政治的立場の異なるツイッターユーザ群の特徴分析

大南勝　　掛谷英紀
s1211093@u.tsukuba.ac.jp　kake@iit.tsukuba.ac.jp
筑波大学

概要　これまで、自然言語処理技術を用いて、文書を政治的イデオロギー別に分類する研究が行われている。しかし、これらの先行研究で機械学習に用いてきた文書は、政治に関係のあるもの、あるいは政治に関係のある人物によるものに限っており、異なる政治的意見を持つ集団が政治的話題以外でどのような特徴を有しているかは十分分析されていない。そこで本研究では、米国 Twitter 社が運営するウェブ上のマイクロブログサービスである Twitter で投稿される「ツイート」と呼ばれる発言を教師信号とすることを考える。本研究では、キーワードを使ってある政治的話題に言及しているユーザを集め、それらのユーザが普段どのようなツイートをしているかを分析する。また、ネット上での暴力的な発言に着目し、暴力的なツイートが見られるユーザ層と、彼らの政治的イデオロギーの間の関連性について検証を行う。

1　はじめに

これまで、自然言語処理技術を用いて、文書を政治的イデオロギー別に分類する研究が多数行われている。畑中らは、新聞社説や国会議事録に着目した文書分類を試みている[1,2]。また、橋本らと東らは、Yahoo!JAPAN が運営する政治評価サイト「みんなの政治」内において、現在はサービスを終了している「みんなの議員評価」と呼ばれる議員評価記事を用いた政治的文書の分類および類似度マップの生成を行っている[3,4]。しかし、これらの手法は政治に関係のある文書のみを学習データに用いるため、政治に関係のない場面での各ユーザの主義主張を関連付けることは難しい。

この問題を解決するため、政治に関係ない情報も混在している言語資源として、米国 Twitter 社が提供するマイクロブログサービスである Twitter [5] (以下、ツイッターと表記)を用いることが考えられる。ツイッターを学習指標として用いるメリットは 2 点ある。1 つは、ツイッターはブログに比べて発言が容易なため、率直な思いが表明されている可能性が高いという点である。もう 1 つは、ユーザの特定が可能な点である。これまで、東らにより、政治家のツイート（ツイッター上の発言）を収集、分類する研究は行われている[6,7]。しかしながら、政治家以外の一般ユーザの政治的意見に注目した研究は行われていない。

そこで、本研究では、一般ユーザを対象に、政治以外の話題のツイートを含めて学習することで、ユーザがある政治的事柄に対して肯定的か否定的かを予測する機械学習を行い、政治的意見が政治以外の関心と何らかの相関を有するか否かを検討する。また、学習の正当性を、学習の結果得られたパラメータおよびクロスバリデーションの正答率で評価する。

さらに、ユーザの発言の暴力性についても分析する。今日、ネット上の誹謗中傷は現代社会において大きな問題となっているが、ツイッターも例外ではない。平和運動家や原発反対を主張する政治家に、「死ね」を含むツイートが見られるように、ツイッター上では普段の主張と相反する人格が表出することがよくある。そこで「死ね」のような暴力的なツイートを投稿するユーザ層と、彼らの政治的イデオロギーの間の関連性について検証を行う。

本論文の構成は次のとおりである。まず、2 節で用いたシステムについて説明し、3 節で実験結果、4 節で暴

力的なツイートの検証を行い、5 節でまとめを行う。

2　システムの概要

本研究では、形態素解析ツールとして、MeCab を用いる[8]。まず MeCab を用いて、収集した文書データを形態素解析し、品詞毎に単語を分割して素性として抽出し、それらから学習データ及びテストデータを作成する。なお、今回の機械学習では素性を名詞と動詞に限った。

学習データを元に、機械学習のプログラムで文書の特徴を学習し、テストデータを使って、開発したシステムの精度を算出する。機械学習には単純ベイズ分類を用いる。システムの概要を図 1 に示す。

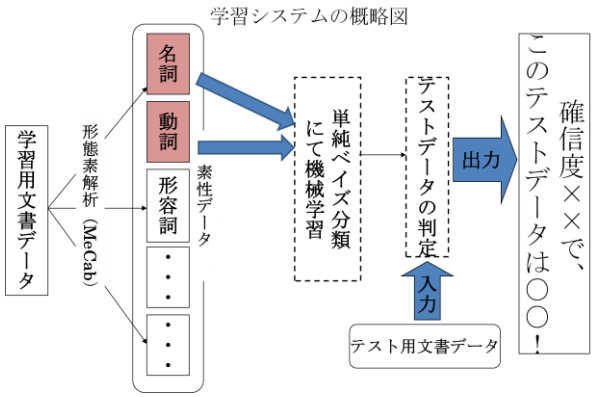

図 1　システムの概要

取得するツイートには URL および、数字も多く含まれる。これらは主義主張に全く関係ないが URL や数字の使用頻度は極めて高いこともあり、判定結果に大きな悪影響を与えると考えられる。そこで、本研究では学習データから URL 及び数字を含む素性はすべて排除している。

3　ツイートの学習

3.1　ツイートの分析

本研究ではツイッターの API を操作してツイートを収集し、学習データに用いる。以下、政治的事柄として安保法制問題を例に説明する。まず「安保法制」をキーワードにツイート内容および、ユーザ名を取得する。この分析対象であるテキストデータにはそのユーザ自身の発言だけでなく、非公式(引用)リツイートの内容を含む。しかし、公式リツイートについては、そのユーザの主義主張が得られない可能性が高いと考えられるので、学習の対象から除外した。

ここで、非公式リツイートと公式リツイートの違いは次のようなものである。非公式リツイートは、ツイッターでは非公式の機能だがユーザに広く利用されている機能である。これは、他人のツイートをコピーし、それに対する自分のコメントを入れて投稿するというものである。これに対して、公式リツイートは Twitter に公式に実装されている機能であり、リツイートすることで他人のツイート原文そのままに自分のタイムライン上に表示することができる機能である。

3.2　安保法制賛成派・反対派の分析

ツイッターでは過去のツイートを取得することが可能である。そこで、3.1 節で述べた方法で取得したユーザに対して、最大過去 100 件のツイートを新たに取得する。ここでも 3.1 節同様に非公式リツイートを含み、公式リツイートは除外する。これらのデータに、各ユーザが安保法案に対して肯定的か否定的であるかのタグ付けを行う。このタグ付けはツイートをもとに明らかに判断できるものだけを抽出し手動で行った。最終的にタグ付けされたユーザ数は安保賛成派が 43、安保反対派が 149 となった。

学習する際に、カテゴリーによって発言の件数に差が大きいと、判定結果が件数の多いカテゴリーに近づいてしまう。そこで本研究の実験では少数派である賛成派に合わせ、各カテゴリー43 ユーザずつ、計 86 ユーザのツイートデータを学習データに採用し、学習・実験を行う。

「賛成派」「反対派」それぞれ 43 ユーザのツイートデータを学習し、10 分割のクロスバリデーションにて判定精度の検証を行ったところ、正解率は 75.7%となった。

ここで、判定システムがどういった素性を手がかりにしてユーザを分類しているかを見るため、賛成派・反対派それぞれで判定に大きな影響を与えた上位の素性を

いくつかピックアップする。(表 1)

表 1　判定に影響を与えた素性の一部抜粋

安保法制賛成派	安保法制反対派
幼稚	再選
半島	強行
在日	賛同
悲惨	広がる
ウイグル	伊勢丹
武装	時事通信
帰国	地球
哀れ	吉永
執行	ヤジ
岡田	原子力

クロスバリデーションによる正解率から、ある程度各カテゴリーの特徴を把握できるといえる。また表 1 の賛成派上位素性には「半島」「在日」といった政治的に右寄りの人が嫌うものが挙がっている。一方、反対派上位素性には「強行」(安保法案の強行採決)、「原子力」といった政治的に左寄りの人が嫌うものが挙がっている。これらから右・左の思想を反映した妥当な分類結果であると考えられる。同時にこれらはどちらも自身の立場に反する相手への批判である。1 節でツイッターでは率直な意見が表明される可能性が高いと述べたが、本研究では率直な意見が相手の批判となって表れる結果が得られていると解釈できる。一方、政治以外について特徴的な素性は、政治以外の話題のツイートを学習対象にしても、上位には現れなかった。これは、今回の学習においては、ある政治的な意見が、政治以外の話題への関心と高い相関をもつといった事実は見出せていないことを意味する。

3.3　辺野古移設賛成派・反対派の分析

次に、沖縄基地問題について、辺野古移設賛成派・反対派それぞれ 60 ユーザのツイートデータを学習し、10 分割のクロスバリデーションにて判定精度の検証を行ったところ、正解率は 64.1%となった。ここでも上位の素性をいくつかピックアップする。(表 2)

表 2　判定に影響を与えた素性の一部抜粋

辺野古移設賛成派	辺野古移設反対派
動物	打倒
敵国	核心
アイドル	図る
理想	調整
大卒	称する
軍艦	資本
タモリ	検察
アニメ	アベノミクス
珊瑚	ファシズム
母国	狂っ

正解率は 3.3 節の安保法制問題の場合と比較すると、約 10 ポイント減少する結果となった。反対派上位素性については、3.3 節同様に、政治的に左寄りの意見を反映する素性が多く見られる。一方賛成派上位素性では、「アイドル」「タモリ」のように政治に関係のない素性も見られている。このことは、辺野古移設反対派は政治運動に積極的な人が多数を占める一方、辺野古移設賛成派には普段政治的な話題にあまり関心のない一般ユーザも多く含まれていることを示唆する。このような分析結果は、1 節で述べた政治的意見と政治以外の関心との関係を見出すという目的に一部適うものである。ただし、クロスバリデーションの正解率と、ジャンルに偏りのない素性の獲得にはトレードオフの関係があると考えられる。この問題を乗り越えることは、この研究における今後の課題である。

4　暴力的なツイートの検証

1 節でも述べたようにツイッター上の暴言等の誹謗中傷は社会問題となっている。そこで暴力的なユーザとイ

デオロギーの関連性についても検証を行った。本研究では「・・しろ」「・・するな」の他、動詞の命令形を多用するユーザを暴力的なユーザと定義している。安保法制問題、及び沖縄基地問題に関して、賛成派・反対派それぞれで命令形を含むツイートの有無により、細かく分類を行った。結果をそれぞれ表 3 と表 4 に示す。

表 3　命令形ツイートの有無によるユーザ振り分け
(安保法制問題)

	安保賛成	安保反対	計
命令形ツイートあり	28	63	91
命令形ツイートなし	15	86	101
計	43	149	192

表 4　命令形ツイートの有無によるユーザ振り分け
(沖縄基地問題)

	移設賛成	移設反対	計
命令形ツイートあり	28	50	78
命令形ツイートなし	34	41	75
計	62	91	153

ここで賛成・反対の主張は命令形ツイートの有無に関連性があるかの検証を行う。これらの要素間に有意差があるかの検証にはカイ二乗検定を用いる。その結果、安保法制問題に関しては $p = 0.0082$、沖縄基地問題に関しては　$p = 0.23$ となった。本研究では 1 回でも命令形を含むツイートが見られれば、命令形「あり」に振り分けている。しかしその条件では、偶然や冗談で使われた可能性を排除しきれていないため、今後より詳細な分析を行う必要がある。

5　まとめ

本研究では、政治的に意見が対立するユーザ群に対して、政治に関係ない情報も混在している言語資源を教師信号として用いて文章を分類することを試みた。本稿ではマイクロブログサービスであるツイッターのツイートと呼ばれる発言を教師信号として学習し、学習結果の有効性を確認した。

またツイッター上では普段の主張と相反する人格が表出することがあることを踏まえ、命令形を含むツイートに着目した検証も試みた。一部トピックでは命令形を含むツイートとの有無と、そのトピックに対する主張に関連があることが確かめられた。

今後の予定としては、より多くのトピックで本研究と同様の検証を行うことを考えている。また、学習対象のデータサイズを大きくすることと、効率の良いデータ収集方法の確立も今後の課題である。

参考文献

[1] 畑中、村田、掛谷：新聞社説・国会議事録に基づく言論のイデオロギー別分類, 言語処理学会第 15 回年次大会, 2009

[2] 畑中、掛谷：国会議事録に基づく言論の政治思考分類, 第 5 回メディア情報検証学術研究会, 2009

[3] 橋本、掛谷：Web 上のレビュー記事のイデオロギー分類とその応用, 言語処理学会第 16 回年次大会, 2010

[4]　東、橋本、掛谷 : Web 上の言語資源に基づく国会議員の分類, 言語処理学会第 17 回年次大会, 2011

[5] Twitter　https://twitter.com/

[6] 東、掛谷：自己組織化マップによる国会議員のツイッター分類, 第 6 回メディア情報検証学術研究会, 2010

[7] 東、掛谷 : 国会議員のツイッター分類とその応用, 言語処理学会第 18 回年次大会, 2012

[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.sourceforge.net/