Vol.2020-MPS-127 No.9
2020/3/2

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]

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)を提案し，その有効性を検証した．具体的には，Twitter からの意見抽出を課題として，まず Tweet データからの教師データセット作成を行い，辞書フィルタを用いた PSSA によるデータの不均衡化の緩和効果を検証した．次に，構築したデータセットを用いて，機械学習モデルの構築と精度評価を行い，従来からの不均衡データに対する手法と比較することで，データセット作成まで含めた機械学習モデル構築手法の比較評価を行い，その有効性を検証した．最後に PSSA に用いるフィルタリングによる Twitter からのサンプリングデータへの影響を分析し，提案手法の有用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 Tweet

YUTA NOZAKI[1]　YOSHITAKA SAKURAI[1]

Abstract: In this paper, we propose an annotation method that relieves imbalance by using mechanical pre-filtering 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

近年企業が Twitter をマーケティングに利用する「ソーシャルリスニング」が広く行われている．しかし，膨大なツイート集合から人手で意見を抽出するのは困難であり，また教師あり学習手法を用いた意見抽出システムも Twitter の特性上，アノテーションにおいて教師データが不均衡になり，抽出精度が落ちるという課題がある．従来手法ではこの課題に対し，アンダーサンプリングやオーバーサンプリングなど教師データレベルで対策を行っているが，本研究ではツイート集合に段階的にフィルタリングをかけることによって，不均衡化が緩和された教師データを作成する手法を提案する．

1 明治大学　総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164-8525, Japan

2. 関連研究

2.1 不均衡データの対策に関する研究

不均衡な教師データを学習させる研究は多く議論されている．紺野ら[1]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．一方，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ーサンプリングの手法は難しいとされていたが，澤崎ら[2]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(Prefilter based Stepwise Sampling for Annotation)はデータに対する不均衡データ対策のアプ

ⓒ2020 Information Processing Society of Japan

ローチであるが，アンダーサンプリングやオーバー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 は既存手法であるアンダーサンプリングや Cost Sensitive Learning と組み合わせて利用することができる．

2.2 意見抽出に関する研究

Twitter のツイートなどは「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[3]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上で，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．立石ら[3]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，文脈上等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．

3. 提案手法

本章では，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手法，PSSA 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．

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が少なくなってしまう場合に，サンプリングする条件を絞り込むことで，少数クラスのデータ数が増えるような機械的フィルタを利用することで，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，アンダーサンプリング等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殆どが学習されず，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全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, 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．PSSA の処理の手順を以下に示す．

(1) ブロック数(n)，教師データ全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($sampling_1$~$sampling_n$)の組を予め定義する．また，$k=1$とする．

(2) 全てのデータ($DATA$)に対し，①で定義した$FILTER_k$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) 作成した全ての教師データ($\sum TRAIN\_DATA_k$)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全てに適用可能である．

ⓒ2020 Information Processing Society of Japan

3.2.1 PSSA を用いた教師データの構築

辞書は小林が構築した評価表現辞書[a]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 5200 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．

PSSA のパラメータは以下のように定義する．

$$n = 3$$
$$ratio = 0.5$$

- 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 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．

$FILTER_k$の k が n に達するまで，よりプレフィルタを厳しくし，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．

3.2.2 学習

前項で作成した教師データで学習を行い，分類モデルを構築する．意見抽出研究，文書分類研究で比較的多く用いられている SVM を用いる．

4. 検証

PSSA が不均衡化を緩和する検証と PSSA を利用した意見抽出モデルの検証を行った．また，従来手法の構築を行い，比較を行った．

4.1 実験条件

4.1.1 データ収集・前処理

最初にツイートを収集した． ツイート収集で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，または略称とした．キーワードは表 1 に示した．また，5 つ星ホテルの内訳は「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」である．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．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 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．

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(2019 年 12 月 5 日更新時点)を利用した．また，収集したツイートデータを「ディズニー」，「USJ，ユニバ」，「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ドメインどうしのツイート件数を揃えることによって特定のキーワードにツイート件数が偏ることを防いだ．

4.2 教師データの不均衡化緩和効果の検証

PSSA によって作成した教師データ(以降，PSSA 教師データとする)の不均衡化が緩和されたかを検証する．

4.2.1 実験

PSSA による教師データの作成を行った．各段階でサンプリングした 1 ドメインあたりのツイート件数を表 1 に示す．また，

表 1　PSSA の各段階でサンプリングしたツイート件数
Table 1 Number of tweets sampled at each stage of PSSA

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

各ドメインの全収集ツイートから，PSSA を適用してドメ

a http://www.syncha.org/evaluative_expressions.html

ⓒ2020 Information Processing Society of Japan

インごとに表 1 で示したツイート数のラベリング済みデータを作成した．

また，アノテーターは 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 種類へ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ーで異なった場合，該当アノテーター間で定義や事例を改めて確認し，判断の統一を行った．

次に，従来手法として収集したツイートからのランダムサンプリングによる教師データ作成を行った．

PSSA の各段階におけるプレフィルタによって抽出されたツイート数が全ツイート数に占める割合(表 2)の分布に基づいて，PSSA で作成した教師データから尤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 3 に示す．

表 2　PSSA の各段階におけるプレフィルタによって抽出されたツイート数が全ツイート数に占める割合
Table 2 Percentage of extracted tweets

フィルタ	割合
FILTER1	0.6427
FILTER2	0.4549
FILTER3	0.1911

表 3　従来手法によって作成された教師データ件数
Table 3 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

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 4 に示す．

4.2.2 結果と考察

表 6 から PSSA を用いて作成した教師データが従来手法で作成した教師データよりも意見の数が多いことがわかった．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割合が少ないが，第 2 段階，第 3 段階では PSSA 側の割合が上回ったことがわかった．第 1 段階で PSSA 側の割合が少なくなった原因として第 1 段階では$FILTER_1$が表 4 の割合から半分以上のツイートがプレフィルタで抽出されているため厳しくなく，また$sampling_1$も 7：3 に設定した

表 4　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 4 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

ことからプレフィルタが強く作用しなかったことが挙げられる．

$FILTER_2$がより厳しくなったことから，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，$FILTER_1$をより厳し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 等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．

4.3 意見抽出モデルの精度検証

PSSA を用いることによって意見抽出モデルの精度が向上するかを検証した．

4.3.1 実験

最初に 4.2.1 で作成した 6664 件のデータから表 4 の割合の分布に基づいて，従来手法によってサンプリングされたデータを擬似的に 666 件作成し，これをテストデータとした．

次に 6664 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 5 に 3 種類のデータ数を示す．

表 5　各教師データ数
Table 5 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

最初にそれぞれの教師データに対して以下の前処理を行った．

・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

ⓒ2020 Information Processing Society of Japan

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，本研究では各ツイートを TF-IDF の手法を用いて特徴量化を行った．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 の TruncatedSVD を用いて 1000 次元へ次元圧縮を行った．

その後，結合したテストデータをそれぞれ分離させ，SVM を用いて学習を行った．ハイパーパラメータは scikit-learn のデフォルトの設定(Normal)で行った．

次に，3 つの教師データに対して多数派データを少数派データと同数になるようにサンプリングを行うアンダーサンプリングの手法を適用した上で同様に学習を行った(Under Sampling).

最後に，学習時に少数派データに対して，多数派データと同等になるように重みを加える Cost Sensitive Learning の手法を適用して学習を行った(Cost Sensitive Learning).

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，アンダーサンプリングは 5 回行った平均の値をとった．以下の表 6 に結果を示す.

表 6 各モデルの精度評価
Table 6 Accuracy evaluation of each model

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	7.7
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

4.3.2 結果と考察

実験によって 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．

Normal において PSSA-Base と baseline の結果を比較すると，全ての評価指標で PSSA-Base が baseline と同等かそれ以上の結果となった．アンダーサンプリングを行った場合と Cost Sensitive Learning を行った場合は時間以外の評価指標で baseline を上回った．

また，PSSA-Large はデータ数が多いため時間は非常に長くなるものの，Normal の Precision 以外と Under Sampling の Recall 以外においては非常に高い結果を得られた．

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．

Normal においては PSSA-Base は不均衡化が緩和されているため，baseline よりも Accuracy と Recall が高い数値を出したと考えられる．

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．

4.4 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，PSSA で作成した教師データがプレフィルタに偏り，プレフィルタでは抽出されなかったが文脈上意見にアノテーションでラベリングされたデータの学習が行われず，抽出が容易なデータに過学習するため，精度が悪化することを検証する．

4.4.1 実験 1

プレフィルタの効果である不均衡化の緩和を意図的に除去するために，アンダーサンプリングを行うことによって均衡化を行った上で，プレフィルタの偏りを検証した．

最初に 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る．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ーサンプリングを行い，均衡なデータをそれぞれ作成する．表 7 に抽出したデータ数を示す．

表 7 抽出したツイート数
Table 7 Number of extracted tweets

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

ⓒ2020 Information Processing Society of Japan

同様に baseline データから表 7 と同じデータ数をそれぞれアンダーサンプリングし，それぞれ学習を行った．表 8 に結果を示す．

表 8 フィルタの影響評価
Table 8 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

4.4.2 実験 2

次に均衡化を行わずにプレフィルタの偏りを検証する．最初に 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．また，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，それぞれ学習を行った. 以下の表 9, 表 10 に抽出したデータ数，学習結果を示す．

表 9 抽出したツイート数と割合
Table 9 Number and percentage of extracted tweets

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 10 フィルタ毎の精度評価
Table 10 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

4.4.3 結果と考察

実験 1 において，$FILTER_k$の k が増えると，PSSA が従来手法よりも精度が悪化した．$FILTER_1$と$FILTER_3$の Recall 以外では従来手法よりも全ての評価指標で精度が下がった．

一方で実験 2 においては$FILTER_1$の全ての評価指標，$FILTER_2$，$FILTER_3$の Precision 以外の評価指標で PSSA の精度が高かった．

$FILTER_1$においては Recall 以外の評価指標で PSSA の精度が高かったが，これは表 4 の$FILTER_1$によって抽出されたツイート数が全ツイート数に占める割合が半分を超えており, プレフィルタとしての性能が低かったと考えられる．$FILTER_k$の k が増えると，プレフィルタが厳しくなり，文脈上「意見」とされるツイートを正しく学習できないデメリットがより作用するが，同時に不均衡化の緩和のメリットが作用し，全体的には精度が高まることが確認された．また, $sampling_k$を調節することによって不均衡化は解消され，プレフィルタの偏りも調節することができる．

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，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，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プレフィルタへ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精度も向上した．

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．また，プレフィルタへの偏りによって生じる精度低下はプレフィルタの不均衡化が緩和される作用によって十分補えることが検証された．

今後の課題として，4 段階目のプレフィルタを追加することによってさらに不均衡化を解消する PSSA を利用した意見抽出モデルの構築が挙げられる．

参考文献

[1] 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 (2018). 一般社団法人 人工知能学会, 2018.
[2] 澤崎夏希, et al. "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)
[3] 立石健二, 石黒義英, and 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) 2001.69 (2001-NL-144) (2001): 75-82.

ⓒ2020 Information Processing Society of Japan