プラットフォームにおける意見クラスタリングの試み

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

インターネットの普及により、製品・サービスの利用者は手軽に評判意見を発信できるようになった。ユーザーから発信された評判意見情報を分類するため、SNSから獲得されたデータを対象に、「意見」に注目した分類が研究されている [5][7]。SNS で交わされるメッセージを対象に、意見に注目し分類を行なうことで、分析者は発生している事象を具体的に把握することができる。例えば特定の食品を対象にした投稿では、「価格が高い」または「味付けが薄い」のような意見が想定される。

意見分類のための SNS として Twitter が利用されることが多いが、Twitter では、評判意見情報だけでなく、雑多な内容の投稿が存在している。そのため、分類の前に、評判意見情報だけを抽出する必要がある。また Twitter では、ユーザーが 1 つの意見を複数投稿に分割して投稿する傾向がある。したがって、複数に分割された投稿を特定し、ユーザーの意見を復元する作業が必要になる。

こうした Twitter 特有の現象が存在しない意見収集プラットフォームとして、我々は不満買取センター [1] を運営している。不満投稿 (以下、不満買取センターに寄せられる投稿を「不満投稿」と呼称する）には文字数制限が存在しないため、ユーザーは任意の文面を投稿することができる。したがって、Tweet で発生するような、意見が複数 Tweet にまたがるという現象は発生せず、ユーザーの意見は必ず 1 不満投稿内に収まる。

本研究では、不満投稿を対象に、意見分類を目的としたクラスタリングを行なった。具体的には、Kevin ら [5] が行なった Tweet クラスタリングを拡張し、クラスタリングを複数の条件で行ない、最適な組み合わせの探索を行なった。

2 関連研究

日本語 Tweet に対して、意見に注目してクラスタリングを行なった研究には鷹栖ら [7] の研究がある。鷹栖らは、Twitter の特性に着目し、意見 Tweet だけを抽出した後、当該ユーザーの関連 Tweet も意見の一部とみなし、意見 Tweet 群を抽出している。しかし、不満投稿においては、ユーザーの意見は 1 投稿内で収まるため、このような前処理を必要としない。また鷹栖らは、クラスタリングに階層型クラスタリングの Ward 法を採用しているが、我々は非階層型クラスタリングである k-means と生成モデルの LDA(Latent Dirichlet Allocation) を採用する。

英語 Tweet に対しては Kevin ら [5] がクラスタリングを試みている。Kevin らは英語 Tweet 集合に k-means と LDA の両方を用い、６クラスタと３０クラスタのクラスタリングを行ない、その性能が十分ではなかったと報告している。

我々の評価実験は Kevin らの実験を拡張する。Kevin らの実験では BOW(Bag-Of-Word) 素性のみを利用してクラスタリングを行なっているが、本研究では、BOW 以外の素性も導入し、評価を行なった。また、k-means にはユークリッド距離以外の距離関数を導入し、LDA ではハイパーパラメタのチューニングを行ない、性能の向上を計った。

3 不満買取センターとは

不満買取センターは 2015 年 3 月にサービスを開始した、意見収集プラットフォームである。一般消費者から、ネガティブな「不満」の意見のみを集め、収集した意見を製品・サービス提供者に販売する事業を行なっている。意見を一般消費者から収集する際には、インセンティブとして金銭的価値のあるポイントを付与している。また同時に、収集した意見を活用するために、分析ダッシュボードと分析レポートの提供を行なっている。

一般消費者が少ない手間で意見を投稿できるように、不満買取センターでは Web、iOS、Android アプリケーションを提供している。投稿の際には、数種のカテゴリ情報と自由記述方式の不満内容を記述するように求められるが、必須記述項目は不満内容だけであり、残りの項目は記述する必要はない。ただし、すべての項目を記述すると、付与されるポイントが増加する仕組みになっており、これにより投稿者は積極的に情報を記述している。表 1 に不満投稿内容の一例を示した。この例の場合では、すべての項目が記述されているので、最大ポイントが付与される。

4 評価実験

4.1 素性

本研究では、3 種類の素性を用意し、その組み合わせを検証する。なお、素性値には各不満投稿で素性が出現した回数を利用する。

4.1.1 BOW 素性

投稿文中に出現した形態素を素性として利用する。ただし、利用する形態素は品詞が名詞、形容詞、動詞のいずれかの品詞に限定する。

[1] http://fumankaitori.com/

表 1: 不満投稿内容の一例

記述項目	記述内容	記述例
不満内容 (必須)	free text	電車が毎日、遅延してばかり。さすがに毎日の遅延はあり得ない。
不満の対象 (オプション)	free text	東京線
サービス・製品提供者 (オプション)	free text	東京鉄道
カテゴリ (オプション)	categorical	駅・電車
サブカテゴリ (オプション)	categorical	公共・環境

表 2: 評価データセットの統計情報

データセット名	件数	投稿ユーザー数	平均文字数	標準偏差	クラスタ数
bottle-coffee	1,598	741	140.959	90.253	5
bodycare-5, 13	1,291	869	52.312	34.605	5, 13

表 3: 組み合わせ一覧

素性	重みづけ	次元削減	アルゴリズム
BOW 述語 述語+述語項 BOW+述語+述語項	なし TF-IDF	なし 100 200 300 400 500	k-means(ユークリッド距離) k-means(cosine 距離) LDA(チューニングなし) LDA(チューニングあり)

4.1.2 述語素性

投稿文中では述語に意見が出現する傾向があると考え、述語を素性として利用する。例えば「新製品の値段が、旧製品よりも高い。」という文であれば、述語である「高い」が意見として現れる。ただし、投稿文中で複数の述語が出現している場合は、すべてを素性として扱う。

4.1.3 述語・項素性

投稿文中で意見の主題と属性が記述されていれば、意見情報として有効な素性になり得る。述語に対する項が活用できると考え、(項, 項タイプ, 述語) のペアを素性とした。例えば「新製品の値段が、旧製品よりも高い。」という文であれば、「高い」という述語に対して、ガ格の「値段」とヨリ格の「旧製品」が項となり、(値段, ガ格, 高い) と (旧製品, ヨリ格, 高い) が素性として生成される。

4.2 重み付けと次元削減

不満投稿文はニュース記事や Web ページほど長くはないため、構築される頻度行列では、ほとんどの素性値が低頻度になる可能性がある。そこで、我々は TF-IDF による重み付けと LSI(潜在意味インデキシング) による次元削減をクラスタリング前に行なう。削減後の次元数がクラスタリング性能に与える影響を調査するため、100 次元から 500 次元までの次元数を 100 きざみで次元削減を行なう。

4.3 クラスタリングアルゴリズム

4.3.1 k-means

Anna[3] は k-means を利用した文書クラスタリングにおいて、距離関数がクラスタリング性能に与える影響を調査し、ユークリッド距離と比較して、*cosine* 距離や *Jaccard* 距離を利用すると、*Purity* と *Entropy* が向上すると報告している。

本研究では、実装の簡便さと実行速度 [2] を考慮し、ユークリッド距離と *cosine* 距離を k-means の距離関数として利用し、その差を観察する。

4.3.2 LDA

LDA(Latent Dirichlet Allocation) はトピックごとにトピック所属確率を計算するので、もっとも所属確率が高いトピックを所属クラスタとみなせば、k-means と同じ出力を得ることができる。所属確率を計算する際には、ハイパーパラメータである α (トピックの生成パラメータ) と β (単語の事前分布) が結果に影響を与える。したがって、LDA を利用したクラスタリングにおいては、適切な α と β を選択する必要がある。

適切なパラメータを選択するため、Bogdan ら [2] は遺伝アルゴリズムを利用し、*Silhouette* 指標を最大化するパラメータチューニングを提案している。本研究では Bogdan らに倣い、*Silhouette* 指標を最大化するような α と β を TPE(*Tree-structured Parzen Estimator Approach*) アルゴリズム [1] で探索する。なお、探索にあたっては α と β ともに 0.1 から 1 までの範囲で探索を行なう。

4.4 評価指標

クラスタリング結果を評価するために、人手でラベル付けを行ったゴールドデータを用意し、評価を行なう。評価には NMI(*Normalized mutual information*) と Pairwise F1[5]、NMI と Pairwise F1 の平均である Average の 3 種類のスコアを用意した。Average は次式で求められる。

$$Average = (NMI + Pairwise\ F1)/2$$

[2] テストデータセットにおいて、*Jaccard* 距離の場合はユークリッド距離の 10 倍の実行時間を要する一方、*cosine* 距離ではほぼ同じ実行時間であった。

なお、NMI は $[-1, 1]$ の値を取り、Pairwise F1 は $[0, 1]$ の値を取るため、Average は $[-0.5, 1]$ の数値を取る。

4.5 実験設定

4.5.1 データセット

本研究では、3 種類のデータセットを用意した。1 つ目は「ボトルタイプの缶コーヒー」の不満を言及したデータセットの “bottle-coffee”、2 つ目と 3 つ目はユーザーの不満投稿のうち、ボディケア用品を言及した不満投稿を集めて作成された “bodycare-13” と “bodycare-5” である。

“bottle-coffee” のデータセットは「不満キャンペーン」と呼ばれる特別枠の収集システム [3] で集められた不満であり、投稿時にユーザーは「容器・パッケージに関する不満」や「味に関する不満」といった意見を 5 カテゴリから選択できるようになっている。“bodycare-13” と “bodycare-5” は、人手で意見内容を分類したデータセットで、前者は意見を 13 分類、後者は 5 分類されている。

両方のデータセットともに意見内容で分類がされている。したがって、クタスタリング結果の average スコアが高いほど、正しく意見クラスタリングができていると言える。

表 2 にデータセットの統計情報を示した。Graham らは日本語 Tweet の平均文字数を調査し、40-50 字程度であると報告している [4]。Tweet と比較すると、“bottle-coffee” は Tweet の 3 倍近くの文字数があり、“bodycare-5, 13” は概ね Tweet と同じくらいの長さであると言える。

4.5.2 組み合わせ

素性と重みづけ、アルゴリズムのもっともよい組み合わせを調べるため、全組み合わせで評価実験を行ない Average スコアを算出した。表 3 に組み合わせ内容の一覧を示す。ただし、表のうち、アルゴリズムに LDA を利用する場合は、次元削減は行わない。素性の組み合わせに 4 通り、重みづけの有無に 2 通り、次元削減の有無と次元数の違いに 6 通り、アルゴリズムに 4 通りで計 112 通りの実験を行なった。

5 実験結果

表 4 に、データセットごとにトップ 3 の組み合わせと、ベースラインとして Kevin ら [5] の実験設定での結果を示す。いずれのデータセットにおいても、チューニングをした LDA がトップの性能であり、また大きくベースラインの結果を上回っている。トップ 1 以降の組み合わせについては、“bottle-coffee” においては k-means を利用した組み合わせが 2 位と 3 位であった。一方で、“bodycare-5”, “bodycare-13” においては 1 位から 10 位あたりまでは LDA を使った組み合わせで、k-means を利用した組み合わせより優れた働きを見せた。

“bottle-coffee” に関しては、クラスタ数を 5 から 13 に変化させると、Average が半分ほどの数値になっている。Kevin ら [5] は、30 クラスタでは 6 クラスタの 3 分の 1 ほどの PairwiseF1 であると報告している。本研究でも分類クラスタ数を増加させると共にクラスタリング性能が同程度に落ちてしまうと言える。

[3] 不満キャンペーンであっても、投稿システムには変わりはなく、投稿者は自由記述で不満内容を記述できるようになっている。

5.1 LDA のチューニングによる性能変化

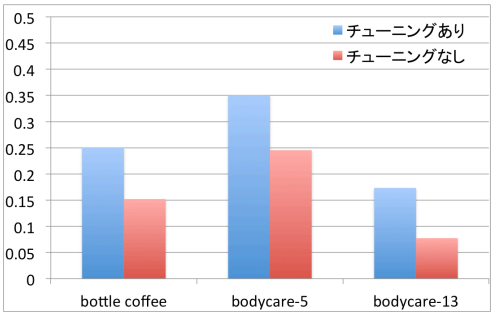

図 1: LDA チューニングの有無による Average スコアの変化 (BOW 素性を TF-IDF 素性で重み付け)

図 1 に LDA のチューニング有無による性能変化を示す。どの場合においても、チューニングをした方がよい Average スコアであり、*Silhouette* 指標による LDA のチューニングは有効であることが言える。

5.2 素性組み合わせによる性能変化

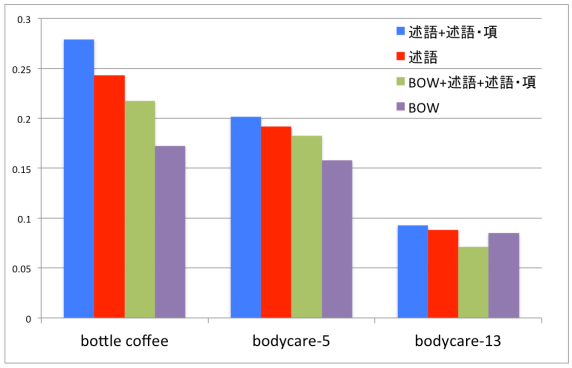

図 2: k-means における素性ごとの Average スコアの変化 (TF-IDF で重み付けし 100 次元に次元圧縮)

素性によるクラスタリング性能の変化は k-means と LDA で異なる挙動を見せた。図 2 に k-means での素性ごとの性能変化を示す。k-means においては、いずれのデータセットでも「述語+述語・項」の素性組み合わせがもっとも良い性能を示し、次いで「述語」素性が良い性能を示している。“bottle coffee” と “bodycare-5” においては「BOW」素性がもっとも悪く、“bodycare-13” では「BOW+述語+述語・項」の組み合わせがもっとも悪い。「BOW」素性がノイズとして働いたため、前者では「BOW」素性の追加と共にクラスタリング性能が悪化し、後者では、組み合わせ素性にした場合にクラスタリング性能が悪化したと考えられる。

図 3 に LDA での素性ごとの性能変化を示す。LDA においては、BOW 素性が有効に働き、一方で「述語+述語・項」の素性組み合わせがクラスタリング性能を悪化させる傾向が見られた。ただし、“bodycare-5” で

表 4: データセットごとの評価結果

データセット	Rank	素性	重みづけ	次元削減	アルゴリズム	Average	NMI	Pairwise F1
bottle-coffee	1	BOW	なし	なし	LDA+チューニング	**0.319**	0.204	0.434
	2	述語 + 述語・項	TF-IDF	100	kmeans+ユークリッド	0.278	0.008	0.470
	3	述語	TF-IDF	100	kmeans+ユークリッド	0.260	0.007	0.441
	base-LDA	BOW	なし	なし	LDA	0.227	0.116	0.338
	base-kmeans	BOW	なし	なし	kmeans+ユークリッド	0.140	0.01	0.286
bodycare-5	1	BOW + 述語 + 述語・項	TF-IDF	なし	LDA+チューニング	**0.370**	0.0945	0.645
	2	BOW	TF-IDF	なし	LDA+チューニング	0.35	0.131	0.569
	3	BOW + 述語 + 述語・項	なし	なし	LDA+チューニング	0.288	0.103	0.474
	base-LDA	BOW	なし	なし	LDA	0.284	0.089	0.478
	base-kmeans	BOW	なし	なし	kmeans+ユークリッド	0.18	-0.008	0.369
bodycare-13	1	BOW	TF-IDF	なし	LDA+チューニング	**0.177**	0.049	0.304
	2	述語 + 述語・項	なし	なし	LDA+チューニング	0.156	0.079	0.234
	3	BOW + 述語 + 述語・項	TF-IDF	なし	LDA+チューニング	0.154	0.002	0.306
	base-LDA	BOW	なし	なし	LDA	0.115	0.044	0.185
	base-kmeans	BOW	なし	なし	kmeans+ユークリッド	0.086	0.007	0.165

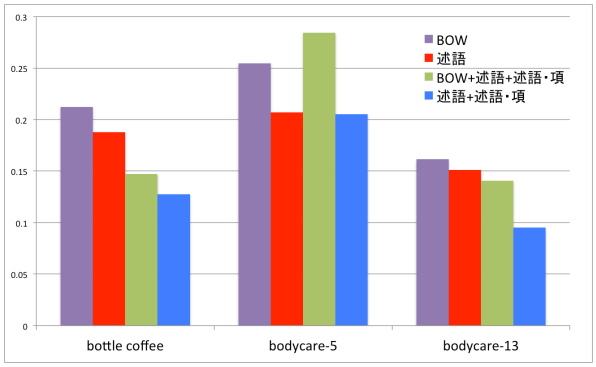

図 3: LDA における素性ごとの Average スコアの変化 (TF-IDF で重み付けし、チューニング)

は「BOW+述語+述語・項」の素性組み合わせがもっとも良い結果を示している。これは LDA による次元削減がうまく働き、「述語」と「述語・項」素性組み合わせのうち、不要な素性が削減された結果、「BOW」素性よりも良い性能であったと考えられる。

6 おわりに

本研究では不満買取センターで収集された不満投稿を意見ごとに分類するため、複数の条件でクラスタリングを行ない、結果の評価を行なった。本研究で利用したいずれのデータセットにおいても、*Silhouette* 指標を最大化するように α と β をチューニングした LDA がもっとも良い分類性能を記録することがわかった。

しかし、まだ十分な分類性能であるとは言えず、さらなる改善が今後の課題である。今後の課題としては、疎な素性への対応と別の指標の利用した LDA チューニングが挙げられる。

k-means では、「述語+述語・項」がもっとも良い結果であったが、同時に述語と項の組み合わせのためにこの素性は疎になりがちである。そこで、WordNet を利用し、項の集合を拡張すると、疎な状態が解消され、クラスタリング性能が改善する可能性がある。

LDA のチューニングには *Silhouette* 指標を利用したが、クラスタリングの結果を評価する Internal 指標には他の指標も利用可能である。Stein ら [6] は *Expected Density* が F スコアと線形な関係になることを示しており、本研究でも有効に働く可能性がある。

参考文献

[1] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems 24*, pages 2546–2554. 2011.

[2] Bogdan Dit, Annibale Panichella, Evan Moritz, Rocco Oliveto, Massimiliano Di Penta, Denys Poshyvanyk, and Andrea De Lucia. Configuring topic models for software engineering tasks in tracelab. In *Proceedings of the 7th International Workshop on Traceability in Emerging Forms of Software Engineering*, pages 105–109, 2013.

[3] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, pages 49–56, 2008.

[4] Graham Neubig and Kevin Duh. How Much is Said in a Tweet? A Multilingual, Information-Theoretic Perspective. In *AAAI Spring Symposium on Analyzing Microtext*, pages 32–39, 2013.

[5] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. Topical clustering of tweets. In *Proceedings of the ACM SIGIR 3rd Workshop on Social Web Search and Mining*, 2011.

[6] Benno Stein, Sven Meyer zu Eissen, and Frank Wißbrock. On cluster validity and the information need of users. In *Proceedings of the International Conference on Artificial Intelligence and Applications (AIA 03)*, pages 216–221, 2003.

[7] 内海彰 鷹栖弘明. 文脈を考慮した観点に基づく意見ツイートクラスタリング. 言語処理学会 第 *21* 回年次大会 発表論文集, pages 433–436, 2015.