社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

マイクロブログ上の流言に対するユーザの態度の分類

藤川　智英†　鍛治　伸裕††　吉永　直樹††　喜連川　優††

†† 東京大学 生産技術研究所　〒 153-8505 東京都目黒区駒場 4-6-1
† 東京大学 大学院 情報理工学系研究科 電子情報学専攻　〒 153-8505 東京都目黒区駒場 4-6-1
E-mail: †{fujikawa,kaji,ynaga,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　近年 Twitter などの影響で様々な情報とともに流言もまた広がりやすくなっており，これらをいち早く検出することが重要と考えられる．しかし，流言かどうかの判断はコンピュータにはもちろん人間にも難しい．本研究では，情報に対するユーザの反応を分類することで，情報の真偽の判断を支援することを目指す．具体的には，疑っている人がどの程度いるか，流言と言える根拠を示している発言はどれかを知るために反応を信疑と根拠の有無で分類することを考える．本稿ではこのような分類を行うための SVM 分類器を構築した．

キーワード　テキスト解析，情報信頼性，データマイニング，Twitter

Classification of users' attitudes toward rumors on microblogs

Tomohide FUJIKAWA†, Nobuhiro KAJI††, Naoki YOSHINAGA††, and Masaru KITSUREGAWA††

†† IIS. The University of Tokyo　4-6-1 Komaba, Meguro-ku, Tokyo 153-8505 JAPAN
† Graduate School of Information Science and Technology with a major in Information and Communication Engineering, The University of Tokyo　4-6-1 Komaba, Meguro-ku, Tokyo 153-8505 JAPAN
E-mail: †{fujikawa,kaji,ynaga,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　In these days, information spread on microblogs (e.g., Twitter) includes tons of (false) rumors. This motivates us to quickly detect such rumors to let people know their unreliability. In this study, we analyze users' attitudes toward such rumors to help people understand the rumors. We classify the responses to the information according to two criteria: 1) doubting or believing the information 2) with or without grounds. We trained two SVM classifiers to solve these two text classification task, using the gold-standard data that is manually built from Twitter.

Key words　Text Analysis, Information Credibility, Data Mining, Twitter

1. はじめに

Twitter などのマイクロブログの登場は情報の共有を手軽にし，現在ネット上では様々な情報が非常に早く広がるようになっている．しかし，それらの中には根拠のない風説（**流言**(注1)）も含まれており，特に災害がなどの非常時には爆発的に流言が広がることが多い．例えば東日本大震災では以下のような流言が少なくとも 80 件広がった [1].

- ホウ酸を食べると放射線を防げる
- 外国人犯罪多発
- 東電の社員が逃走した

こうした流言は人間の安全や名誉に関わるものも多いため，迅速に発見し，サイトなどを通じてその存在を明らかにすることが重要と考えられる．

しかし，Web 情報をユーザがどう信じるかに関する研究 [2] では，正しい情報というのは必ずしも客観的に定義できないため，最終的な判断を行うのはユーザであり，正しさをシステムによって完全に客観的に決定することはできないと指摘している．

そこで，本研究では情報が正しいかどうかを直接判定するのではなく，情報に対する反応を「疑いの有無」と「根拠の有無」に分類してユーザに提示することで，ユーザが情報の真偽を判断するための支援を行うことを目的とする．情報に対するユー

(注1)：流言という用語は従来，真疑不確かな情報一般に対して用いられうるが，本稿では文脈上誤解のない場合，誤った情報を参照する意味で用いる．

ザの反応を分類することには，次のような利点がある．まず，どの程度の割合の人がその情報を疑っているかを知ることができる．もし通常より多くの人が疑っている情報なら，それは誤情報である可能性が示唆されるため，情報の真偽を判断するための材料の一つとなりうる．また，流言に対する反応の中には，それが誤りまたは真実であると判断する根拠を示している発言も多いため，そうした発言を集めることが出来れば，ユーザに情報の真偽を判断するための判断材料を提供することができる．

本研究では代表的なマイクロブログである Twitter 上の発言を対象とする．実際に流言を検出するシステムを構築する場合，その概略は以下のようなものを想定している．

（1） Twitter からツイート（投稿）を収集する
（2） どのような話題が流れているかを判別する
（3） その話題に対する反応を収集する
（4） 反応を分類する
（5） 分類結果を元に，流言の可能性がある情報を判別する

このうち，1-3 については，Twitter が提供する API を利用するなどして対応するものとし，本稿では 4 について詳しく述べる．

本稿の構成は以下のとおりである．まず，第二章で関連研究について紹介する．第三章では提案手法について述べ，第四章では評価データの作成方法を述べる．第五章では実験内容とその結果について述べ，第六章でまとめを行う．

2. 関 連 研 究

マイクロブログに限らず，ウェブには，真偽不明の情報が多く含まれることから，情報の真偽を直接的に判断する研究 [3] や間接的に真偽の判断材料をユーザに提示する研究 [2] がこれまで行われている．

Castillo ら [3] は Twitter 上の誤情報を対象として，情報の真偽を直接的に判断する手法を提案している．彼等は既存のテキスト分類で利用される手がかりに加え，分類対象の情報を拡散しているユーザの属性や情報の拡散の仕方などを手がかりとして用いることで情報の真偽を判断している．

一方，情報の正しさは必ずしも客観的に定義できないことから，正しさを直接判断するのではなく，正しさを判断するための材料を提示することを目指す研究も存在する．山本 [2] は「エジソンは電気を発明した」のようなファクト型知識に対してそれを支持するページがどの程度存在するかなどを提示することでユーザの判断を手助けする手法を提案している．また，情報分析システム WISDOM (注2)では，「電気自動車は環境に良い」「裁判員制度の是非」など賛否の分かれる主張を入力すると，それに対する賛成意見，反対意見がどの程度存在するか，また，具体的にどのような人がどのような意見を述べているかなどを一覧表示することでユーザが主張の是非を判断する手助けを行うことを目指している．

本研究は後者の研究と同様の立場に立ち，ユーザーに情報の真偽を判断するための材料を提示するという立場を取り，ある流言に対してどのような意見が存在するかを提示することを目指す．

(注2)：http://ici.wisdom-nict.jp/

表 1 前 処 理
Table 1 Preprocessing

Original string	Replaced string
Html tags such as <a href=“ … ”>	(Delete)
http://...	URL
@username	USERNAME
RT USERNANME	RTorQT

図 1 Twitter における情報伝播の例
Fig. 1 Example of information spread on Twitter

3. 提案手法

3.1 問題設定

本研究では，流言に対する Twitter 上での反応を，情報に対するユーザの態度（信疑），その根拠の有無という二つの観点で分類するテキスト分類問題を考える．以下で，各タスクの具体的な内容について順に説明する．

まず，情報に対するユーザの態度（信疑）の分類基準について説明する．一口に疑いと言っても完全に疑ってるものから半信半疑，「本当ならひどい話です！」のように一応疑っているがかなり信用よりのものまで幅広い．しかし，ここでは，わずかでも疑っているものは疑いに分類する．なぜなら，細かく分類するとモデルが複雑になり，ラベル付の際にも判断に個人差が出てくるためである．また，根拠のはっきりした情報なら「本当なら」のように留保をつけることもないと思われるため，真偽不明の情報特有のツイートであると考えられるからである．

例えば有名人が死んだという流言が流れたとき，「○○（有名人）が死んだって本当？」「デマだろ」のような発言を”疑い”，「○○（有名人）が交通事故でなくなったそうです．ショック……」のように，まったく疑っておらず，普通のニュースと同様に受け取っている場合は”信用”と分類する．

次に，各発言におけるユーザの態度（信疑）に対する根拠の有無の判断基準について述べる．根拠としては，具体的な URL やツイートの引用など，その根拠の内容が明白なものに限らず，発言の文面からユーザが確信を持って判断する材料を持つと考えられる場合に根拠ありとする．例えば「○○が死んだって話，ジョークサイトの嘘記事ですよ.」という発言では，単にそれが誤りであると指摘するだけでなく，なぜ誤りと言えるかの根拠も示しているため，”根拠あり”と分類する．

以下で Twitter 上で，情報がユーザの反応と共にどのよう伝搬するかを簡単に説明する．Twitter では図 1 のように，ユーザを指し示すのに「@username」という書式が使われる．ま

た，あるツイートに対してそれを引用して発言する非公式 RT という書式が存在し，この場合,「{投稿者自身の発言} RT @username: {引用する発言}」という形になる．Twitter ではこのような形でひとつの情報が連鎖的に広がっていくことが多い．

非公式 RT の場合，”RT”以前の部分に意見が書かれているため，最後の RT 以前の部分のみを分類に使用する．図 1 の 2 番目，3 番目ともに,「RT @XX: Twitter が情報統制」の部分は使用しない．ここで，最後の RT 以前の部分を使用するのは，3 番目の発言ように疑っている意見を信じている場合でも，最初部分（「そうなんだ」）だけを取り出して「信用している」と誤って判断しないためである．

3.2 分類手法

我々は，テキスト分類問題で広く使われる分類器である SVM を，教師付きデータから学習し，各分類タスクに適用する [5]. この際，文脈と関係ない特定文字列は表 1 のように置換する．

流言に対するツイートを観察した結果，おおよそ次のような特徴が見られた．

- ユーザが情報を信じている場合は「ひどい」「感動した」などの感情的表現が含まれることが多い
- 流言であると指摘しているツイートは「流言」「ガセ」など直接的な言葉を使っているツイートが多い
- 根拠を示しているツイートは，何らかの事実を提示するため,「○○は××のはず」「××じゃない」など特定の言い回しのパターンが好んで用いられる

これらの特徴を考慮して，出現する単語や文体を素性とすれば分類が可能と考えられる．実際に用いた素性の詳細は次のとおりである．

- 単語の 1,2,3-gram
- 品詞（第一階層と第二階層）の n-gram
  - 例：“ 東京都が株主 ”
  - ∗ 第一階層： 名詞-助詞-名詞
  - ∗ 第二階層： 固有名詞-格助詞-名詞
- 文章長（$n \leqq 1$ となるように正規化）

ここで，品詞 (POS) の n-gram を使用しているのは，品詞の連なりは文体を表すためである [6]．根拠を提示している文では，次のように話題によって提示される根拠も異なり，当然のことながら含まれる単語も異なる．

- 株主は東京都です．都知事じゃありません．
- 現在ではザイールという国は存在しません．

そのため，共通する要素である文体を捉えることが重要と考えられる．

例えば感情的な文なら,「ひどい！」「すごい！」→「形容詞-記号」,「許せない」→「動詞-助動詞」のようになるし，事実を提示する文では,「○○は××」→「名詞-助詞-名詞」という形になる．

ここで，より長い品詞の連なりを使用すればそれだけ文体をとらえやすくなるが，その場合素性の数が爆発的に増えてしまう．そこで，長さを 3 までに固定した固定長品詞 n-gram を使用する場合と、長さは固定しないが、Mukherjee ら [6] が提案するアルゴリズムを使用して、出現するドキュメントの割合が閾値 (minsup) 以上の可変長品詞 n-gram だけを使用する場合で実験を行った．

3.3 素性の重み付け

単語，品詞の素性の重みとして，tf 及び tfidf を検討する．これらは以下の式で表される．

$$\mathrm{tfidf} = \mathrm{tf} \cdot \mathrm{idf} \tag{1}$$

$$\mathrm{tf_i} = \text{単語 } i \text{ の出現回数} \tag{2}$$

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : d \ni i\}|} \tag{3}$$

$|D|$ は総ツイート数，$|\{d : d \ni i\}|$ は単語 i を含むツイート $d \in D$ の数である．

idf は，その単語が出現するツイートの数が小さいほど大きくなる関数形になっている．そのため，特定のツイートにしか登場しない単語の重要度を上げる一般語フィルタとしての役割を持つ．例えば新聞記事のカテゴリ分類では「は」「が」のようなどんな文章にも出現するような単語を使用してもあまり意味はなく,「株価」「粉飾決算」など特定のカテゴリ (ここでは経済) にしか出現しない単語のほうが分類にあたって有用であるため，これが用いられる．

4. 信疑／根拠タグ付きデータの作成

学習･評価用データの作成には Togetter (http://togetter.com/) というツイートをまとめたサイトを用いる．このサイトでは誰でも議論や様々な話題に関するツイートをまとめることができ，毎日 200 件以上のまとめが作られている．また，各まとめには何に関するものなのか分かりやすいようにタグ付けが行われている (図 2).

今回の実験では，様々なページの中で,「デマ」というタグが付いたページを使用する．多様なサンプルを得るため，流言に対する 10 人以上の反応をまとめたページを用い，手作業でラベルづけを行った．

なお，この時,「マジかよ」のような極端に短く，信じている場合にも半信半疑にも使用される，人間でも判断が難しい反応に関してはラベル付けを行わなかった．

対象とした特定の流言に関するまとめページは 35 で，ラベル付けを行ったのは 1986 ツイート，内 761 が「疑い」で，410 が「根拠あり」だった．

5. 実験結果

ラベルづけしたデータから分類器を作成し，その精度を調べた．

5.1 実験設定

実験ではラベルづけしたデータのうち,「東京電力株主 5 位は石原氏」というトピックに関する流言をテストデータ，それ以外を訓練データとして用いた (表 2, 3).

図 2　Togetter
Fig. 2　Togetter

表 2　信疑の分類のツイート数
Table 2　Number of tweets about doubt or belief

	Doubt	Belief	All
Training data	743	1210	1953
Test data	211	120	331

表 3　根拠の有無の分類のツイート数
Table 3　Number of tweets about with or without doubtion

	With grounds	Without grounds	All
Training data	340	1613	1953
Test data	145	186	331

ツイートの形態素解析には MeCab (注3)，辞書は ipadic を使用した．また，SVM 分類器の実装には liblinear (注4) を用いた．

SVM を用いて分類器の学習を行う場合，ソフトマージンパラメタ C の値が分類精度に大きく影響することが知られている．そこで，C を $2^{-s}(s = -4, -2, 0, ..., 14, 16)$ の範囲で動かし，**陽性** (疑い，根拠あり) に対する F-measure という精度が最も高くなる C を選んだ．また，可変長品詞 n-gram における閾値 minsup に関しても，0.05 から 0.30 まで 0.05 刻みで動かし，陽性に対する F-measure が最も高くなる minsup を選んだ．

比較のため，全ての発言を疑いまたは根拠ありと分類した場合を Baseline とする．

5.2　分類精度

信疑の分類と根拠の有無の分類の二タスクについて，単語素性のみを用いた場合と，固定長の品詞を素性に加えた場合の比較を行った．さらに重み付けを tf とした場合と tf-idf 値とした場合の比較を行った．

分類結果の評価には以下の 4 つの指標を用いた．

Precision　陽性と判定されたツイートのうち，実際に陽性だったツイートの割合

Recall　テストデータに含まれる陽性の全ツイートのうち，正しく判定されたツイートの割合

Accuracy　全体のうち正しく判定できたツイートの割合

(注3)：http://mecab.sourceforge.net/
(注4)：http://www.csie.ntu.edu.tw/ cjlin/liblinear/

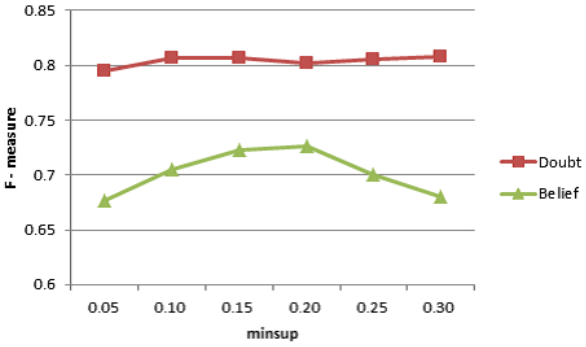

図 3　minsup と F-measure の関係
Fig. 3　Relation of minsup and F-mesure

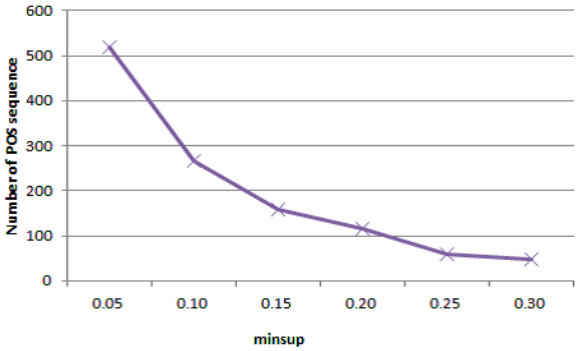

図 4　minsup と 素性数 の関係
Fig. 4　Relation of minsup and number of pos sequence

F-measure　Precision と Recall の調和平均

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

表 4, 5 に分類実験の結果を示す．文体を表す品詞の n-gram を使用した場合，信疑分類の精度に大きな変化は見られない (表 4) が，根拠の有無の分類精度は大きく向上している (表 5)．これは前述したように，根拠を提示している文は共通の単語はあまりないが文体は共通しているためと考えられる．

しかし，tfidf に関しては，信疑の分類でもあまり変化がなく (表 4)，根拠の有無の分類では精度がかえって下がってしまった (表 5)．これは，前述したように同じカテゴリであっても使われている単語に多様性があり，出現割合の少ない単語だからといって重要なわけではないためと考えられる．

次に，品詞の可変長 n-gram を用いた場合に minsup のみを変化させたときの結果を示す．図 3 は minsup と F-measure の関係を示す．また，minsup と素性数の関係を 図 4 に示す．これらのグラフから，minsup の値を適切に選ぶことによって，少ない素性数で精度の向上が実現されていることが確認できた．

図 5,6 は C と各指標の関係を示すが，C が一定以上になると指標にほとんど変化は見られなくなり，また，最大値との差も小さいため，C を変化させて指標との関係を調べることができない場合でも，C を一定以上の大きさにすれば最適値に近い結果が出ると予想できる．

5.3　議　　論

表 6 と表 7 に，素性の上位 10 件を示す．これらの表から次

表 4　信疑の分類結果

Table 4　Result of doubt or belief

	Baseline	tf			tf-idf		
		Only words n-gram	＋ Fixed length POS n-gram	＋Flexible length POS n-gram	Only words n-gram	＋ Fixed length POS n-gram	＋Flexible length POS n-gram
Accuracy	0.630	0.749	0.785	0.758	0.779	0.779	0.789
Precision	0.630	0.830	0.872	0.839	0.848	0.852	0.841
Recall	1 .00	0.763	0.777	0.767	0.796	0.791	0.825
F-measure	0.773	0.795	0.822	0.801	0.822	0.820	0.833

表 5　根拠の有無の分類結果

Table 5　Result of with or without grounds

	Baseline	tf			tf-idf		
		Only words n-gram	＋ Fixed length POS n-gram	＋Flexible length POS n-gram	Only words n-gram	＋ Fixed length POS n-gram	＋Flexible length POS n-gram
Accuracy	0.438	0.704	0.749	0.785	0.692	0.737	0.722
Precision	0.438	0.758	0.804	0.825	0.795	0.845	0.811
Recall	1.000	0.476	0.566	0.648	0.400	0.490	0.476
F-measure	0.609	0.585	0.664	0.726	0.532	0.620	0.600

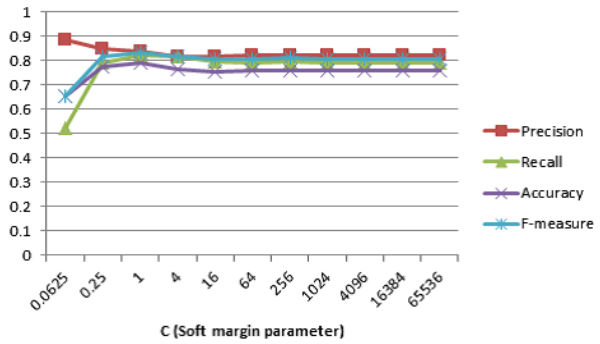

図 5　C と信疑の分類の精度の関係

Fig. 5　Relation of C and index of Classification according to doubt or belief

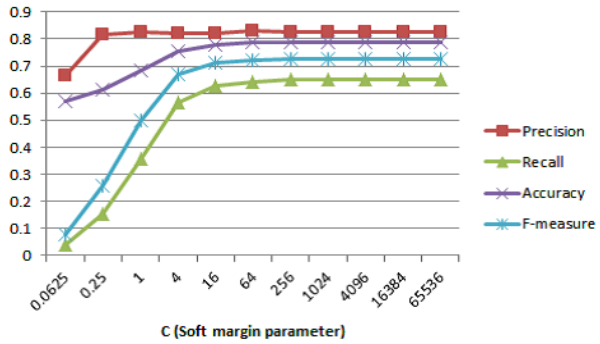

図 6　C と根拠の有無の分類の精度の関係

Fig. 6　Relation of C and index of Classification according to with or without grounds

のことがわかる．まず，信疑分類に関する素性では，「デマ」「本当」「？」など直感的にも関係があると思われる素性が並んでいる．

しかし，根拠の有無の方に関しては，重複した素性や「コンゴ」のように特定のトピックとしか関連しない単語に大きな重

表 6　疑いの有無について重みの大きい素性の一覧

Table 6　Features with beg weights of classification according to doubt or belief

	Doubt		Belief	
1	デマ	7.08	！	-2.68
2	ソース	3.96	中国	-2.61
3	本当	3.00	統制	-2.59
4	という	2.74	…	-2.57
5	確認	2.64	ね	-2.35
6	ガセ	2.50	新宿	-2.18
7	嘘	2.35	情報統制	-2.15
8	記事	2.06	事	-2.13
9	.	2.01	歴史	-2.10
10	ほんとに	1.99	日本	-2.04

みが与えられてしまっている．これは，根拠を示す文の場合，トピックによって出現する単語がまったく違うこと．さらに，あるトピックについて根拠を示す側はある単語を多用するが，そうでない側はほとんど使わない場合，その単語を根拠を示す際に使われる単語と分類器が誤認識してしまうためだと思われる．また，大きな重みを与えられた品詞 n-gram についても，その妥当性は直感的に判断が難しい．これに関しては今後さらに調査を進めていく予定である．

表 8, 9 に判定されたラベルと実際のラベルの数の関係を示した Confusion Matrix を示す．表 9 を見ると，根拠の有無の分類については，偽陰性が誤りの大きな要因となっていることが分かる．そのため，今後は，根拠が提示されている文を正しく認識するための新しい素性の導入などに努めていきたい．

図 7, 8 に，SVM を用いて分類を行ったときのマージンに対するスコア超平面からの距離とテストデータの分布を示す．

信疑タグ付きデータ，根拠タグ付きデータともに，陽性のほうがピークが右よりになっており，スコアの大きさが実際のラベルと連動していることがわかる．

表 7　根拠の有無について重みの大きい素性の一覧

Table 7　Features with beg weights of classification according to with or without grounds

	With grounds		Without grounds	
1	/	4.27	:	-3.23
2	.	3.49	を	-3.17
3	係助詞_サ変接続名詞	3.47	一般名詞_一般名詞_一般名詞	-3.00
4	形容詞_助動詞_名詞	3.46	0	-2.96
5	コンゴ	3.28	!	-2.96
6	代名詞名詞_一般名詞	3.15	:トルコ	-2.75
7	ザイール	3.11	:トルコが	-2.75
8	「	2.92	名詞_名詞_記号	-2.71
9	固有名詞_固有名詞_固有名詞_格	2.91	USERNAME:トルコ	-2.68
10	助詞_自立動詞	2.89	う	-2.65

表 8　疑いの有無の分類の Confusion Matrix

Table 8　Confusion Matrix of classification according to doubt or belief

	Belief	Doubt
Classified to belief	96	47
Classified to doubt	24	164

表 9　根拠の有無の分類の Confusion Matrix

Table 9　Confusion Matrix of classification according to with or without grounds

	Without grounds	With grounds
Classified to without grounds	166	63
Classified to with grounds	20	82

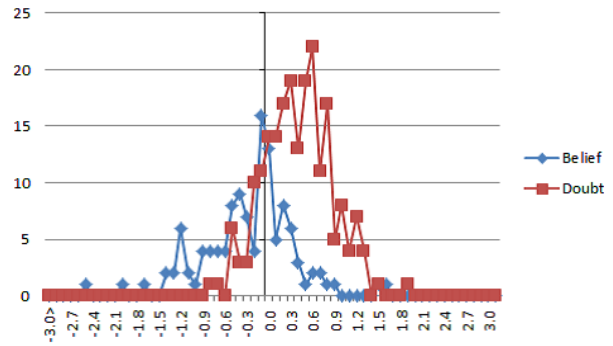

図 7　マージンとテストデータの分布 (信疑タグ付きデータ)

Fig. 7　Distribution of margin and test data(annotated according to doubt or belief)

6. ま と め

本研究では，流言に対するマイクロブログ上のユーザの態度を分類するという問題を提案し，それを機械学習の手法を用いて解く方法を示した．実験では，Twitter 上での流言に対する態度の分類を行い，提案手法の有効性を示した．

今後の課題としては，他に分類に寄与する素性を新たに探すことが挙げられる．また，情報の受け手の態度を自動的分類，集約することによって，ウェブ上の流言を自動的に検出できるようになる可能性があるとも考えており，これについても今後研究を行っていきたい．

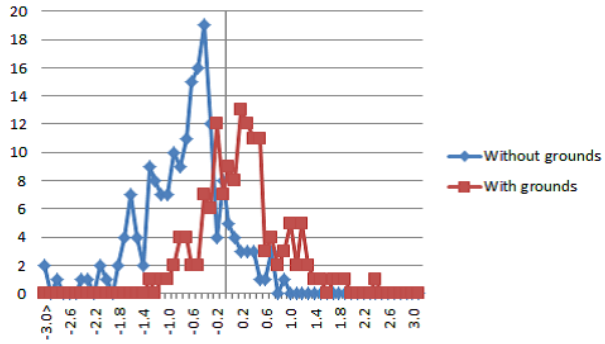

図 8　マージンとテストデータの分布 (根拠の有無タグ付きデータ)

Fig. 8　Distribution of margin and test data(annotated according to with or without grounds)

文　献

[1] “震災後のデマ 80 件を分類整理して見えてきたパニック時の社会心理,” 2011. http://news.livedoor.com/article/detail/5477882/.

[2] 山本祐輔，“ウェブ情報の信憑性分析に関する研究,” PhD thesis, Kyoto University，2011.

[3] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on twitter,” WWW 2011, pp.675–684, Hyderabad, India., April 2011.

[4] “情報信頼性判断支援システム”．http://ici.wisdom-nict.jp/.

[5] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[6] A. Mukherjee and B. Liu, “Improving gender classification of blog authors,” Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp.207–217, MIT, Massachusetts, USA, 9-11., Oct. 2010.