情報処理学会論文誌　数理モデル化と応用　Vol.13　No.2　69–83 (Aug. 2020)

Twitterからの意見抽出モデル構築のための教師データ作成手法

野崎 雄太[1]　櫻井 義尚[1,a)]

受付日 2020年1月13日，再受付日 2020年4月2日，
採録日 2020年4月23日

概要：本論文では，教師データセットの作成において，事例をランダムに選び，アノテーションすると不均衡データになってしまう課題に対して，機械的なプレフィルタリングを用いたサンプリングにより，不均衡化を緩和するアノテーション手法 PSSA（Prefilter based Stepwise Sampling for Annotation）を提案する．また，辞書フィルタを用いた PSSA による Twitter からの意見抽出モデルを構築し，提案手法の有効性を示した．まず，辞書フィルタを用いた PSSA による，不均衡化の緩和効果の検証のため，ツイートのアノテーション実験を行い，次にアノテーション段階での不均衡データ対策の有効性を検証するため，意見抽出モデルを構築し，アノテーション手法と前処理，機械学習構築手法の組合せの違いによるモデル精度の違いを検証した．最後に，アノテーションを行うサンプル選択に辞書フィルタを用いることによる影響を分析するため，各辞書フィルタを適用した場合とフィルタリングしなかった場合のモデル精度を比較した．以上の比較実験を通して，提案するアノテーション手法の優位性を多角的に検証した．

キーワード：アノテーション，不均衡データ，教師データ，機械学習，自然言語処理

Dataset Creation Method for Constructing Opinion Mining Model from Tweets

Yuta Nozaki[1]　Yoshitaka Sakurai[1,a)]

Received: January 13, 2020, Revised: April 2, 2020,
Accepted: April 23, 2020

Abstract: In this paper, we propose an annotation method that relieves imbalance by using mechanical prefiltering to solve the problem that randomly selected cases and annotated result in imbalanced data when creating a training dataset. We proposed PSSA (Prefilter based Stepwise Sampling for Annotation) and verified its effectiveness. Specifically, we first created a training dataset from Tweet with the task of opinion mining from Twitter, and verified the effect of mitigating data imbalance by PSSA. Next, we construct a machine learning model using the constructed dataset and evaluate its accuracy, and compare it with the conventional method for imbalanced data to evaluate and compare the machine learning model construction method including dataset creation. And verified its effectiveness.

Keywords: annotation, imbalanced data, training data, machine learning, natural language processing

1. はじめに

近年 SNS の利用が増えており，特に Twitter[*1]は，誰もが自由に思ったこと考えたことを，ツイート（つぶやき）と呼ばれる短文投稿機能によって発信できる，利用者も多く知名度も高い SNS の 1 つである．ツイートの中には商品やサービスに対する批評や要望などの「意見」も含まれており，このような消費者の生の声を収集，分析する「ソーシャルリスニング」という手法は，企業が消費者のニーズを把握し，マーケティング戦略を決定するための重要な手

[1] 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
[a)] sakuraiy@meiji.ac.jp

[*1] https://twitter.com

段となっている．実際，2013 年の NTT データ経営研究所の企業に対するアンケート調査 [1] においては，「自社の商品・サービスに関する投稿数やポジティブ・ネガティブ件数の定量的な把握を実施していますか」という設問に対し，回答総数 408 件のうち「把握している」が 39.5%，また「把握したい」という回答を含めると 63.8%に達しており，多くの企業がソーシャルリスニングを重視していることが分かる．

しかし，Twitter は誰もが自由にツイート（投稿）できるため，ツイート数が膨大であり，多様な意見が含まれる反面，企業側にとって重要な「意見」の占める割合は少ない．そのため，収集した膨大なツイート集合を 1 つずつ確認しながら，「意見」ツイートを人手により探し出すことは，時間的なコストが非常に大きくなる．そのため，SNS からの意見抽出の研究が進められている [2]．

しかしながら，人の言語表現は多様であり，単純な表現パターンのみで「意見」を判別することは難しい．このような課題に対しては，事例データに基づいた機械学習による判別が行われている．この事例データは，教師データと呼ばれ，例題と答えについてのデータである．Pak ら [3] は，実際の Twitter 投稿から教師データを作成し，感情分析のための機械学習モデルを構築している．このように，Twitter は自然言語処理のための教師データ，コーパスとして様々な研究に利用されている．

意見抽出モデルの教師データ作成をする場合，収集した各ツイートに対して，「意見である」と「意見でない」のラベル（答え）を付与する作業が必要になる．このラベル付け作業は「アノテーション（注釈付け）」と呼ばれ，そのラベル精度は．そのまま機械学習モデルの精度に大きく影響するため，重要な工程である．

しかしながら，意見抽出の場合，実際に Twitter からランダムにサンプリングしてきたツイート集合にアノテーションを行うと，「意見である」ラベルの付与されるツイートの全体に占める割合は，「意見でない」ラベルの付与されるツイートと比べて圧倒的に小さくなり，分類する各クラスのデータ数の比率が偏る．これをそのまま学習すると，モデルの精度が著しく落ちることが知られている．これは，不均衡データ問題と呼ばれ，実問題への機械学習適用において頻繁に起こる課題である．

このような不均衡データに対するアプローチとしては，大きく以下の 2 つが提案されている（**図 1**）．

- サンプリングベースのアプローチ：オーバサンプリングやアンダサンプリングなど教師データの削除をしたり，増やしたりすることでデータの割合を調整，均衡化するデータに対しての手法 [4], [5]
- モデルベースのアプローチ：モデルの学習時に，損失関数などの重みをデータ数の偏りに応じて調整する，学習アルゴリズムに対しての手法（Cost Sensitive Learning）[6], [7]

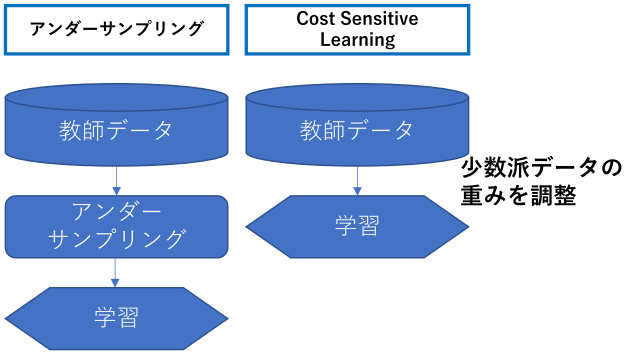

図 1　アンダサンプリングと Cost Sensitive Learning の比較
Fig. 1　Comparison between undersampling and cost sensitive learning.

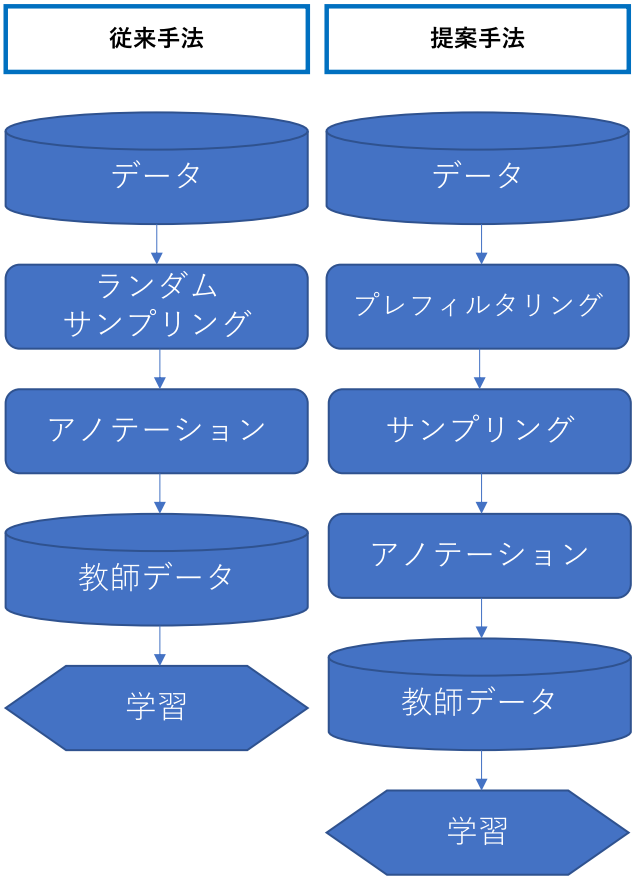

図 2　従来手法と提案手法の比較
Fig. 2　Comparison between conventional method and proposed method.

これらの手法は，アノテーション済みの不均衡な教師データに対するアプローチであるが，本論文では，アノテーションによる教師データ構築段階での不均衡データ対策手法を提案する．具体的には，アノテーション作業の対象となるデータ収集時に機械的なプレフィルタリングを用いたサンプリングにより分類クラスのデータ偏りを調整しながら収集することで，不均衡化を緩和する手法 PSSA（Prefilter based Stepwise Sampling for Annotation）を提

案する．図 2 に提案手法の従来とのアプローチの違いを示す．

また，この PSSA を意見抽出モデルの構築に適用するため，辞書マッチングによるプレフィルタを用いた PSSA を提案，実際にアノテーションとモデル構築を行い，精度検証を通して，その有用性を検証する．

本論文の構成は以下のとおりである．以下，2 章では，不均衡データ対策手法と意見抽出手法，教師データ構築についての関連研究について説明する．3 章では提案するアノテーション手法とその意見抽出モデルについて述べ，4 章では，アノテーション実験，学習モデルの比較実験，フィルタリングによる影響分析の結果を示し，提案手法の有効性について評価する．最後に，5 章で本論文をまとめる．

2. 関連研究

本章では，提案するアノテーション手法と意見抽出モデル構築に関する既存研究について述べる．

2.1 不均衡な教師データに対するアプローチに関する研究

分類問題において，各クラスのデータ数が均衡な教師データで学習した分類モデルは，不均衡な教師データで学習した分類モデルよりも分類性能が高いとされる [8]．さらに，He ら [9] によると，分類問題において，教師データが不均衡データであると，多数派クラスのデータに偏り，過学習するという問題があるとしている．

例をあげると，*Positive* のデータ数が 1，*Negative* のデータ数が 99 の計 100 件の不均衡データを，特に対策をせずに教師あり学習させると，すべてのデータを *Negative* と予測しても正解率（*Accuracy*）は 0.99 となり，非常に高い値が得られる．一方で，実際に *Positive* であるデータのうち，モデルによって *Positive* と予測したデータの割合を示す再現率（*Recall*）は 0 になる．また，モデルが *Positive* と予測したデータのうち実際に *Positive* であったデータの割合を示す適合率（*Precision*）も 0 になる．反対にこのモデルが完全に正しく予測できるとすれば，*Accuracy* は 1 となり，*Recall* も 1，*Precision* も 1 になる．このように不均衡なデータを学習した分類モデルに対する評価指標は正解率（*Accuracy*）だけでなく，再現率（*Recall*），適合率（*Precision*）が重要であることが分かる．

本論文で使用するモデルの評価指標は *Accuracy* のみではなく *Precision*，*Recall* も利用する．また，*Prescision* と *Recall* は相反の関係にあることから，*Precision* と *Recall* との相乗平均である，*F* 値（*F-measure*）も評価指標として利用する．以下にそれぞれの評価指標の数式を示す．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F\text{-}measure = \frac{2Recall \cdot Precision}{Recall + Precision}$$

TP ：*True Positive* 正しく *Positive* と予測

FP ：*False Negative* 誤って *Positive* と予測

FN ：*False Negative* 誤って *Negative* と予測

TN ：*True Negative* 正しく *Negative* と予測

このような不均衡な教師データから学習させる様々な手法が提案されている．Haixiang ら [10] は不均衡データに対する対応手法として，大きく，データの前処理と Cost Sensitive Learning の 2 つをあげている．

データの前処理に関する手法では，多数派データと少数派データのデータ数のバランスを再調整するサンプリングによるアプローチが多く用いられている．サンプリング手法は多数派データ数を少数派データ数に近づけるアンダサンプリングや少数派データをかさ増しし，多数派データ数に近づけるオーバサンプリング，またこれら 2 つを同時に利用したハイブリッド方式があげられる．

ランダムアンダサンプリングは多数派データからランダムにサンプリングすることによって少数派データとのバランスを調整する手法で，最も単純で多く用いられている [10]．また，この改良手法として，クラスタベースによるサンプリングや距離ベースでのサンプリング手法が提案されている [11], [12]．Yen ら [11] は不均衡データセットに対し，クラスタリングを行い，各クラスタ内の多数派データと少数派データのバランスを考慮して，各クラスタから多数派データのサンプリングを行い，他のアンダサンプリング手法と比べて優れていることを示した．また，Ku ら [12] は，ランダムアンダサンプリング手法は単純であり使いやすいとする一方で，有用なデータが失われ，過学習する可能性があるとし，ファジイ距離ベースでアンダサンプリングを行う手法を提案している．

また，オーバサンプリングの手法である SMOTE はランダムなオーバサンプリングと比較してより精度が高く，過学習が起きにくいとされている [10]．紺野ら [13] は深層学習を用いて少数派のデータをかさ増しする手法を提案し，画像分類において高い精度を示した．文書分類などの自然言語処理ではデータのかさ増しを行うオーバサンプリングの手法は難しいとされていたが，澤崎ら [14] は単語の入れ替えと文節の入れ替えを行うことによって，少数派の文書データのかさ増し手法の提案を行った．

本研究で提案する PSSA はデータに対する不均衡データ対策のアプローチであるが，アンダサンプリングやオーバサンプリングのように既存の不均衡なラベル付きデータセットから学習に用いる教師データをサンプリングする手法ではなく，収集したデータから不均衡化が緩和された

ラベル付きデータセットを作成するアノテーション手法である．このため，PSSA はアンダサンプリングや Cost Sensitive Learning などの不均衡化した教師データに対する対策手法と組み合わせて利用することができる．

2.2　意見抽出の関連研究

乾ら [15] は意見抽出研究の題材となるテキストの種類について以下のようにまとめている．

- 意見の収集，集約が目的となっているテキスト
  - 社会調査などによる自由回答アンケート
  - カスタマーサポートセンターにおける「お客様の声」
  - レビュー
- 潜在的に意見を含むテキスト
  - チャット
  - Web 掲示板
  - Weblog

前者は「意見抽出」自体が目的となっているテキスト集合のため，有用な情報が比較的多い傾向がある．前者のようなテキストから意見表現などを抽出する研究は多い．小林ら [16] は「意見」を〈対象，属性，評価値〉という 3 つ組で定義し，意見表現を，共起パターンを用いて収集する手法を提案している．佐野ら [17] は Web 経由で収集した地域苦情データに対して，経験的パターンや係り受け関係，機械学習の手法を用いて要望表現の識別を行い，Yu ら [18] は，感情分析を行うために，ニュース記事から論説などの「意見」と「事実」にそれぞれ分類し，Bayes 分類モデルを用いた意見抽出を提案している．Pang ら [19] は感情分析に用いるため，映画のレビューを主観に基づく表現を含むテキストと事実を述べているテキストを分類し，機械学習の手法を用いて「主観抽出」を行っている．

一方で，Twitter などの SNS を含む後者は「意見抽出」を目的とされたテキスト集合ではないため，有用なデータが非常に少なくなる．立石ら [20] はインターネット上からの意見を効率的に抽出する研究がこれまで存在しなかったことを明らかにしたうえで，評価表現辞書を用いて意見を抽出するシステムを提案した．また，橋本ら [21] は Twitter から政策やサービスへの評判や評価の傾向を予測する研究を行った．川島ら [22] は Twitter からの意見抽出に関する研究で，辞書と半教師あり学習の Distant supervision の手法を用いて教師データを半自動的に構築し，SVM を用いて分類，抽出した．立石ら [20]，橋本ら [21]，川島ら [22] は評価情報を含む表現のある文章を，辞書を用いて抽出しているが，本研究では辞書に含まれている意見表現が明示されているツイートだけでなく，辞書に含まれていないが，文脈上などで意見とされるツイートに対してもサンプリングを行い，教師データを作成する手法を提案する．

また，前者のような意見収集が目的のテキストでは主語や述語など文法的に整っているテキストが多い傾向にあるが，後者のようなテキストでは文法的に誤っているなど，テキスト自体の問題も多く [23]，共起パターンや係り受け関係など文法的な抽出のみでの高精度での分類は困難である．

2.3　データセット作成に関する研究

特定の情報源からデータを収集し，アノテーションを行ってデータセットを作成する，教師データセット構築についての研究は数が少ない．

筒井ら [24] は公開されている地方議会の会議録のデータを政治学や言語学や情報工学などの様々な学問分野において研究対象になっていることを明らかにしたうえで，これらのデータの形式がそれぞれ異なっており，収集作業に労力がかかることなどの問題を取り上げ，これを解決するためにコーパス構築を行っている．宮崎ら [25] は Web 上のレビュー文書から 10 名のアノテータによって注釈付き評判情報コーパスを構築している．この研究のなかで，宮崎らは各アノテータによる判断の一致率が低いことを問題として取り上げ，アノテーション事例参照利用を提案し，判断の揺れ削減に効果があると報告した．本研究でもアノテーション時にアノテータに事例を複数あげることによって，教師データの品質の確保を行った．

アノテーション作業対象のデータ数が多くなると，それにともなってアノテータを増やすことが必要である．組織によっては労働力の確保が難しいことがある．これを解決するために，アノテーション作業をクラウドソーシングし，データセットを構築した研究が存在する [26]．しかし，クラウドソーシングにおけるアノテーション作業者（アノテータ）はやる気にばらつきがあり，報酬目的の不誠実な作業者も存在することから低品質な成果物がもたらされることがあり，それを仕組み的に改善する研究が行われている [27]．本研究でもアノテータに人数を十分確保することが難しいため，クラウドソーシングによるアノテーションを行っている外部企業に実際のラベリング作業の一部を委託した．

3.　提案手法

本章では，あるデータ集合から教師データ（ラベル付きデータセット）を作る際，ランダムにラベル付けするデータを選択すると不均衡データになってしまう場合に，フィルタリングを活用して段階的にサンプリングすることで，データの不均衡化を緩和するアノテーション手法，PSSA（Prefilter based Stepwise Sampling for Annotation）を提案する．また，PSSA を意見抽出問題へと適用した辞書フィルタを用いた PSSA，これを用いた意見抽出モデルの構築手法を提案する．

分類問題の教師データを作る際は，各クラスのデータ数が均衡している方が精度の高いモデルが構築できる．しか

しながら意見抽出問題の場合，たとえばTwitterからランダムに事例を集めてくると，「意見」の事例は「意見でない」事例より圧倒的に少なくなってしまう．このような自然言語処理では，辞書によるキーワードマッチングがよく使われており，意見の文書に共通するような言語表現を集めて辞書を作り，マッチングすることで，かなりの意見が抽出可能である．このように辞書フィルタを使えば，意見の事例を集めることができる．しかしながら，これで集められた事例は，単純な事例ばかりで，文脈で表現された複雑な意見の事例が集められない．

そこで，すべてが「意見」でなくても良いので，意見が多く含まれるような緩いフィルタを利用することで，「意見」の含まれる割合を上げることを考える．また，そのフィルタも実際に利用してみなければその効果が分からないため，絞込条件が異なる複数のフィルタを段階的に実施することで，不均衡度合いを緩和しつつ，分類精度の向上に重要な事例も含まれるような事例サンプリングを実現する．

3.1 PSSA（Prefilter based Stepwise Sampling for Annotation）

PSSAは，ランダムサンプリングすると一部の分類クラスのデータ割合が小さくなってしまう場合に，少数クラスのデータ割合が増えるような機械的フィルタを利用し，サンプリングする条件を絞り込むことで，データの不均衡化を緩和する．しかし，機械的フィルタにより絞られたデータは，中に含まれる学習データのパターンが単純化するなどの悪影響が考えられる．そこで，絞込み効果の異なる機械的フィルタを複数用意し，（悪影響の少ない）弱い効果のフィルタから順に適用し，段階的にアノテーション対象となるデータをサンプリングしていくことで，不均衡化を緩和した教師データを構築する手法である．機械的フィルタには，ルールベース，辞書マッチングなどの条件ベースのものから，学習済みの弱学習器など追加の教師データが不要であれば適用可能である．実際のフィルタについては次節で説明する．以下，PSSAの詳細な手順について述べる．

事例をランダムにサンプリングを行い，アノテーションを行うと不均衡な教師データとなってしまう問題に対して，アンダサンプリングなどの手法では少数派ラベルのデータ数に多数派ラベルのデータ数を合わせるため，多数派ラベルの教師データのほとんどが学習されず，効率が悪くなるという問題がある．また，教師データを作成するアノテーションは時間的コストがかかるため，効率の良いデータサンプリング手法が求められる．

PSSAは最初に，事例をランダムにサンプリングすると少数派となるデータ（以下，少数派ラベルデータとする）の特徴をプレフィルタとして構築する．構築したそれぞれのプレフィルタをすべてのデータに適用し，それぞれプレフィルタリングで *Positive* と判別されたデータ集合，プレフィルタリングで *Negative* と判別されたデータ集合の2つに分ける．*Positive* と判別された集合と *Negative* と判断された集合から一定の割合でサンプリングを行い，アノテーションを行い，教師データを作成する．また，プレフィルタ，サンプリングの割合を複数構築し，データ分布への悪影響が少ない，絞込効果の弱いプレフィルタから順に適用し，上記の内容を繰り返す．これによってプレフィルタリングによって少数派ラベルデータを多く抽出できる．また，プレフィルタリングで *Positive* と判別された集合とプレフィルタリングで *Negative* と判断された集合両方からサンプリングを行うことによって教師データの特徴がプレフィルタに偏ることを防ぐことができる．

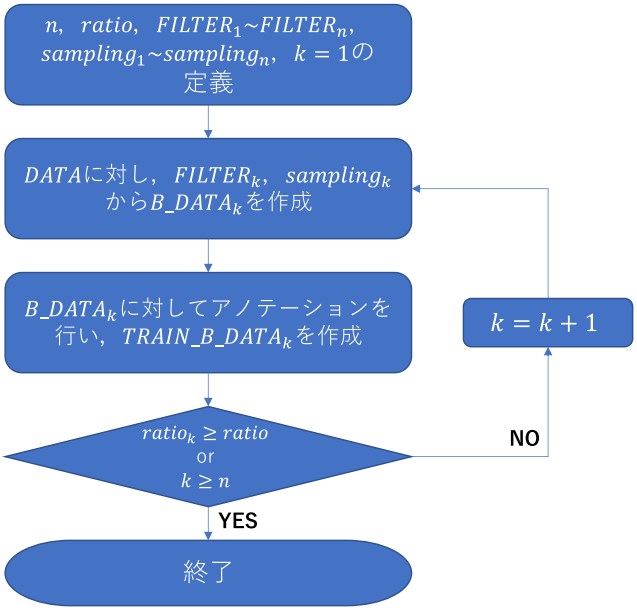

図 3　PSSA の処理の流れ
Fig. 3　PSSA processing flow.

PSSAの処理フローを**図 3** に示すとともにその処理手順①～④を以下に示す．

なお，ブロック数（n）とは，適用するプレフィルタの数であり，プレフィルタによるサンプリングを行う回数である．目標値（$ratio$）とは，少数クラスの全体に占めるデータ割合の目標とする割合である．

① ブロック数（n），教師データすべてに含まれる少数派ラベルデータの割合の目標値（$ratio$），プレフィルタ（$FILTER_1$～$FILTER_n$）とサンプリング割合（$sampling_1$～$sampling_n$）の組をあらかじめ定義する．また，$k=1$とする．

② すべてのデータ（$DATA$）に対し，①で定義したプレフィルタ（$FILTER_1$～$FILTER_n$）のうちの $FILTER_k$ でプレフィルタリングを行った後，$sampling_k$ でサンプリングを行い，ブロックデータ（B_DATA_k）を作成する．

③ B_DATA_k にアノテーションを行い，教師ブロックデー

タ（$TRAIN\_B\_DATA_k$）を作成する．

④ 作成したすべての教師データ（$\sum TRAIN\_DATA_k$）のうち，少数派ラベルデータが占める割合（$ratio_k$）が事前に定義した $ratio$ に到達する，または k が n に達するまで k を 1 ずつ増やしながら②，③を繰り返す．

3.2　辞書フィルタを用いた PSSA による意見抽出モデル

本章では，PSSA を用いた意見抽出モデルの構築について説明する．プレフィルタとしては，評価表現辞書によるマッチングを用いた．モデルの構築には，PSSA により作成された教師データを用いて学習するモデルの選択が必要だが，本手法は一般的な教師あり機械学習モデルすべてに適用可能である．

3.2.1　Twitter からの意見抽出

ツイートの収集を行い，「意見」の定義に基づいてアノテーション作業を行い，意見抽出モデルのための教師データを構築する．アノテーション作業には膨大な人手と時間的コストがかかるため，より信頼性の高い教師データの構築手法が求められる．しかし，従来手法では収集したツイートからランダムサンプリングを行ってアノテーション作業を行うことによって教師データを構築するが，この手法は Twitter などの SNS の特性上，「意見」ではないツイートが非常に多く抽出され，不均衡データになり，学習が難しくなる可能性が非常に高い．

3.2.2　「意見」の定義

本研究で提案する意見抽出システムの目的は企業がマーケティングに利用する上で「有用」な「意見」を抽出することであり，一般的に「意見」とされる定義の中でも「要望」や具体的な「批評」を述べているツイートを抽出することが必要である．また，アノテータの負担増によって教師データの信頼性が失われることを防ぐために必要最小限で定義を行うことも必要である．

国立研究開発法人情報通信研究機構旧知識処理グループ情報信頼性プロジェクトによって開発された意見（評価表現）抽出ツール（以下ツールとする）[*2]では「意見」を「感情」,「批評」,「メリット」,「採否」,「出来事」,「当為」,「要望」と定義しているが，本研究はこれらのうち，「意見」を「感情」「批評」「要望」に限って定義する．また，このうち「感情」と「批評」はツールでは「主観的でかつ，感情的な評価表現」と「主観的ではあるが，感情的ではない評価表現」が類似しているため，本研究においては「感情・批評」の 1 つにまとめ，「発言者の主観的な感情や批評のある表現」として，「感情・批評」「要望」「その他」の 3 つのラベルを指定してアノテーション作業を行った．しかし，「感情・批評」は非常に広範囲に定義されている表現であり，たとえば「ディズニー行きたい」や「USJ 楽しい」などの「なぜそのような感情や批評に至ったのか」が示されていないような，マーケティング上有用ではないツイートが「意見」として抽出されてしまう可能性が高い．これを防ぐため，「感情・批評」を「根拠・理由の示されている発言者の主観的な感情や批評のある表現」とすることによって該当する表現を限定した．

「要望」の定義はツールでは「要望を表す評価表現」と，具体的に定義されていないため，川島ら [19] の研究を参考にした．川島ら [19] は Twitter からの要望抽出を行う研究で「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」,「希望非断定」,「当為」,「当為非断定」,「不満」を要望表現と定義しているが，本研究では前述のように必要最小限で定義を行うため，これらのうち最も普遍的な表現である，「発言者が相手に強制するする表現」とする「命令」,「発言者が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」とする「依頼」を引用した．

以下で具体例を示しながら「意見」として抽出する「感情・批評」,「要望」の定義を整理する．

・「批評・感情」の定義

理由が示されている感情・批評：根拠・理由の示されている発言者の主観的な感情や批評のある表現

　例：今度ハロウィンイベントがあるからディズニー行きたい

　例：キャストさんが親切でディズニー楽しかった

　例：USJ で財布を落としたら親切な人が拾ってくれたみたい．本当に感謝．

・「要望」の定義

命令：発言者が相手に強制するする表現

　例：ディズニーは早く新しい作品を出せ

依頼：発言者が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：ディズニーの年パスもう少し安くしてほしい

　例：ディズニーのキャストは再度教育すべき

また，客観的な事実を述べた「出来事」や，発言者が個人ではなく，広告である「広告」や前述した「理由の示されていない感情・批評」はマーケティング上有用性が低いため本研究で「意見」に分類しないとした．具体例を以下に示す．

出来事：客観的な事実を述べた表現

　例：ディズニー行ってきた

　例：ディズニーの年パス買ってきた

広告：発言者が個人ではなく，広告である表現

　例：ディズニーのチケットが今だけ 30%引き！

理由が示されていない感情・批評：なぜその感情になったのか，書かれていないもの．

　例：ディズニー行きたい

　例：ディズニー楽しかった

　例：○○さんと一緒に行ったディズニー楽しかった

[*2] https://alaginrc.nict.go.jp/opinion/index.html

3.2.3　評価表現辞書を用いたプレフィルタリング

意見表現をまとめた辞書を構築し，それを収集したツイートにプレフィルタリングして抽出されたツイートにアノテーション作業を行う立石ら [20]，橋本ら [21] の研究や抽出されたツイートに半教師あり学習の Distant Supervision の手法を適用して網羅的に教師データを収集した川島ら [19] の研究が存在するが，辞書に収録されている表現を含むツイートのみが教師データとして学習され，辞書に収録されていない表現ではあるが「意見」として抽出されるべきツイートが教師データに含まれず，抽出漏れが起きる可能性がある．また，反対に辞書でプレフィルタリングして抽出されたツイートすべてが必ずしもアノテーション作業で「意見」と判定されるわけではない．

本研究では，評価表現辞書を用いたプレフィルタリングを用いる．この辞書には，小林が構築した評価表現辞書[*3]を用いる．この辞書は評価を表す可能性のある表現を集めた約 5,200 語からなる辞書であり，ある程度ドメイン横断的に使用可能としている．辞書には，「高い」，「安い」など特定の対象に対する評価の値を表す表現や，「好き」，「嫌い」などの書き手などの評価保持者の感情を表す表現が含まれている．

本提案手法では，評価表現辞書を基に下記のような 3 段階のプレフィルタを提案し，これを適用した PSSA を用いて不均衡化を緩和する．ブロック数 n の値を増やすほどプレフィルタリングの回数が多くなり，細かくフィルタリングを制御できるが，それだけ多くのフィルタを設計する必要があり，アノテーション作業を小刻みに調整する労力と作業時間がかかるため，適用する分野にもよるが，現実的には，$n = 3$〜5 程度を推奨する．

本研究課題である Twitter からの意見抽出タスクにおいては，プレフィルタに評価表現辞書を用いて，評価表現の品詞のバリエーションの違いと文字数の制限によって，プレフィルタを構築した．品詞の組み合わせ的には，$n = 10$ 段階以上のフィルタも可能だが，クラウドソーシングを活用してアノテーションを実施する際，日単位などの粒度での調整は難しいため，1 段階のフィルタの結果まとめを 2 週間程度で実施すると想定し，バッファ期間を含めて 2 カ月以内にアノテーションを完了するため，$n = 3$ 段階でのフィルタリングを構築した．

また，2 値分類問題において，それぞれのラベルのデータ数が全教師データに占める割合が 0.5 であれば最も精度が高いため，*ratio* の推奨値は 0.5 とする．ただし，ある程度の不均衡はアルゴリズムでの不均衡対策により，対処可能なので，目標値 *ratio* を 0.5 とすることが困難な場合には，少し小さめに設定し，教師データ中の事例が偏らないようにフィルタを設計した方が結果的に分類精度が高くなる場合も考えられる．本提案手法では，推奨値を利用することとし，PSSA のパラメータは以下のように定義する．

$$n = 3$$

$$ratio = 0.5$$

プレフィルタの絞込条件が細かくなると，「意見」ラベルのデータが抽出される割合が増える一方，教師データの事例の偏りも顕著になるため，PSSA の段階が進むにつれて，プレフィルタで抽出されたツイートの割合が低くなるようにサンプリング割合 *sampling* を決定した．

以下に提案手法の各段階のプレフィルタとサンプリングを示す．

・第 1 段階

$FILTER_1$：小林の評価表現辞書を収集したツイートにプレフィルタリングする．

$sampling_1$：抽出されたツイートと抽出されなかったツイートそれぞれ 7：3 になるようにサンプリングを行う．

・第 2 段階

$FILTER_2$：小林の評価表現辞書に MeCab [*4]で形態素解析を行い，「形容詞」，「副詞助詞類」，「助動詞」，「名詞副詞接続」，「名詞形容動詞語幹」に該当する評価表現のみを取り出した新しい辞書を作成する．また，収集したツイートに対して，新しい辞書でプレフィルタリングする．

$sampling_2$：抽出されたツイートと抽出されなかったツイートそれぞれ 8 : 2 になるようにサンプリングを行う．

・第 3 段階

$FILTER_3$：小林の評価表現辞書に MeCab で形態素解析を行い，「形容詞」，「副詞助詞類」，「助動詞」，「名詞副詞接続」に該当する表現のみを取り出した新しい辞書を作成する．また，収集したツイートに対して，新しい辞書でプレフィルタリングし，71 文字以上の文字数があるツイートのみをプレフィルタリングする．

$sampling_3$：抽出されたツイートと抽出されなかったツイートそれぞれ 5 : 1 になるようにサンプリングを行う．

以下に上記プレフィルタを用いた処理手順①〜③を示す．

① $k = 1$ として，収集したツイート（*DATA*）に対し，$FILTER_k$ と $sampling_k$ からブロックツイート（B_DATA_k）を作成する．

② B_DATA_k に対してアノテーションを行い，ラベル付きブロックツイート（$TRAIN\_B\_DATA_k$）を作成する．

③ 作成したすべてのラベル付きツイート（$\sum TRAIN\_DATA_k$）のうち，「意見」ラベルデータが占める割合が *ratio*（$= 0.5$）を超える，または k が n（$= 3$）になるまで k を 1 ずつ増やしながら上記①，②を繰り返す．

$FILTER_k$ の k が進むにつれて，よりプレフィルタの絞

[*3] http://www.syncha.org/evaluative_expressions.html

[*4] https://taku910.github.io/mecab/

込条件を細かくし，少数派ラベルデータである「意見」データが多く収集される．また，$sampling_k$ において，プレフィルタで抽出されたデータだけでなく，プレフィルタで抽出されなかったデータからもサンプリングすることによって，プレフィルタに偏らないデータを収集することができる．

4. 検証

本章では，提案手法の有効性を検証するため，5 つの検証実験を行った．まず実験 1 において，辞書フィルタを用いた PSSA による，不均衡化の緩和効果の検証のため，ツイートのアノテーション実験を行う．

次に実験 2-1，2-2 において，アノテーション段階での不均衡データ対策の有効性を検証するため，実験 1 で作成された教師データを用いて意見抽出モデルを構築し，アノテーション手法と前処理，機械学習構築手法の組み合わせの違いによるモデル精度の違いを検証する．

最後に実験 3-1，3-2 において，アノテーションを行うサンプル選択に辞書フィルタを用いることによる影響を分析するため，各辞書フィルタを適用した場合とフィルタリングしなかった場合のモデル精度を比較する．

4.1 実験条件

4.1.1 ツイートデータ収集・前処理

まず，アノテーション対象となるツイート収集について説明する．今回の実験では，意見抽出の目的を「大規模サービス施設におけるサービス改善に役立つ意見の収集」と設定した．ツイート検索を行うためのキーワードは，ツイート数が一定数以上存在するテーマパークや東京都内の 5 つ星ホテルなど合計 19 施設の名称，またはその略称とした．ツイート収集対象とした施設のキーワードと収集したツイート数を**表 1** に示す．表中の都内 5 つ星ホテルは「アマン」,「グランドハイアット」,「コンラッド」,「リッツ・カールトン」,「ペニンシュラ」,「シャングリラホテル」,「パークハイアット」,「マンダリンオリエンタル」の 8 施設を選定した．収集期間は 2018 年 5 月 1 日～2019 年 4 月 30 日までの 365 日間とした．なお，不要なツイートを除外するため，収集時に以下の前処理を行った．

- 重複しているツイート，リツイート（RT），リプライ（@付きツイート）は収集対象から除外した．
- URL が含まれているツイートは該当部分を「〈URL〉」の文字列に置き換えた．
- Python ライブラリ neologdn [*5] を利用して文字表現の正規化を行った．
- 大文字アルファベットは小文字に統一を行った．

neologdn は SNS などの日本語テキストにおいて表記ゆれの正規化を行う．具体例をあげると，全角英数字記号や半角カナなどを半角英数字記号，全角カナへ変換，また「スーパーーーー」や「無駄無駄無駄ァ」などの連続して重複する文字列を「スーパー」や「無駄ァ」などへ正規化を行う．

その後，収集したツイートに対して MeCab を利用して分かち書きを行った．また，MeCab の辞書は標準の IPADIC と Web 上の新語が追加されている NEologd [*6]（2019 年 12 月 5 日更新時点）を利用した．収集されたツイート数を表 1 に示す．

表 1 のように，「ディズニー」と「USJ，ユニバ」をキーワードとして収集したツイート件数がそれら以外のキーワードの件数と比べて非常に多いことが分かった．そこで，特定のキーワードに対するツイート件数の割合が偏ることを防ぐため，収集したツイートデータを「ディズニー」,「USJ，ユニバ」（以降「USJ」とする）,「その他」の 3 つのドメインに分割し，収集したツイートから教師データを構築する際に，各ドメインのツイート件数の割合が極端に偏らないように調整を行った．**表 2** に 3 つのドメインに集約した後の各ドメインの収集ツイート件数を示す．

表 1　各キーワード名と収集したツイート数
Table 1　keyword and number of tweets collected.

キーワード名	ツイート数
ディズニー	83100
USJ，ユニバ	31346
ピューロランド	2944
ハウステンボス	1544
(5つ星ホテル)	1108
レゴランド	984
キッザニア	798
ハワイアンズ	549
東京ドイツ村	470
サマーランド	306
日光江戸村	112
リニア鉄道館	59
合計	123320

表 2　各ドメインのツイート数
Table 2　Number of tweets for each domain.

ドメイン名	ツイート数
ディズニー	83100
USJ	31346
その他	8874

[*5] https://github.com/ikegami-yukino/neologdn

[*6] https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md

4.1.2　実験環境

ここでは，実験に使用した計算機のスペックやプログラミング言語，使用したパッケージのバージョンについて述べる．

まず，プログラムを実行する計算環境のスペックは以下のとおりである．

CPU：Intel XeonW-2123（3.60 GHz，4Core）
メモリ：128 GB
GPU：NVIDIA TITAN V
OS：Ubuntu 16.04 x86_64

プログラム実装に用いた言語は Python 3.7.4 である．また，今回の実験に際して使用した Python パッケージとそのバージョンは以下のとおりである．なお，実行における仮想環境構築には Anaconda を使用した．

scikit-learn 0.22
imbalanced-learn 0.6.1
numpy 1.16.5
pandas 0.25.1
mecab-python3 0.996.3
neologdn 0.4

4.2　教師データの不均衡化緩和効果の検証

収集したツイートデータに PSSA を適用してアノテーションをすることで教師データを作成し，その不均衡化の緩和効果を検証する．

4.2.1　実験 1

ツイートデータから意見分類のための教師データ作成実験を行う．本実験において，提案手法である辞書フィルタによる PSSA を用いた場合と，従来手法であるランダムサンプリングを用いた場合の教師データの不均衡度合いを比較することで，PSSA による不均衡化緩和の効果を分析する．

今回，PSSA は 3 段階のフィルタリングを用いる．各段階において，アノテーション対象としたツイート件数と各ドメインのデータ数を**表 3** に示す．

次に比較対象として，従来手法であるランダムサンプリングにより作成される教師データを構築する．これは，PSSA の各段階における辞書フィルタによって抽出された意見ツイート候補の全体割合（**表 4**）の分布に基づいて，PSSA で作成した教師データからもっともらしいツイート数になるように作成した．従来手法によって作成されたツイート件数を**表 5** に示す．

アノテータは 20 代から 60 代までの男女 35 人で，与えられたツイートに対し，「感情・批評」，「要望」，「その他」の 3 カテゴリのラベリング作業を行った．「感情・批評」と「要望」両方に該当する表現があった場合は両方にラベリングを行った．ラベリングの判断がアノテータで異なった場合，該当アノテータ間で定義や事例を改めて確認し，判断の統一を行った．

表 3　PSSA の各段階でアノテーション対象としたツイート件数
Table 3　Number of tweets targeted for annotation.

PSSA段階	ディズニー	USJ	その他	合計
第1段階	764	770	780	2314
第2段階	246	248	256	750
第3段階	1200	1200	1200	3600
合計	2210	2218	2236	6664

表 4　PSSA の各段階における意見ツイート候補の割合
Table 4　Percentage of extracted tweets.

フィルタ	割合
FILTER1	0.6427
FILTER2	0.4549
FILTER3	0.1911

表 5　従来手法によって作成された教師データ件数
Table 5　Number of teacher data created by the conventional method.

PSSA段階	ツイート数
第1段階	1928
第2段階	727
第3段階	740
合計	3395

表 6　PSSA によって作成された教師データのうち，「意見」ラベルが占める割合
Table 6　Percentage of opinion labels.

手法	PSSA段階	割合	ツイート数	意見数
PSSA	第1段階	0.2035	2314	471
	第2段階	0.1613	750	121
	第3段階	0.2822	3600	1016
	合計	0.2413	6664	1608
従来手法	第1段階	0.2293	1928	442
	第2段階	0.1155	727	84
	第3段階	0.2014	740	149
	合計	0.1988	3395	675

4.2.2　実験 1 の結果と考察

アノテーション実験により求められた，PSSA で教師データを作成した場合と従来手法で教師データを作成した場合の「批評」・「要望」ラベル（以下，「意見」ラベルとする）の教師データが全教師データに占める割合を以下の**表 6** に示す．

表 6 の結果から，PSSA を用いて作成した教師データの方が従来手法で作成した教師データよりも意見の割合が高くなり，不均衡度合いが改善されることを確認した．

段階ごとに見ると，第 1 段階では PSSA 側が従来手法

側よりも意見の割合が小さいが，第 2 段階，第 3 段階では PSSA 側の意見の方が割合が大きくなった．第 1 段階で PSSA 側の意見の割合が小さくなった原因として，第 1 段階のプレフィルタ $FILTER_1$ が，ランダムサンプリングと比べてもそれほど意見の絞込効果が低い可能性がある．実際，表 4 に示したように，$FILTER_1$ により抽出された意見ラベルの候補は 0.64 と高い．また $sampling_1$ も 7 : 3 に設定したことからプレフィルタが強く作用しなかったことがあげられる．

$FILTER_2$ では，より絞込条件が細かくなったことから，第 2 段階での意見が全体に占める割合が PSSA 側は従来手法側よりも高い値を示したと考えられる．そのため，$FILTER_1$ をより絞込条件が細かいものに変えることで従来手法側よりも高い値が出せると考えられる．また，$sampling_1$ を 8 : 2 などプレフィルタで抽出されたツイートの割合をさらに増やすことでも従来手法側よりも高い値が出せると考えられる．本実験では，後の実験において，フィルタリングとランダムサンプリングとの比較を行うため，ランダムサンプリンによるデータを一定数確保する必要があり，$sampling_1$ を 8 : 2 としたが，その比較が不要な場合は，$sampling_1$ を 9 : 1 などに設定することで，より不均衡の緩和が行える．

4.3　意見抽出モデルの精度向上効果の検証

アノテーション段階からの不均衡データ対策手法である PSSA を適用することによる，意見抽出モデルの精度向上効果の検証を行う．実験 1 で作成された教師データを用いて意見抽出モデルを構築し，アノテーション手法と前処理，機械学習構築手法の組合せの違いによるモデル精度の違いを明らかにすることで，PSSA の優位性を確認する．

4.3.1　実験 2-1

教師データを，辞書フィルタを用いた PSSA により作成した場合と，ランダムサンプリングにより作成した場合，それぞれのデータセットに対して，不均衡データ対策（Under-sampling，Cost Sensitive Learning）を行ってモデル化した場合，対策を行わずにモデル化した場合の分類精度を評価した．

まず，データセットの作成方法について述べる．最初に 4.2.1 項で作成した 6,664 件のデータから表 4 の割合の分布に基づいて，ランダムサンプリングにより選択された場合のデータを擬似的に 666 件作成し，これをテストデータとした．次に 6,664 件のデータからテストデータ 666 件を除いたデータ（以降 PSSA-Large データとする）から，表 4 の分布に基づいて擬似的に従来手法（ランダムサンプリング）による教師データ（以降 baseline データとする）を作成した．また，baseline データのデータ数と同じ数を PSSA-Large データからサンプリングを行ったデータ（以降 PSSA-Base データとする）を作成した．**表 7** に 3 種類のデータセットのデータ数を示す．

表 7　実験 2-1 の各教師データ数と意見割合
Table 7　Number of teacher data.

データセット	割合	ツイート数	意見数
PSSA-Large	0.2446	5998	1467
PSSA-Base	0.2075	2275	472
baseline	0.1837	2275	418

次に，各データセットに共通で適用したベクトル化までの前処理について説明する．

自然言語のベクトル化を以下の手順で実施した．

- ツイート内に記号が含まれている場合は半角スペースに置き換えた．
- 半角カナは全角カナに置き換えた．
- MeCab で形態素解析を行い，助詞を削除した．

次に教師データは文書データであるため，各ツイートを TF-IDF の手法を用いて特徴量化を行った．TF-IDF は文書内に含まれる単語の重要度に基づいてベクトル化する手法であり，以下の数式で表される．

$$tfidf_{a,X} = tf_{a,X} \cdot idf_a$$

$$tf_{a,x} = \frac{\text{文書 X における単語 a の出現回数}}{\text{文書 X における全単語の出現回数の和}}$$

$$idf_a = \log\left(\frac{\text{全文書数}}{\text{単語 a の出現する文書の数}}\right)$$

tfidf を計算することによって，各ツイートをベクトルとして表すことができる．PSSA-Large，PSSA-Base，baseline それぞれのデータセットとテストデータを結合し，3 つのデータセットを新たに作成し，各データセットにおいてツイートの *tfidf* を求めた．

次に特徴量化を行った教師データに対して scikit-learn [*7] の TruncatedSVD [*8] を用いて 1,000 次元へ次元圧縮を行った．

以上の手順で作成された特徴ベクトルを入力とし，SVM（ハイパーパラメータは scikit-learn のデフォルトの設定）による意見抽出モデルを以下の 3 種類構築し，精度評価を行った．

- Normal：不均衡データ対策なし
- Under Sampling：アンダサンプリング（多数派クラスのデータ数を少数派クラスのデータ数まで削減）を適用
- Cost Sensitive Learning：損失関数の定義において，少数派クラスの誤分類のペナルティを多数派クラスの誤分類のそれよりも重くするコスト考慮型学習を適用

精度評価は前述の *Accuracy*，*Precision*，*Recall*，*F-measure*，学習と予測の合計時間で行った．また，アン

[*7] https://scikit-learn.org/stable/
[*8] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

表 8　ランダムサンプリングとの比較評価
Table 8　Accuracy evaluation for comparison with random sampling.

手法	評価指標	Normal	Under Sampling	Cost Sensitive Learning
PSSA-Large	Accuracy	0.8694	0.7775	0.8679
	Precision	0.8750	0.4376	0.6496
	Recall	0.3415	0.7154	0.6179
	F-measure	0.4912	0.5430	0.6333
	Time	35.03	11.93	40.01
PSSA-Base	Accuracy	**0.8363**	0.6949	0.8153
	Precision	**1.0000**	0.3487	0.5000
	Recall	0.1138	**0.7496**	0.4390
	F-measure	0.2044	**0.4758**	0.4675
	Time	7.09	1.83	7.70
baseline	Accuracy	0.8318	0.6931	0.8153
	Precision	**1.0000**	0.3444	0.5000
	Recall	0.0894	0.7317	0.4146
	F-measure	0.1642	0.4684	0.4533
	Time	7.11	**1.53**	7.47

ダサンプリングは 5 回行った平均の値をとった．以下の**表 8** にそれぞれのモデルの評価結果を示す．

4.3.2　実験 2-1 の結果と考察

まず，ランダムサンプリング（baseline）と PSSA（PSSA-Base は baseline とデータ数を揃えたもの）の比較を行うと，すべてのモデルにおいて *Accuracy*，*F-measure* ともに baseline より PSSA-Base の評価値が上回っており，計算時間も大きく変わらないことから，PSSA の優位性が確認できた．なかでも PSSA のデータセットに Under Sampling を適用したモデルが最も性能が高かった．また，不均衡データ対策を実施しない場合（Normal）は，最も PSSA-Base と baseline の精度差が大きくなった．

また，Recall は Under Sampling の PSSA-Base が，Precision，Accuracy は Normal の PSSA-Base が，時間では baseline が一番高い結果となった．

また，データ数が大きい PSSA-Large では，最も性能が高かったのは，PSSA のデータセットに Cost Sensitive Learning を適用したモデルで，データ数が多いため時間は非常に長くなるものの，*Precision* と *Recall* の偏りも小さく非常にバランスの良いモデルとなっている．

Normal と Under Sampling，Cost Sensitive Learning を比較すると，全手法で Under Sampling と Cost Sensitive Learning の *Precision* が下がり，*Recall* が上がった．Cost Sensitive Learning の方が Under Sampling よりも *Precision*，*Recall* の変動は少なかった．

また，Normal の PSSA-Large では PSSA-Base と比較して *Precision* が下がり，*Recall* が上がっており，Under Sampling の PSSA-Large では *Precision* が上がり，*Recall* が少し下がっていることから，完全に不均衡化が緩和されている状態でツイート数を増やすと *Recall* は一定の数値で収束するが，*Precision* は上がると考えられる．

4.3.3　実験 2-2

次にランダムサンプリングではなく，クラスタリングベースのサンプリングを用いた教師データ作成手法との，精度比較を行った．これは，フィルタリングの代わりにクラスタリングを行い，各クラスタから均等にサンプリングする手法であり，辞書フィルタのような対象となるデータについての事前知識を必要とせず，距離計算が可能なベクトルデータであれば適用可能なことがメリットである．

今回，クラスタリング手法には，非階層的手法の代表手法である，k-means を利用した．収集したツイート集合に k-means を適用し，各クラスタから同じデータ数をサンプリングし，アノテーションを行った教師データの不均衡化の緩和効果の検証と，これを利用した意見抽出モデルの精度検証を行った．k-means のクラスタ数 k については，x-means で最適なクラスタ数を求めた結果，すべて 2 であったため，$k = 2$ を基本として，参考のため，$k = 3$，$k = 5$ の場合も実施した．

以下，教師データの作成手法について述べる．教師データは新たにアノテーションを行うのではなく，PSSA-Large データから擬似的に作成し，再現した．

最初に PSSA-Large データの段階ごとにフィルタで抽出されたツイートとフィルタで抽出されなかったツイートに分類した．次にそれぞれのデータに実験 2-1 と同様の前処理，TF-IDF による特徴量化，次元圧縮を行った．

次に k-means を適用し，各データ内でクラスタごとのデータ数が同じになるようにアンダサンプリングを行った．最後にこれらのデータを表 4 の分布に基づいてもっともらしいツイート数になるように調整し，教師データを作成した．

比較のため，本実験で作成したクラスタリングベースのサンプリングと教師データ数が同じになるように，PSSA-Large データや baseline データのデータ数を調整（ランダムサンプリング）した．**表 9** に各教師データのツイート数と「意見」クラスの割合を示す．

続いて，作成したそれぞれの教師データに対し，実験 2-1 と同様に意見抽出モデルの精度検証を行った．結果を**表 10** に示す．また，本実験ではアンダサンプリング手法として，Clustering besed UnderSampling を適用した．これは，通常のランダムなアンダサンプリングと異なり，クラスタリングを行い，得られた各クラスタから均等にアンダサンプリングを行う手法である．

4.3.4　結果と考察

まず，表 9 に基づき，意見の割合を比較する．最適なクラスタ数 $k = 2$ の場合は，クラスタリングベースのサンプリングは 0.17 とランダムサンプリング（baseline）の 0.15

表 9　実験 2-2 の各教師データ数
Table 9　Number of teacher data for experiment 2-2.

手法	*k=2*			*k=3*			*k=5*		
	割合	ツイート数	「意見」数	割合	ツイート数	「意見」数	割合	ツイート数	「意見」数
baseline	0.1593	653	104	0.1629	669	109	0.1789	587	105
PSSA	0.2848	653	186	0.2735	669	183	0.2572	587	151
クラスタリング	0.1776	653	116	0.1854	669	124	0.1431	587	84

表 10　クラスタリングベースサンプリングとの比較評価
Table 10　Accuracy evaluation of each model for comparison with clustering base sampling.

手法	評価指標	*k=2*			*k=3*			*k=5*		
		PSSA	Normal	クラスタリング	PSSA	Normal	クラスタリング	PSSA	Normal	クラスタリング
Normal	Accuracy	0.8213	0.8198	0.8183	0.8243	0.8213	0.8228	0.8228	0.8198	0.8168
	Precision	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Recall	0.0325	0.0243	0.0163	0.0487	0.0325	0.0407	0.0406	0.0243	0.0081
	F-measure	0.0629	0.0476	0.0320	0.0930	0.0630	0.0781	0.0781	0.0476	0.0161
	Time	2.41	2.35	2.42	2.46	0.82	2.49	1.99	2.08	1.82
Under Sampling	Accuracy	0.6997	0.6497	0.6276	0.6957	0.6306	0.6063	0.7126	0.6367	0.4051
	Precision	0.3363	0.2947	0.2978	0.3303	0.2909	0.2780	0.3450	0.2888	0.2237
	Recall	0.6406	0.6631	0.6263	0.6260	0.6796	0.6878	0.5951	0.6472	0.8537
	F-measure	0.4401	0.3986	0.4216	0.4320	0.4058	0.3946	0.4340	0.3970	0.3505
	Time	1.09	0.52	0.59	1.04	0.57	0.67	0.93	0.50	0.38
Cost Sensitive Learining	Accuracy	0.8188	0.8188	0.8153	0.8185	0.8168	0.8183	0.8168	0.8175	0.8153
	Precision	0.5492	0.6029	0.5000	0.5497	0.5333	0.5625	0.5200	0.5940	0.5000
	Recall	0.1257	0.0642	0.0244	0.1255	0.0650	0.0732	0.1056	0.0417	0.0163
	F-measure	0.2024	0.1144	0.0465	0.2008	0.1159	0.1295	0.1757	0.0773	0.0315
	Time	2.39	2.35	2.32	2.54	2.32	2.44	2.09	2.05	1.99

より割合が改善しているが，PSSA の 0.28 には遠く及ばない結果となった．また，クラスタ数が $k = 5$ の場合は，0.14 と baseline より悪化しており，クラスタ数に結果が影響されることが分かった．

次に，表 10 に基づいて，モデルの精度比較を行う．PSSA はランダムサンプリング（Normal）とクラスタリングベースのサンプリング（Clustering）と比べて，すべて F-measure が高くなっており，その優位性を確認することができた．Clustering は，最適なクラスタ数である $k = 2$ のとき，Under Sampling を適用したモデルが最も精度が高く，Normal より精度が高いが，$k = 5$ では逆に Normal より精度が落ちており，意見の割合の比較と同様に，クラスタ数に結果が影響されることが分かった．

また，PSSA，Normal，Clustering ともに，このデータ数では，Under Sampling との組合せが，精度が高くなっている．

4.4　プレフィルタによる影響の検証

PSSA のプレフィルタリングによって，教師データの不均衡化が緩和される一方，フィルタリングによって判別にとって重要な事例が教師データ中から失われる可能性がある．つまり，フィルタの性能によっては，判別が比較的簡単な事例ばかりになり，判別が難しい事例が含まれ難くなる危険性がある．本論文で用いた辞書フィルタの場合では，単純な意見表現による事例が増えてしまい，複雑な文脈により表現された意見の事例が少なくなることで，意見抽出モデルの精度が下がってしまう可能性がある．これらの悪影響を評価するため，各フィルタを適用した教師データとフィルタリングをしなかった場合を比較する．

4.4.1　実験 3-1

PSSA はプレフィルタにより分類クラス間のデータ不均衡を緩和することにより，モデルの精度を高める手法であるが，本実験では，この PSSA のプレフィルタによる悪影響のみを評価するため，すべてのモデルにおいて，アンダサンプリングを行い，均衡化を行ったうえでモデル構築を行う．これにより，各クラスのデータ割合は均等，教師データ数も同じ条件で，PSSA と従来手法（フィルタリングなし）のモデル精度の比較を行うことになる．違いは，教師データのなかの事例のみで，フィルタリングしたツイート集合から選択した事例とフィルタリングなしのツイート集

表 11　実験 3-1 のために抽出したツイート数
Table 11　Number of extracted tweets for experiment 3-1.

フィルタ	ツイート数
FILTER1	714
FILTER2	620
FILTER3	420

表 12　フィルタによるモデル精度低下の影響評価
Table 12　Impact assessment of filters.

フィルタ	評価指標	PSSA	従来手法
FILTER1	Accuracy	0.7027	0.6862
	Precision	0.3469	0.3054
	Recall	0.6911	0.7398
	F-measure	0.4620	0.4323
FILTER2	Accuracy	0.6547	0.6682
	Precision	0.3082	0.3250
	Recall	0.6992	0.7398
	F-measure	0.4279	0.4516
FILTER3	Accuracy	0.5450	0.6772
	Precision	0.2568	0.3284
	Recall	0.7724	0.7154
	F-measure	0.3854	0.4501

合から選択した事例の比較となる．フィルタリングによる教師データ内の事例の偏りの影響により，PSSA は従来手法より精度が低くなることが予想されるが，各フィルタでの精度の低下度合いから，その悪影響を評価する．

最初に PSSA の教師データを以下のように作成した．まず，baseline データに対して，PSSA の 3 つのプレフィルタをそれぞれかけ，プレフィルタによって抽出されたデータに対してそれぞれアンダサンプリングを行い，均衡なデータをそれぞれ作成した．**表 11** に抽出したデータ数を示す．

次に従来手法（フィルタリングなし）の教師データを作成するため，baseline データから表 11 とそれぞれ同じデータ数になり，かつ各クラスのデータ数が同じになるように，ランダムサンプリングしてきた．

上記で作成した教師データから意見抽出モデルを学習し，精度評価を行った結果を**表 12** に示す．

4.4.2　実験 3-2

実験 3-1 では，プレフィルタの悪影響のみを評価するため，アンダサンプリングによる均衡化を行ったうえでモデルを学習したが，実験 3-2 では，教師データ作成段階から同条件で PSSA と従来手法（フィルタリングなし）を比較する．つまり，教師データの数は同じでも，各クラスのデータ割合は異なる．これにより，プレフィルタによる悪影響だけでなく，不均衡緩和によるモデル精度への効果も含めて評価を行うことができる．

表 13　実験 3-2 のために抽出したツイート数と割合
Table 13　Number and ratioof extracted tweets for experiment 3-2.

フィルタ	ツイート数	PSSA		従来手法	
		割合	意見数	割合	意見数
FILTER1	1444	0.2472	357	0.1766	255
FILTER2	1116	0.2778	310	0.1783	199
FILTER3	688	0.3052	210	0.1628	112

表 14　フィルタごとの精度評価
Table 14　Accuracy evaluation for each filter.

フィルタ	評価指標	PSSA	従来手法
FILTER1	Accuracy	0.8303	0.8243
	Precision	1.0000	1.0000
	Recall	0.0813	0.0488
	F-measure	0.1504	0.0930
FILTER2	Accuracy	0.8303	0.8228
	Precision	0.9167	1.0000
	Recall	0.0894	0.0407
	F-measure	0.1630	0.0781
FILTER3	Accuracy	0.8273	0.8198
	Precision	0.8333	1.0000
	Recall	0.0813	0.0244
	F-measure	0.1481	0.0476

最初に PSSA の教師データ作成のため，baseline データに対して，3 つのプレフィルタをそれぞれかけ，データを抽出した．次に従来手法の教師データ作成のため，baseline データから同じデータ数をそれぞれランダムサンプリングした．実験 3-1 のように各クラスのデータ数を同じにするような処理は実施していないため，それぞれ意見クラスのデータ割合は異なる．**表 13** に抽出したデータ数と割合を示す．

上記で作成した教師データからそれぞれ意見抽出モデルを学習し，精度評価を行った結果を**表 14** に示す．

4.4.3　実験 3-1 と実験 3-2 に対する結果と考察

まず，PSSA のプレフィルタによる悪影響，教師データの事例の偏りによるモデルの精度低下を評価する．実験 3-1 の結果を見ると，$FILTER_k$ は k が増えるほど，絞込条件が細かいフィルタであるため，モデルの精度低下が予想されるが，表 12 の結果は，そのとおりになった．しかしながら，$FILTER_1$ の段階では，悪影響はなく，従来手法より PSSA の方が高い評価となっている．$FILTER_2$ では，*F-measure* が 0.03 ほど下がり，$FILTER_3$ では，0.07 下がっている．

しかしながら，実験 3-2 における，不均衡緩和によるモデル精度への効果も含めての評価を見ると（表 14），PSSA

がすべてのフィルタにおいて高い評価を得ており，PSSA の適用によるモデル精度向上効果が確認された．

5. おわりに

本論文では従来のデータ収集，アノテーション手法では不均衡データ学習になる，Twitter からの意見抽出タスクにおいて，不均衡化を緩和する手法 PSSA を提案し，その有効性を検証した．PSSA はサンプリングを行う前に，段階的に複数のプレフィルタリングを行うことによって教師データの中の事例の過度な偏りを防ぎ，不均衡化を緩和することができ，作成した教師データを学習させたモデルの精度も向上することを確認した．

具体的には，PSSA を用いると教師データに含まれる「意見」ラベルデータ数の割合が増加することが確認され，教師データの不均衡化が緩和されたことから，学習時の精度も向上した．また，フィルタリングの代わりにクラスタリングベースのサンプリングを行った場合よりも，モデル精度向上への効果が高いことを確認した．また，フィルタリングによる事例の偏りの影響についても分析し，その影響は不均衡緩和による精度向上のより小さい事を確認した．

評価に *Accuracy*，*Precision*，*Recall* のいずれかを用いることによって最良のモデルは変わるが，*Accuracy* や *Precision* を評価指標として利用すると，PSSA 単独のモデルが最良な結果が出ることが検証された．網羅性を高めるために *Recall* を評価指標とすると，PSSA と Under Sampling を併用したモデルが最良の結果が出ることが検証された．

今後の課題としては，より効果的なプレフィルタの設計と評価，他のドメインへの拡張手法の研究などがあげられる．

謝辞　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] NTT データ経営研究所：入手先 ⟨https://www.nttdata-strategy.com/aboutus/newsrelease/130805/⟩（参照 2019-11-15）

[2] 新井範子：ソーシャルメディアを活用したマーケティングに関する研究，専修大学博士論文 (2012).

[3] Pak, A. and Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining, *LREc*, Vol.10, No.2010 (2010).

[4] Peng, Y. and Yao, J.: AdaOUBoost: Adaptive over-sampling and under-sampling to boost the concept learning in large scale imbalanced data sets, *Proc. International Conference on Multimedia Information Retrieval* (2010).

[5] Naganjaneyulu, S. and Kuppa, M.R.: A novel framework for class imbalance learning using intelligent under-sampling, *Progress in Artificial Intelligence*, Vol.2, No.1 pp.73–84 (2013).

[6] Elkan, C.: The foundations of cost-sensitive learning, *International Joint Conference on Artificial Intelligence*, Vol.17, No.1, Lawrence Erlbaum Associates Ltd. (2001).

[7] Zhou, Z.-H. and Liu, X.-Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem, *IEEE Trans. Knowledge and Data Engineering*, Vol.18, No.1, pp.63–77 (2005).

[8] Weiss, G.M. and Provost, F.: The effect of class distribution on classifier learning: An empirical study (2001).

[9] He, H. and Garcia, E.A.: Learning from imbalanced data, *IEEE Trans. Knowledge and Data Engineering*, Vol.21, No.9, pp.1263–1284 (2009).

[10] Haixiang, G. et al.: Learning from class-imbalanced data: Review of methods and applications, *Expert Systems with Applications*, 73, pp.220–239 (2017).

[11] Yen, S.-J. and Lee, Y.-S.: Cluster-based under-sampling approaches for imbalanced data distributions, *Expert Systems with Applications*, Vol.36, No.3, pp.5718–5727 (2009).

[12] Ku-Mahamud, K.R., Zorkeflee, M. and Mohamed Din, A.: Fuzzy distance-based undersampling technique for imbalanced flood data, pp.509–513 (2016).

[13] 紺野友彦，藤井秀明，岩爪道昭：深層学習抽出特徴量から生成した擬似特徴量を用いた不均衡データ多クラス画像分類，人工知能学会全国大会論文集第 32 回全国大会，一般社団法人人工知能学会 (2018).

[14] 澤崎夏希ほか：量的不均衡データに対する学習精度改善のための文書かさ増し手法，ARG WI2, No.11 (2017).

[15] 乾　孝司，奥村　学：テキストを対象とした評価情報の分析に関する研究動向，自然言語処理，Vol.13, No.3, pp.201–241 (2006).

[16] 小林のぞみほか：意見抽出のための評価表現の収集，自然言語処理，Vol.12, No.3, pp.203–222 (2005).

[17] 佐野優太，峯　恒憲：地域苦情データにおける現状，要望表現の識別，人工知能学会論文誌，Vol.32, No.5, AG16-B_1 (2017).

[18] Yu, H. and Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, *Proc. 2003 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics (2003).

[19] Pang, B. and Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *Proc. 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics (2004).

[20] 立石健二，石黒義英，福島俊一：インターネットからの評判情報検索，情報処理学会研究報告自然言語処理（NL），Vol.2001-NL-144, pp.75–82 (2001).

[21] 橋本和幸ほか：センチメント分析とトピック抽出によるマイクロブログからの評判傾向抽出，電子情報通信学会論文誌 D，Vol.94, No.11, pp.1762–1772 (2011).

[22] 川島崇秀，佐藤哲司，神門典子：半教師あり学習を用いた要望ツイートの抽出手法の評価，マルチメディア，分散協調とモバイルシンポジウム 2016 論文集 2016，pp.38–43 (2016).

[23] 峠　泰成：ドメイン特徴語の自動取得による Web 掲示板からの意見文抽出，言語処理学会第 11 回年次大会，2005 (2005).

[24] 筒井貴士ほか：地方議会会議録コーパスの構築および政治情報システム構築を目標としたアノテーションの一提案，自然言語処理，Vol.21, No.2, pp.125–155 (2014).

[25] 宮崎林太郎，森　辰則：注釈事例参照を用いた複数注釈者による評判情報コーパスの作成，自然言語処理，Vol.17, No.5, pp.5_3–5_50 (2010).

[26] Paperno, D. et al.: The LAMBADA dataset: Word prediction requiring a broad discourse context, arXiv

preprint arXiv:1606.06031 (2016).
[27] 梶村俊介ほか：列挙型クラウドソーシングタスクのための品質管理法，人工知能学会論文誌，K-F79 (2016).

野崎 雄太

2020 年明治大学総合数理学部ネットワークデザイン学科卒業．同年筑波大学大学院理工情報生命学術院博士前期課程，現在に至る．機械学習，自然言語処理の研究に従事．

櫻井 義尚　（正会員）

明治大学総合数理学部ネットワークデザイン学科准教授．2000 年電気通信大学電気通信学部電子情報学科卒業．2002 年同大学大学院博士前期課程修了．2005 年同博士後期課程単位取得済み退学．同年博士（工学）．2005 年 4 月東京電機大学情報環境学部情報環境学科助手．2010 年 4 月同大学助教．2013 年 4 月より現職．機械学習，進化計算の理論及びマーケティング，自然言語処理の応用に関する研究に従事．IEEE，情報処理学会，人工知能学会，日本マーケティング学会，日本知能情報ファジィ学会，電気学会，日本オペレーションズ・リサーチ学会，各正会員．