# Twitterからの消費者ニーズの抽出手法に関する研究

川島 崇秀

近年，著しく普及してきたソーシャルメディア上には，製品やサービスに関する口コミが日々大量に投稿されている．これらの口コミ情報は，企業のマーケティング活動等に有益であり，膨大な書き込みの中から企業にとって価値のある口コミを抽出する技術が求められている．ソーシャルメディアの中でもTwitterは近年顕著な普及を見せており，リアルタイム性の高さ・ユーザの多様性・投稿量の多さといった視点から分析対象として大きな注目を集めている．

Twitterを対象とした口コミ分析に関する代表的な研究としてセンチメント分析が知られている．センチメント分析とは，Twitter上のレビューを肯定と否定の2つのカテゴリに分類することである．しかし，この分類では，投稿を感情と いう視点てで肯定的か否定的かの2値に区分する為，消費者の顕在的なニーズを表す要望を含む投稿を，直接的に抽出することはできない．また，Twitterを対象とした従来の辞書ベースの要望抽出手法では，Twitter上に投稿される多様な文章表現への対応が難しく，十分な分類性能を達成することができなかった．

本研究では，Twitterのような短文かつ省略の多い文章を対象に，ビジネス応用が期待される要望抽出を高精度に実現する手法を提案する．提案法の特徴は，これまで情報抽出の分野で有効性が確認されている教師あり機械学習のアルゴリズムSVM(Support Vector Machine)を要望抽出に適用する点にある．教師あり学習で必要となる教師データの収集は，人手による収集ルールの生成と機械学習を融合した半教師あり学習手法Distant Supervisionを適用することで低コストなな学習データの収集を試みる．

評価実験では，ソーシャルゲームに関する投稿を対象とし，提案手法と従来法の分類精度の比較を行うことによって，本手法の有効性を検証した．教師データの収集時に使用したゲームタイトル名を使用する場合と，使用しない場合の2通りの方法でソーシャルゲームに関する投稿を各1000件ずつ収集し，人手でラベル付けを行ったものを評価用データとして用意した．評価データに対して，構築した分類器を用いて分類を行った結果，いずれの方法で収集した評価用データに対しても，適合率，再現率，F値において提案手法が高い評価を示し，本手法の有効性が確認された．

（指導教員　佐藤 哲司）

# Twitterからの
# 消費者ニーズの抽出手法に関する研究

2016年3月

201413127

川島崇秀

筑波大学情報学群

知識情報・図書館学類

# 目 次

# 第1章　はじめに

## 1.1　背景

近年，ソーシャルメディアの普及により，誰でも簡単に情報発信ができるようになった．ソーシャルメディア上では，製品やサービスに関する口コミが日々大量に投稿されている．こういった背景から，企業においてソーシャルメディア上の口コミ情報を市場調査や反響測定といったマーケティング活動に活かそうという試みが注目されている [?]．特にソーシャルメディア上の口コミ情報を分析し，トレンド予測，ブランド・商品に対する評判の理解や改善に活かす試みは，ソーシャルリスニングと呼ばれており，従来のマーケティング調査では難しかった，消費者が日常の中で発する自然な声を聞くことができるといったメリットがある．

ソーシャルメディアの中でも近年顕著な普及を見せる Twitter は，リアルタイム性の高さ・ユーザの多様性・投稿量の多さから分析対象として大きな注目を集めている [?]．しかし，Twitter 上では，日々何千万ものクチコミ情報が秒単位で行われており，これらの投稿を人手で分析することには膨大なコストが掛かる．したがって，Twitter 上の投稿をビジネス活動に活用していくためには，自動で口コミ情報の抽出を行い，要約を行うなどの分析を自動化するツールの開発が必要不可欠である．

Twitter 上の投稿を自動で抽出する研究としては，センチメント分析が知られている [?]．センチメント分析とは，ユーザが書くレビューに対して，肯定的なものと否定的なものの２つのカテゴリに分類する手法である．しかし，この手法では，投稿を感情という視点で肯定的か否定的かの２値に分類する為，感情を含まないがビジネス活動に於いて価値のある情報を抽出することが困難である．例えば，要望を含む投稿などは消費者のニーズを直接的に表す重要な情報であると考えられるが，従来の手法で分類することは難しい．

そこで本研究では,「要望」という点に着目して，消費者の要望を含むツイートの抽出を試みる．従来の手法で抽出の対象としていなかった要望を含むツイートの抽出を行い，それらのデータをビジネス活動に活かしていくことで，サービスの品質改善や新規事業の創造といった活動の支援につなげていくことが期待できる．

しかしながら，Twitter 上の投稿から商品やサービスに関する要望を含むツイートの抽出をする際に課題となるのは，Twitter 上の投稿における多様な文章表現である．Twitter 上には文法的に崩れた表現や多様な文章表現が非常に多く存在している．それ故，従来手法で提案されている辞書ベースの手法 [?] を適用した場合，これらの多様な表現への対応が困難であった．

## 1.2 目的

本研究の目的は，企業の消費者ニーズの理解を支援するために，従来手法と比較して，より高い精度での要望抽出を実現することにある．最終的な展望としては，消費者ニーズの理解を支援することによって，サービスの品質改善や新規事業の創造といった，ビジネス活動の促進に繋げていくことが期待される．そこで本研究では，要望を含むツイートの抽出に機械学習のアルゴリズムを適用することで，従来手法と比較してより高い精度での抽出を試みる．また，学習データの収集に半教師あり学習の一つである「Distant Supervision」の考えを適用することで，より低コストな要望表現の抽出方法を提案する．

## 1.3 本論文の構成

以下，本論文では 2 章で関連研究を紹介し研究の位置づけを明確にする．3 章で本研究で対象とする要望の定義を行い，4 章で提案手法を詳述する．5 章で評価した結果を述べ，考察する．最後に 6 章でまとめと今後の課題を示す．

# 第2章　関連研究

本研究は，レビュー記事に書かれたクチコミ情報から要望を抽出する研究である．特に短文を投稿する Twitter を対象とすることに技術的な課題がある．また，機械学習の手法を用いるという点において，学習コストの低減が大きな課題となる．

## 2.1　口コミ分析に関する研究

口コミ分析に関する研究としては，weblog を対象としたものが主流であった．鈴木ら [?] は，半教師あり機械学習の手法を用いて，Weblog から評価情報の抽出を行った．奥村ら [?] は，weblog を定期的に監視し，評価情報を自動抽出するシステムを開発した．

Twitter 上の口コミ情報の抽出に関する研究も盛んに行われている．Twitter 上の投稿に対してセンチメント分析を行った研究としては野畑ら [?, nobata] 研究が挙げられる．野畑らは教師あり機械学習の手法を用いて Twitter 上の投稿をポジティブなものとネガティブなものの２つのカテゴリに分類した．

## 2.2　要望の抽出に関する研究

要望文の抽出に関する研究ではアンケートの自由記述欄から意見や要望を抽出する試みが行われている．山本ら [?] は，アンケー トの自由回答欄から要望を抽出する手法として，自由回答の記述の何文目に要望が書かれる傾向にあるのかを分析し，要望文 を自動抽出する手法を提案した．大塚ら [?] は自由回答アンケートにおいて間接的な要求を抽出するための基準として「〜てほしい」に言い換え可能か否かという基準を提案し，機械学習の手法を用いて要求を含むテキストを抽出している．

また，Twitter 上から要望を含む投稿を抽出する研究としては，栗原ら [?] の研究が挙げられる．栗原らは Twitter 上から地方自治体に関する要望を含む投稿の抽出を試みている．手法としては，あらかじめ作成した要望表現の特徴を含む辞書を用いたパターンマッチングを用いている．山本 [?] らは，Twitter を用いて生活に関連する単語からなる辞書を作成し，特定の地域の要望を含む，生活情報を抽出する手法を提案した．

要求表現の言語学的な定義に関する試みも行われている．大森 [?] は要求表現の定義として,「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」，「当為非断定」,「希望非断定」の態度を帯びる文は要求文であるとした．大森はさらに，要求表現の文法的な特徴として,「〜しろ」「〜たい」「〜ほしい」といった文末表現を挙げている．

## 2.3 Distant Supervision を用いた情報抽出に関する研究

Web からの情報抽出のタスクに Distant Supervision を用いた研究も行われている．Distant Supervision[?] とは，半教師あり学習の手法の一つであり，知識ベースから取得した少数の手がかり表現を用いることで，半自動的な教師データの収集を可能にする学習方法である．M. Mintz[?] らは Web テキストからの関係性抽出のタスクに Distant Supervision の考え方を適用し，少量の知識ベースから大量の学習データを収集している．この際，知識ベースとして FreeBase を用いている．三浦ら [?] は，Twitter の投稿に対するセンチメント分析のタスクに Distant Supervision の考え方を適用することで，教師あり機械学習の低コスト化に成功している．この際，学習データ収集に用いる手がかり表現として顔文字を用いている．山本ら [?] は，Web 上のニュース記事からの企業間関係抽出のタスクに Distant Supervision を適用している．ここでは，あらかじめ作成した教師データから，その判断の決め手となった語を抽出し，手がかり表現としている．

これらの研究では Web 上からの情報抽出に関するタスクにおいて，分類性能を低下させることなく，教師あり機械学習の低コスト化に成功している．これらの結果から，適切な手がかり表現の定義が可能であれば，情報抽出のタスクにおいて「Distant Supervision」の考えを適用することで，低コストで教師データを取得できると考えられる．

## 2.4 本研究の位置づけ

本研究では，機会学習のアルゴリズムを用いることによって，Twitter 上から要望文を対象とした自動抽出を行う．また，半教師あり学習の手法の一つである「Distant Supervision」を用いることによって，少数の手がかり表現から半自動的に学習データを収集する．そのため，要望という視点から Twitter 上のクチコミ情報を抽出するという点で，感情という視点から投稿の抽出を試みた野畑らや三浦らの研究と異なる．また要望の抽出手法においても，機械学習の手法を用いたという点で，辞書ベースの手法を用いた栗原らの手法とは異なっている．

本研究は，Twitter 上の要望文の抽出というタスクに対して，機械学習の手法を適用し，学習データの収集に半教師あり学習の手法の一つである「Distant Supervison」を用いることで，低コストかつ高い精度な要望抽出を試みた点に新規性があると考える．

# 第3章　要望の定義

関連研究で述べた大森と栗原らの論文を参考に，本研究で扱う要望を定義する．大森は要求とは，文に描かれている事態であるとし，次の４つの条件を満たす表現であるとした．

**条件 1** 当該事態は，それを捉える当事者にとって望ましい事態である
**条件 2** 当該事態は，当該当事者にとって未実現である
**条件 3** 当該事態の実現主体として，個人，集団，組織など，意志を持つ主体が存在する
**条件 4** 当該当事者は，当該事態の実現を，当該事態の実現主体に求めている

大森はさらに，直接的な要求を表す文章は,「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」，「当為」,「当為非断定」,「希望非断定」のいずれかの態度を帯びるとし，それぞれの文法的な特徴を明らかにしている．
また，栗原らは大塚らの研究を参考に,「〜てほしい」「〜てください」「〜てくれ」といった，日本語母語話者のほとんどが「要求」と判断できる表現を「直接要求」表現とし,「〜べき」「〜がベストだと思う」「 が必要」といった,「〜てほしい」に言い換え可能な表現を「要求意図」表現とした．栗原らはさらに，Twitter 上の投稿は自由回答アンケートと異なりユーザの独り言や愚痴が投稿される傾向があることに注目し，直接要求や要求意図に当てはまらない場合でも，その内容が要望の動機になる否定的なテキストを「不満」と定義し,「直接要求」「要求意図」「不満」の３つに該当するテキストを要望と定義した．

本研究では以上の先行研究を踏まえ「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」,「当為」,「当為非断定」,「希望非断定」の態度を帯びる表現と，これらに該当しないが要望の動機となる否定的な表現を「不満」とし，まとめて要望と定義する．

#### 命令

相手が意志的に制御できる動作を，相手に強制する表現

例 1) つまらん心配はしないてで早く行け
例 2) はやくバグ修正しろ

#### 依頼

相手の意志を尊重して，相手にある動作をするよう頼む表現

例 3) あなたはやく帰ってきてちょうだい
例 4) ちょっと，その婆さんに会ってみ てくれないか?

**禁止**

相手にある動作をしないこと，あるいは，ある事態が生じないように努力することを命令する表現

例 5 ) そういうことに，やたら興味を持つな
例 6 ) いちいちアップデートすんな

**誘いかけ**

聞き手に，話し手と同様の行動をとるように要求する表現

例 7) やりましょう，松田さん熊谷さん
例 8) 一緒にゲームしましょう！

**希望**

話し手自身に関わる事態の実現を 希望する，あるいは他者がある事態 を実現することを希望する表現

例 9) 千葉へいってもらいたい
例 10) 早く返金して欲しい

**希望非断定**

希望の態度を断定することを控える表現

例 11) 音楽というコンテンツを手に入れたら，通勤の時に電車で iPod やその他携帯音楽プレイヤーで聴きたいかもしれない
例 12) Windows も Mac も辞書データをひっくるめて月額制でお安くしておきますよという，プミアムコース を作ってもらいたいかもしれない

**当為**

ある事態が望ましいとか，必要だ，というように事態の当否を述べる当為の態度のうち，「〜べきだ」,「〜なければならない」のような述語の 基本形をとって表される表現

例 13) 日本は早急に貿易黒字を減らすべきだ
例 14) 君は，あの時彼と別れるべきだった

**当為非断定**

当為の態度を断定することを控える表現

例 15) 日本は早急に貿易黒字を減らすべきだろう
例 16) 君は積極的になったほうがいいかもしれない.

**不満**

「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」, 「当為」,「当為非断定」,「希望非断定」に該当しないが, 要望の動機となる否定的な表現

例 17) 横浜市営地下鉄の始発遅い, 最悪
例 18) 市役所の対応悪いわ

# 第4章　要望の抽出手法の提案

## 4.1　提案手法の概要

Twitter から要望を含む投稿を抽出するに当たって課題となるのは，Twitter 上の投稿における多様な文章表現である．Twitter は気軽に投稿を行えるといった特徴から，非常に多くのユーザに親しまれている．しかし，投稿に 140 字という字数制限があるため，短文の投稿が多く，またその気軽さ故に，文法的に崩れた表現や多様な文章表現が非常に多く存在している．その為，従来手法で提案されている辞書ベースの手法を適用した場合，これらの多様な表現への対応が困難である．

より高い精度で要望を含む投稿を抽出する方法としては，教師あり機械学習を用いた手法などが挙げられるが，分類器を学習させる為に，多大な人的コストがかかるという点に課題があった．教師あり機械学習の手法を用いる場合，分類器を学習させるための教師データを人手で用意する必要がある．それ故，リアルタイム性が非常に高い Twitter に対応する為には，人手により頻繁にラベル付けを行い，教師データを更新し続ける必要があった．

そこで本研究では，要望を含む投稿の抽出に機械学習のアルゴリズムを適用し，教師データの収集に半教師あり学習の手法を用いることで，低コストかつ高い精度での要望抽出を実現する手法を提案する．

#### 1． SVM(Support Vector Machine) を用いた要望の抽出

本手法の一つ目の特徴は要望を含む投稿の抽出に機械学習のアルゴリズムである SVM-Support Vector Machine) を用いる点である．SVM は 1995 年に，AT & T の V．Vapnik によって統計的学習理論の枠組みで，提案された 2 クラスのパターン認識手法の一つだ．SVM では，2 種類のクラスのデータと，分離超平面との間の距離 (マージンと呼ぶ) が最大になるような分離超平面が，最も汎化能力 の高い超平面になるということを利用している．クラスの特徴ベクトルを非線形変換して，その空間での線形の識別を行う「カーネルトリック」と呼ばれている方法を,「マージン最大化」という基準で行う．それ故，現在知られている多くの手法の中でも，高い分類性能と汎化能力を有している学習モデルとして知られている．本研究では，SVM を用いて要望を含む投稿の抽出を行うことで，より高い精度での要望抽出を試みる．

#### 2． Distant Supervision を用いた教師データの収集

2 つ目の特徴は，教師データの収集に Distant Supervision の考えを適用する点である．Distant Supervision とは，半教師あり学習の手法の一つであり，知識ベースから取得した少数の手がかり表現を用いることで，半自動的な教師データの収集を可能にする学習方法

である．先行研究において，Distant Superviosion の考えを Web テキストからの関係性抽出のタスクに用いることの有効性が確認されている．本研究では知識ベースとして，大森らの論文中に記述されている文末表現リストを用いることで，半自動的に教師データを収集している．本手法を用いることで，低コストな機械学習を実現する．

### 4.1.1 要望抽出手法の枠組み

提案法を実現する要望抽出システムの概要を図 4.1 に示す．この図では，①および②のステップにおいて，要望抽出の対象となる商品名/サービス名に関するユーザの投稿を後述する要望表現辞書と n-gram 判定の 2 段階の処理によって教師データとして収集し，分類器の学習を行う．③のステップにおいて構築した分類器を用いて要望を含む投稿の抽出を行っている．以降 4.2 節では，教師データの収集に用いる要望表現辞書の作成方法について述べる．4.3 節では，4.2 節で作成した要望表現辞書を用いて，教師データを収集する手順について説明する．

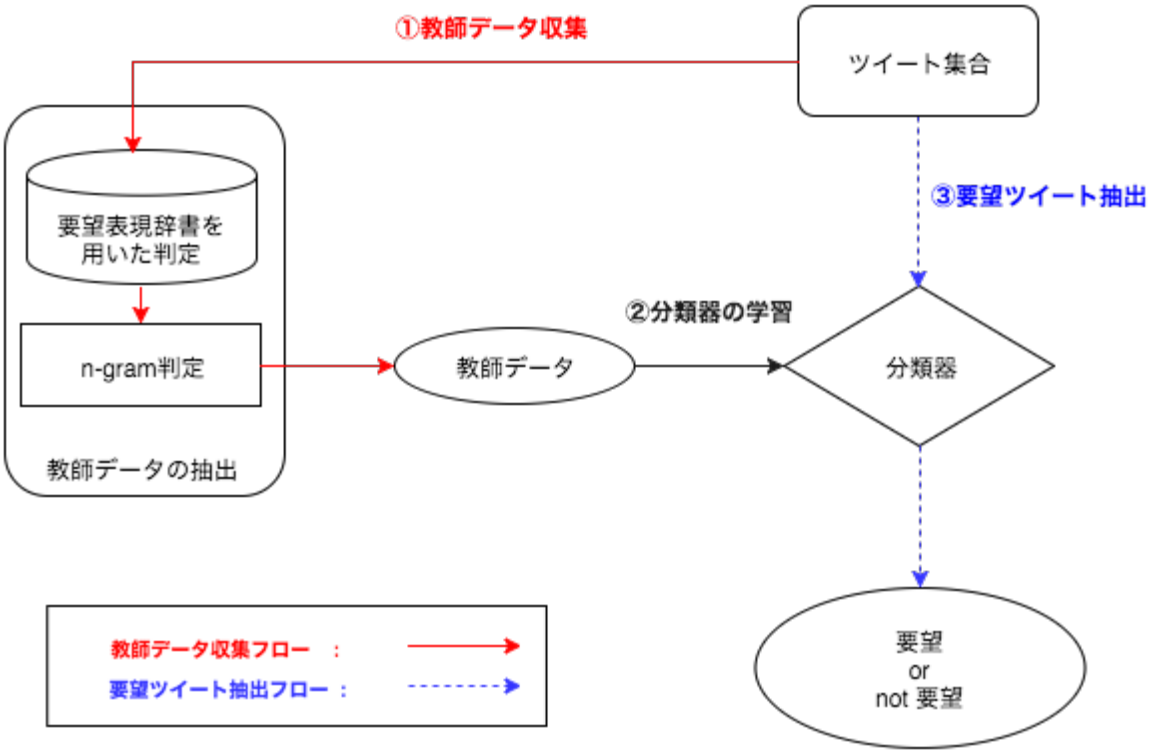


図 4.1: 提案手法の概要

## 4.2 要望表現辞書の作成

Distant Supervision の手法を用いた教師データの収集では，予め，教師データの判別の手がかりとなる表現を決定しておき，それらの表現的な特徴を含むデータを収集することによって，半自動的な教師データの収集を可能にする．Distant Supervision の手法を用いて教師データを収集するには，まず，教師データの特徴を決定する必要がある．以降，本論文では，教師データの特徴となる表現のことを手がかり表現と呼称する．

先行研究において要望を含む文には「〜しろ」「〜たい」「〜ほしい」といった特徴的な文末表現が出現することが知られている．本研究では，大森らの論文中に記述されている要望文の特徴表現リストを参考に，合計 19 個の特徴的な表現を定義し，手がかり表現とする．これらの手がかり表現をデータベースに格納することで，要望表現辞書とした．

本研究で使用する手がかり表現一覧

(命令) 〜しろ，〜して
(依頼) 〜ください，〜下さい，動詞否定形+でくれ，〜てくれ，〜て頂戴 (ちょうだい)
(禁止) 〜するな，〜やるな
(誘いかけ) 〜しよう
(希望/希望非断定) 動詞連用形+たい ，動詞テ形+欲 (ほ) しい，〜たい，ほしい
(当為/当為非断定) 〜べき
(不満) 〜過ぎ，〜にくい，〜づらい，〜ない

## 4.3 教師データの収集

4.2 節で作成した要望表現辞書を用いて，教師データを収集する．要望表現辞書を用いて教師データを収集するに当たって重要となるのは，手がかり表現の出現位置である．要望表現辞書に定義した手がかり表現は文末に出現する傾向が極めて高く，文末以外で出現する場合では，要望とはならない可能性が高い．そこで本研究では，より高い精度で教師データを収集する為に，手がかり表現の出現位置を考慮した，以下の 2 段階の処理で教師データの収集を行う．

### 4.3.1 手がかり表現を含むツイートの抽出

4.2 節で作成した要望表現辞書を用いて，TwitterAPI から対象の商品名/サービス名を含み，かつ手がかり表現を含むツイートを抽出する．この際，以下の除去対象となる語を含むツイートはスパム/ボットによる投稿である可能性が高いとして，収集の対象外とした．

除去対象となる語の例

定期，拡散，速報，価格，最安，料金，Line，Amazon，DM，ゲット，iTunes，彼氏

### 4.3.2　手がかり表現が文末から 3-gram 以内に出現するツイートの抽出

先行研究において，要望を含む投稿には，文末に特徴的な表現が現れることが確認されている [?]．Twitter 上の手がかり表現を含む要望を収集した結果,「〜し/て/欲しい/なぁ」や「〜しろ/よ/アホ」など，手がかり表現が，文末から 3-gram 以内に出現する傾向が確認された．

この結果を元に，4.3.1 節で抽出したツイートを文単位に分割し，文末から 3-gram 以内に手がかり出現するものを選択し，教師データとした．ツイートの文単位への分割には，句点・空白・顔文字といった表現を区切り文字として使用し，これらの表現が出現した位置でツイートを分割し，1 文と見なしている．また，3-gram 以内に出現するかの判定には，形態素解析器である Mecab を使用した．MeCab を用いてツイートを形態素に分解し，手がかり表現が，文末の形態素から 3 形態素以内の距離に存在しているかを確認することで判断している．

## 4.4　分類器の生成

4.3 節で収集した教師データを用いて分類器の構築を行う．分類器のアルゴリズムには情報抽出のタスクにおいて有効性が知られている SVM を用い，実装には Python の機械学習ライブラリである scikit-learn を使用した．

素性には，単語の出現頻度などによって文書をベクトルで表現する形式である Bag of Words(BoW) を用いる．ただし，品詞が，名詞・動詞・形容詞・形容動詞・副詞・助動詞のいずれかに該当しない単語は除外した．また，素性の構築時には　文章内において一定以上の割合で出現する単語を頻出語として除去している．

# 第5章　評価実験

## 5.1　評価対象

ソーシャルゲームは，近年急速な普及を見せており，多くの要望が Twitter 上に投稿されている．そこで本研究ではソーシャルゲームに関する投稿を評価対象とした．また，分類器を構築する際に使用する教師データとして，表 5.1 に示すソーシャルゲーム 10 タイトルに関する投稿を使用する．

評価用データは，2015 年 8 月から 2ヶ月間に投稿されたツイートを以下の 2 通りの方法で収集した．この際，4.3.1 節で除去対象としたキーワードを含むツイートの除去を事前に行っている．

**手法 (i) 教師データで使用したゲームタイトルを使用する**
教師データとして使用したソーシャルゲーム 10 タイトルのタイトル名を含むツイートを各 200 件ずつ，合計 1000 件のツイートを収集した．この際，ゲームの公式 Twitter アカウントに対するリプライも収集対象としている．結果，要望を含む投稿は 76 件得られた．

**手法 (ii) 教師データで使用したゲームタイトルを使用しない**
教師データとして使用していないソーシャルゲームタイトル「白猫プロジェクト（白プロ，＃白猫）」を含むツイートを 1000 件収集した．この際，ゲームの公式 Twitter アカウント（@wcat__project）に対するリプライも収集対象としている．結果，要望を含む投稿は 79 件得られた．

以上の手法により収集したツイートに対して，クラウドソーシングサービスのランサーズを用いて，要望のラベルを人手で付与した．各ツイート毎に 5 名の参加者に回答してもらい，もっとも一致率の高い解答を正解ラベルとして付与した．この際，回答の質を向上させる為に，100 ツイート毎に解答難度の低いダミーデータを用意し，ダミーデータへの回答を誤った参加者の解答を事前に除去している． 回答者の判別の一致度を示す k 係数は，手法 (i) で 0.468，手法 (ii) で 0.548 となり，５人の解答はおおむね一致していることが分かる．本研究では，以上の手順により作成したラベル付きツイートを評価用データとして使用する．

表 5.1: 教師データの収集に使用するゲームタイトル一覧

| ゲームタイトル | 検索時に追加で使用したキーワード |
|---|---|
| パズル&ドラゴンズ | パズドラ，@pad __ sexy |
| モンスターストライク | モンスト，@monst __ mixi |
| 刀剣乱舞 | とうらぶ，@TOUKEN __ STAFF |
| 艦隊これくしょん | 艦これ，@KanColle __ STAFF |
| アイドルマスターシンデレラガールズ | モバマス，# imas __ cg，アイマス CG |
| ラブライブ! スクールアイドルフェスティバル | スクフェス，@lovelive __ SIF |
| Fate/Grand Order | fatego，FATEGO，@fgoproject，FateGo |
| ブレイブフロンティア | ブレフロ，@brave __ pr |
| 魔法使いと黒猫のウィズ | 黒ウィズ，黒猫のウィズ，@colopl __ quiz |
| グランブルーファンタジー | グラブル，@gbf __ vee，@granbluefantasy |

## 5.2 評価方法

提案手法と従来法の分類精度の比較を行うことによって，本手法の有効性を検証する．提案手法の有効性を検証するには，抽出したツイート集合がどれだけ正解しているかという正確性と，抽出した記事集合が全ての正解のうち，どれだ け正解を含んでいるかという網羅性の 2 つの観点からの評価が必要となる．本論文では，正確性を適合率 (precision)，網羅性を再現率 (recall)，適合率と再現率の調和平均である F 値 (F-measure) によって提案手法の抽出精度を評価する．それぞれの計算方法について，以下に示す．

$$precision = \frac{\text{抽出した正解ツイート数}}{\text{抽出したツイート数}} \tag{5.1}$$

$$recall = \frac{\text{抽出した正解ツイート数}}{\text{全ての正解ツイート数}} \tag{5.2}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{5.3}$$

## 5.3 実験設定

### 5.3.1 SVM の設定

SVM のカーネルには線形カーネルを使用した．また，パラーメタの設定では，C パラーメタの値を 1．0 から 10．0 まで 0．1 刻みで試して調整を行い，1．0 という値に設定している．

### 5.3.2 従来法の実装

比較対象である辞書ベース分類器の実装に当たっては，栗原らの論文を参考に，大塚らの論文中に記述されている要望表現リストを辞書の作成に使用した． 手法 (i)，手法 (ii) の手順により収集した評価用データに対して，作成した辞書内の要望表現とマッチするかどうかの判定を行い，一致する場合に要望を含む投稿であると判定する．

## 5.4 実験結果

評価データとして，教師データで使用したゲームタイトルを用いた場合の実験結果を，表 5.2 に示す．また，評価データとして，教師データで使用したゲームタイトルを用いなかった場合の実験結果を，表 5.3 に示す．

表 5.2: 手法 (i) 教師データの収集に使用したゲームタイトルを用いる場合

| | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| 提案法 | 0.22 | 0.46 | 0.30 |
| 従来法 | 0.2 | 0.06 | 0.12 |

表 5.3: 手法 (ii) 教師データの収集に使用したゲームタイトルを用いない場合

| | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| 提案法 | 0.24 | 0.57 | 0.34 |
| 従来法 | 0.19 | 0.09 | 0.12 |

## 5.5 考察

5.4 より，いずれの手法の評価データを用いた場合でも，ベースライン手法と比較して，適合率，再現率，F 値が向上していることが分かる．これにより，要望を含む投稿の抽出に SVM と「Distant Supervision」を用いた本手法の有効性が証明された．

特に再現率は，評価指標の中でも大幅な精度向上を確認することが出来た．この要因としては，Distant Supervision の手法を用いて収集した大量の学習データが，Twitter 上の多様な文章表現への対応を可能にした為であると考えられる．また，手法 (i)，手法 (ii) を比較すると，教師データの収集時に使用していない未知のゲームタイトルに関しても，教師データの収集時と同様のゲームタイトルを使用した手法 (i) と同等以上の分類性能を発揮していることが確認できる．逆に，手法 (i) よりも手法 (ii) の場合に高い分類性能を発揮している．この要因としては，学習データの収集時に使用したゲームジャンルの影響が考えられる．今回の実験は，Twitter 上の全てのソーシャルゲームに関する要望に，共通

する単語出現分布があるという前提の上で行っている．しかしながら，ソーシャルゲームには，ゲームジャンルが存在しており，各ジャンルごとにゲームシステム上の特徴が異なるケースがある．従って，ゲームシステム上の差異から，ジャンルごとに要望の種類も異なってくる可能性は十分に考えられる．今回の実験では，学習データのゲームジャンルの違いを考慮していない為，使用したゲームのジャンルには偏りがあるが，ゲームジャンルの違いを考慮し，バランスよく学習データを収集することで，より汎用的な分類器の構築が可能になると考えられる．

また，適合率に関しては，従来手法と比較して，大幅な精度向上を達成することは出来なかった．この要因としては，学習データの収集時に，一定数のノイズが混入してしまった為であると考えられる．今回の実験では，先行研究を元に，手がかり表現の出現位置を文末から３形態素以内に設定したが，各手がかり表現ごとに，正解データとなる要望を高い精度で獲得可能な値は，異なっている可能性がある．各手がかり表現ごとに最適な出現位置を設定し，ストップワード, 評価極性といった複数のルールを組み合わせて学習データを収集することで，さらなる適合率の向上が期待できる．

提案手法によって得られた分類精度に関しても，先行手法と比較して優位な結果を得ることが出来たが，実用レベルには達していない．今後，分類精度を向上させていく方法としては，新たな手がかり表現の追加が挙げられる．学習データの収集時における詳細なルール設定に加えて、新たな手がかり表現を追加することで、より高い精度での要望抽出が期待できる．

# 第6章　まとめ

本論文では，Twitter 上からより高い精度で要望に関する投稿を抽出することを目的に，Twitter から消費者の要望を含む投稿の抽出手法を提案した．本手法では，要望表現の抽出に教師あり機械学習のアルゴリズムである SVM を用いることで，従来手法と比較して，より高い精度での要望抽出に取り組んだ．また，教師データの収集に半教師あり学習の一つである「Distant Supervision」を適用することで，低コストな機械学習の実現を試みた．

「Distant Supervision」の手法を用いた教師データ学習データの収集に当たって，先行研究を元に手がかり表現を定義し，データベースに保存することで要望表現辞書とした．また，手がかり表現が文末に出現する傾向があることから，本研究では，手がかり表現が文末から 3 グラム以内に出現するものを教師データとして収集している．また，SVM の素性としては，Bag of Words を使用している．この際，品詞が，名詞・動詞・形容詞・形容動 詞・副詞・助動詞のいずれかに該当しない単語は除外した．

評価実験では，ソーシャルゲームに関する投稿を対象とし，提案手法と従来法の分類精度の比較を行うことによって，本手法の有効性を検証した．教師データの収集時に使用したゲームタイトル名を使用する場合と，使用しない場合の 2 通りの方法でソーシャルゲームに関する投稿を各 1000 件ずつ収集し，人手でラベル付けを行ったものを評価用データとして用意した．評価データに対して，構築した分類器を用いて分類を行った結果，いずれの方法で収集した評価用データに対しても，適合率，再現率，F 値において提案手法が高い評価を示し，有効性を確認できた．

今後の課題としては，まず，学習データの収集の為の詳細なルール設計が挙げられる．今回の評価実験で十分な適合率を達成できなかった要因の一つとして，学習データの収集時に，一定数のノイズが混入したことが考えられる．今回の実験では，先行研究を元に，手がかり表現の出現位置を文末から 3 形態素以内に設定したが，各手がかり表現ごとに，正解データとなる要望を高い精度で獲得可能な値は，異なっている可能性がある．各手がかり表現ごとに最適な出現位置を設定し，ストップワード, 評価極性といった複数のルールを組み合わせて学習データを収集することで，さらなる適合率の向上が期待できる．また，新たな素性の追加や，手がかり表現の追加といった，分類性能を向上させる施策も行っていきたい．

# 謝辞

佐藤哲司教授（筑波大学大学院図書館情報メディア研究科）は，研究の細部にわたり懇切丁寧な指導をしてくださいました．心より感謝いたします．また，日常の議論を通じて多くの知識や示唆を頂いた佐藤研究室の皆様に感謝します．