SNS からの重要意見抽出のためのデータセット構築及び LLM による分類検証

矢口 一晟[1]　櫻井 恵理子[2]　櫻井 義尚[3]
[1]明治大学大学院 先端数理科学研究科　[2]産業能率大学 経営学部　[3]明治大学 総合数理学部
cs233017@meiji.ac.jp　SAKURAI_Eriko@hj.sanno.ac.jp　sakuraiy@meiji.ac.jp

概要

本研究では，重要意見抽出のためのデータセット構築および大規模言語モデル（以下：LLM）を用いた重要意見抽出の手法を提案し，その有効性を検証した．そこで，人間がラベリングを実施せずに SNS から企業がマーケティング活動に必要となる重要意見抽出モデルが構築可能であることを示す．これにより，重要意見データセットの作成や抽出にかかる時間的かつ金銭的コストの削減が期待できる．データセットの構築検証ではワークショップを開催することで重要意見に関連する議論を実施し学生による重要意見データの作成が可能であるかを検証した．また重要意見の抽出検証では，LLM による Zero-shot 分類および擬似データを作成した後，ファインチューニングによる重要意見分類の精度を比較した．

1 はじめに

Twitter（現 X）などの SNS にはユーザーから発信される投稿データが蓄積されており，その中には企業のサービス向上や改善などに有益となる情報が存在し，マーケティング調査に活用されている．一方で，投稿データの中でも有益な情報が含まれている割合は少なく，人手でそれらを抽出することは大変な人的コストを要する．そこで本研究では， LLM を用いることで重要な意見を含む投稿の抽出を効率化する手法を提案し，その有効性を検証する．

具体的には，まず重要意見の抽出検証の先駆けとして X と Google レビューデータを用いた重要意見ワークショップを実施することで検証用データとなる重要意見データの作成を行う．ワークショップでは計 14 名の学生とともに 4 つの観点に対して，個人およびグループワークを行い，重要意見データセットの構築を図った．次に，ワークショップで作成した重要意見データから，得られる特徴を LLM によって抽出する．最後に抽出した特徴を用いて，2 つの手法による重要意見の抽出効果を検証する．

2 関連研究

SNS にはユーザーから発信される膨大な投稿データが蓄積されており，その中には企業のマーケティング調査に有益な情報が存在する．そのため，SNS を用いた意見抽出の研究が進められている［1］．

また，これまでの自然言語処理分野において，意見の中でも「要望」「不満」「現状認識」などの意見自体の役割を担う部分の抽出は扱われてこなかった．そこで，山本ら［2］の研究では人々が何を求めているかを調査するため機械学習モデルである Support Vector Machine を用いて，アンケート回答に記載されている意見の中から要望が述べられている「要望文」を抽出する研究を行った．山本らの研究では「要望」に着目し，横浜環状線北西線の建設に関する業務における要望文を抽出するシステムを提案しているが，本研究では人々の共感性に着目し，意見データの中から企業がマーケティング調査を行う上で必要となる重要意見の抽出を目指す．

さらに，近年 LLM に関する研究が盛んに行われており Brown ら［3］は GPT-3 の開発を行い，LLM に特定のタスク説明文と少量の教師データを入力することで高い精度を達成し，日本語タスクにおいても高い性能を発揮できる場合があることが報告されている．そこで，藤井ら［4］は LLM によって生成した擬似データを用いて学習した小規模モデルが 6 個の日本語タスクにおいて LLM との比較検証を実施した結果，フォーマルなテキストを入力としたときの分類タスクにおいて Zero-shot および Few-shot Prompting を上回る結果となった．このアプローチは，LLM による重要意見抽出といった特定のタスクにおいても有効性を持つことが考えられる．

本研究では，大規模言語モデルである GPT-4o を用いた分類および品質の良い擬似データセットの作成による分類手法を提案し，重要意見の抽出タスクに

おける有効性を検証する．

3 データセット

本研究では，重要意見データセットの構築におけるワークショップおよび重要意見抽出効果の検証において分類用の検証データでは，Twitter（現 X）と Google レビューから合計 196 件のデータを用いた．また，本研究は重要意見抽出の対象をディズニー施設としているため，ディズニーに関するサービス向上・改善に繋がる重要意見抽出を行う．データの内訳は表 1 の通りである．

表 1：ワークショップ用データセットの内訳

	Twitter（現 X）	Google レビュー
データ数	160	36

4 提案手法

本章では，ワークショップによる重要意見データセットの構築および重要意見抽出の 2 つの手法を説明する．

4.1 重要意見データセットの構築

重要意見のラベリングは個人の主観や意見によって偏るという特性があるため，グループでの議論を実施し，役に立つ又は結果的に議論が巻き起こった意見をラベリングすることが重要意見のラベルになるという仮説のもと，ワークショップを実施した．具体的には，産業能率大学の学生 14 名と有識者をお招きし，4 つの観点（図 1）での 5 値（3 値）ラベリングおよびコメントワークを行い，196 件のデータセットをもとに重要意見データセットを作成した．最終的に，学生によるデータセット構築が，どの程度有識者が作成した重要意見データを網羅できているかを確認するために，図 2 の流れで 5 つの分析を実施した．

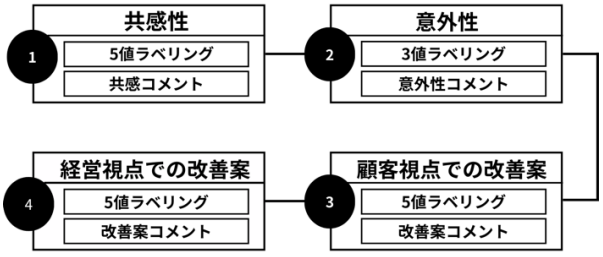

図 1：ワークショップの流れおよび 4 つの観点

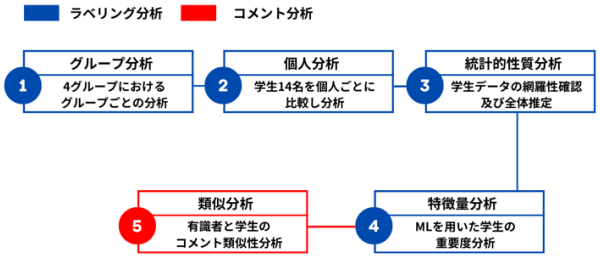

図 2：5 つの分析フロー

4.2 LLM による重要意見抽出

LLM による重要意見の抽出検証では，GPT-4o を用いた Zero-shot 分類（図 3）と GPT-4o により作成した擬似データを，京都大学が公開している Deberta-v3-base-japanese に Fine-Tuning させることで重要意見抽出モデルによる分類（図 4）の 2 つの手法で検証した．後者の擬似データ作成においては同条件下で 1000 件，2000 件，3000 件の擬似データを作成し Fine- Tuning させ，比較する．検証時はワークショップで構築した重要意見データを検証用データとし，評価指標である Accuracy, Recall, Precision, F1-score を用いてモデル評価を行う．

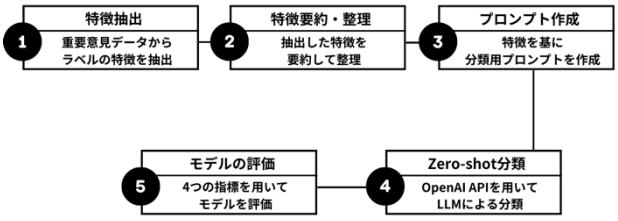

図 3：GPT-4o による Zero-Shot 分類手法

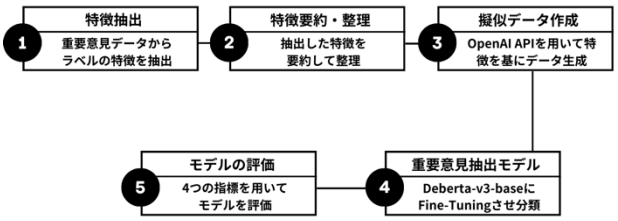

図 4：擬似データ Fine-Tuning による分類手法

また，Deberta-v3-base-japanese は BERT[5]モデルを基盤としつつ Deverta-v3 [6]を日本語に特化させたモデルとなっている．

5 実験 1:重要意見データセット構築

5.1 実験概要

本実験では，4 つの観点からラベリングに関する個人ワークおよびグループワークを実施し，有識者意見と比較することで，学生によるデータセット構築が可能であるかを検証した．分析の先駆けとなるワークショップでは，図 1 の 4 つの観点から個人ワークとグループワークを行い，それぞれ 5 つのラベリングと，その観点から表 2 のラベル 1 と 2 に該当するデータに対してコメントを行うものである.また，分析に関しては図 2 のラベリング分析およびコメント分析に関する 5 つの分析を行うことで，学生による重要意見データの構築可否を検証した.

表 2：4 つの観点における 5 値ラベルの説明

	共感性	**意外性**
ラベル 1	とても共感	とても意外性あり
ラベル 2	やや共感	やや意外性
ラベル 3	わからない	意外性はなし
ラベル 4	やや共感しない	
ラベル 4	全く共感しない	
	顧客視点_改善案	**経営視点_改善案**
ラベル 1	とても改善できる	とても改善できる
ラベル 2	やや改善できる	やや改善できる
ラベル 3	わからない	わからない
ラベル 4	やや改善できない	やや改善できない
ラベル 5	全く改善できない	全く改善できない

また，4 つの観点を 2 値ラベルに変換する際には，表 2 の「ラベル 1・2」に該当するデータを「ラベル 1」，「ラベル 3〜5」に該当データを「ラベル 0」とする.

さらに，実験を行う上で学生 14 名をそれぞれ Student1〜14 に置き換え，有識者のラベリングデータを正解データと仮定し，検証を実施した.

5.2 実験結果

5.2.1 ワークショップの結果

ワークショップでは 4 つの観点に対し各 1 時間の時間を設けラベリングを実施した結果，「ラベル 1・2」と判定された，つまり票数が多く議論が活発になったのは「共感性＞意外性」「経営視点での改善案＞顧客視点での改善」となった．また，有識者のラベリングも同様の結果となり，学生および有識者ラベリングでは「**共感性**」のラベリング数が最多となった.

5.2.2 グループおよび個人分析

本実験では 14 名の学生を 4 つのグループに分け，個人ワークを実施した後，グループワークを実施した．結果として，ワークショップの結果に付随しグループでの議論も共感性に対する議論が最多であったが，グループ意見として出力された結果は個人ワークで得られた意見の要約であり新たな意見が創出された事例はごく僅かとなった．そのため，共感性に着目し学生 14 名と有識者を含めた各個人同士の分析を実施した結果，学生の共感票数（2 値ラベルにおけるラベル 1 と判定されるもの）が多いものは有識者の共感を得ていることが分かった.

5.2.3 網羅性の確認および特徴重要度分析

学生の共感判定が有識者の正解データをどの程度網羅できているか，また Student1〜14 の中で，どの学生が網羅性に寄与しているのかを確認した．そこで，14 名分の学生票数の閾値を半数である 7 件に設定し比較検証した結果と，Random Forest を用いての予測および Feature importance での Student の重要度分析の結果を以下に示す.

表 3：網羅性と Random Forest での予測

	網羅性	Random Forest
Accuracy	**70.9%**	**84.8%**
Precision	0.696	0.895
Recall	0.571	0.820
F1-score	0.627	0.831

表 4：Feature importance の値（上位 4 名抜粋）

	値
Student 12	**0.880**
Student 13	0.400
Student 4	0.390
Student 9	0.270

5.2.4 類似分析

本実験では，学生と有識者が共感すると判定（2 値ラベルにおけるラベル 1）したコメントに着目し，両方のコメントを，Sentence BERT を用いて類似度分析を実施した．以下の表 5 は有識者が共感判定且つコメントをした 82 件/196 件に対して，類似度の

閾値を0.5以上と0.6以上に分けて分析した結果である．

表 5：Sentence BERT を用いたコメント類似度

閾値	該当件数/82 件
0.5 以上	43 件
0.6 以上	27 件

5.3 考察

本実験では．ワークショップの結果から共感性のラベリングおよびコメントが多く存在し，重要な指標であることが示された．そこで共感性に着目し，5つの分析を実施した．表 3 より網羅性の確認および Random forest を用いた正解データ予測では Accuracy にて 70〜85%の精度となり高い割合で学生のラベリングが正解データを判定できているといえる．また，学生の中でも表 4 より Student12 は正解データに重要となることが示された．ワークショップの事前アンケート結果から Student12 はディズニー施設内でのキャスト経験があるため，ドメイン知識の豊富さによる特定のキーワードへの理解度の深さが影響していると考えられる．最後に，コメントの類似度分析においては，閾値が 0.5 以上のとき 1/2 以上，0.6 以上では 1/3 の精度となっており，学生のラベリングが有識者の正解データ（重要意見データセット）の構築に寄与できることが示された．

6 実験 2：LLM による重要意見抽出

6.1 実験概要

本実験では GPT-4o による Zero-shot 分類と GPT-4o により作成した擬似データを Fine-Tuning させることで重要意見抽出モデルによる分類の 2 つの手法を検証した．2 つの手法ではそれぞれ OpenAI の API を用いており，GPT-4o による分類では言語モデルに対して重要意見の特徴を含むプロンプトを与えることで，重要意見であるか否かを判定させた．この手法を Base とする．また，Fine-Tuning によるモデル構築での分類では，言語モデルを使用することで 1000 件，2000 件，3000 件の擬似データを作成し，重要意見抽出タスクに特化させた形で重要意見判定を実施した．

3 種類の件数で Fine-tuning させたモデルをそれぞれ FT_1000，FT_2000，FT_3000 として表記する．

6.2 実験結果

表 6：2 手法による重要意見抽出の検証結果

	Acc.	Pre.	Rec.	F1.
Base	**69.90%**	**0.736**	0.464	0.569
FT_1000	60.09%	0.603	0.255	0.331
FT_2000	59.69%	0.541	0.455	0.490
FT_3000	59.45%	0.556	0.358	0.431

6.3 考察

表 6 から Fine-tuning させたモデルは，学習件数を増やしても Base と比べて，4 つの指標全てにおいて劣ることが分かった．そのため，擬似データの Fine-Tuning による重要意見抽出の有用性は検証できなかった．しかし，Base による分類では約 7 割の正解率となり，Precision に関してはモデルがデータを重要意見であると予測したもののうち実際に重要意見である（正例）割合が高いことを示しており，予測の信頼性が高いといえる．また，Fine-Tuning モデルに着目すると FT_2000 モデルで F1-score の値が高いことが分かる．しかし，データ数を増やした FT_3000 モデルでは全体的に値が下がっていることから，単にデータ数を増加させるだけでは精度が向上しないことが示された．これは，生成されたデータの中で類似データが多く含まれていることによる過学習への対応や適切なデータセットのサイズを選定する必要があると考える．

7 おわりに

本研究では，ワークショップの実施による重要意見データセットの構築およびワークショップから得られた特徴を用いて重要意見を抽出するシステムを提案した．実験 1 では，学生による重要意見ラベリングが正解データに対して Accuracy が 70%〜85%の精度を達成することができ，コメントを含めた類似度では半数以上のデータを構築可能であることを検証できた．また，実験 2 では，重要意見指標を用いた 2 つの手法において LLM による分類にて正解率が 70%の精度を達成できたが，擬似データ作成による Fine-Tuning での分類では精度向上が見られなかった．これは，擬似データの質の低さが要因であると考える．そのため，今後は，LLM に入力するプロンプトの工夫やデータ増幅に対する過学習への対応などにより，モデルの分類精度向上を目指す．

謝辞

本研究は，ワークショップに参加してくださった産業能率大学経営学部櫻井ゼミの皆様のご協力により実現しました．心より感謝申し上げます．

本研究は JSPS 科研費 20K11960 の助成を受けたものです．

参考文献

[1] 川島崇秀，哲．佐藤 ，典．神門，“Twitter からの消費者ニーズの抽出手法に関する提案,”第 8 回データ工学と情報マネジメントに関するフォーラム，2016.

[2] 山本瑞樹，孝．乾，大．高村，聡．丸本，裕．大塚 ，学．奥村，“文章構造を考慮した自由解答意見からの要望抽出,” 言語処理学会 12 回年次大会併設ワークショップ「感情　評価　態度と言語」，2006.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child and A. Ramesh, "Language Models are Few-Shot Learners," 2020.

[4] 藤井巧朗 ，智．勝又，“日本語タスクにおける LLM を用いた疑似学習データ生成の検討,”言語処理学会 第 30 回年次大会 発表論文集，2024.

[5] J. Devlin, Ming-Wei Chang, Kenton Lee , Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2019.

[6] P. He, Jianfeng Gao , Weizhu Chen, “DEBERTAV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing,” 2023.