修　　士　　論　　文

マイクロブログに現れる社会的影響力を持つ情報カスケードの分析及び検知に関する研究

Analysis and Detection of Information Cascades with Social Influence from Microblogs

指導教員　　豊田 正史准教授

東京大学 大学院情報理工学系研究科
電子情報学専攻

氏　名　　48-146414 川本 貴史

提　出　日　　平成28年2月4日

概 要

マイクロブログではユーザ間での情報共有が連鎖することによる情報カスケードがしばしば観測され，その中には商品の回収につながり得る風評の拡散や災害への対処方法の共有など社会的影響力を持つものも存在する．このような情報カスケードに迅速に対応するために，本稿では教師あり学習に基づく分類器を用いて Twitter における情報カスケードから社会的影響力のある情報カスケードを早期に検知するというタスクを新たに提案し，これを機械学習に基づく分類器により解く手法を提案する．我々はまず，複数人によるアノテーションを行い問題の妥当性の検証を行った．その後，本文に含まれるテキスト情報や情報カスケードのグラフ構造の特徴量，情報カスケード毎のユーザ分布といった社会的影響力の有無の早期検知に有効であると考えられる多様な特徴量を設計し，カスケードが大規模に成長するかどうかの分類を行った後，社会的影響力の有無を分類する二段階の分類手順を踏む方法とそれを同時に分類する方法の二種類の手法を提案し比較を行った．実験では，実際のツイートデータセットにおいて検知する対象の情報カスケードのサイズ（初期リツイート数）を変化させて社会的影響力の有無を分類する実験を行い，どれだけ早期に検知が可能であるかを明らかにし，提案した特徴量の有効性を評価した．

謝辞

はじめに指導教官である豊田正史准教授に感謝いたします．豊田先生には，研究方針から発表練習に至るまで，ギリギリになってから準備を始める私に厳しくも優しく相談に乗っていただきました．また言語処理学会に参加させていただいて，様々な外部研究室との交流の場への参加ができたこと，Twitter の実データを利用した研究ができたことはこの喜連川・豊田研究室という恵まれた研究室に所属できたおかげだと感じています．このような環境を提供してくださった豊田正史准教授，喜連川優教授に心より感謝いたします．

次に吉永直樹特任准教授に感謝いたします．共著に入っていただいてからは吉永先生の力に頼りっぱなしになってしまいました．特に原稿の締め切り前では WebDB の時も TOD の時も日付が変わる締め切り直前まで手伝っていただきました．車まで出していただき，本当にありがとうございました．また，NLP グループの輪読会にも参加させていただき，様々な NLP の技術を学ばさせていただきました．プログラミング演習など自分のコードの汚さ，遅さを実感させられる非常に有意義な機会だったと感じています．また，青春 18 切符の旅行などに連れて行っていただき，様々な「安くて美味しいもの」を教えていただきました．特にビールは色々な種類を飲まさせていただいたと思います．卒業してからも研究室の飲み会やビアバーなどでご一緒させていただきたいです．

次に，伊藤正彦特任准教授，横山大作特任助教，と鍛治伸裕特任准教授 (現・Yahoo! JAPAN 研究所 上席研究員) に感謝します．伊藤先生，横山先生には専門分野が少し異なるのにも関わらず週一回のミーティングの際に鋭い質問をいただいたり，自分の全然知らない新しい知識をいただき，研究に活かすことができました．また，サーバーや会議室等の研究室の備品の管理の面でも様々なご迷惑をお掛けしてしま

いましたが，研究のバックアップをしていただきありがとうございました．また，鍛治先生には Yahoo! JAPAN 研究所に移られるまで週一回のミーティング以外にも NLP グループの輪読会でお世話になりました．勉強会のスライドなど何度も直していただき，ありがとうございました

次に，研究室の先輩の鈴木有さん，槇佑馬さん，伊東直弘さん，清水翔太さん，劉さん，長谷川大起さんに感謝いたします．先輩方に研究室での過ごし方や，特に NLP グループの方々には様々なツールの使い方から研究テーマなどまで教えていただきました．また，鈴木さんには特にお世話になりました．元の研究室の先輩でもあり，学部時代のクラスの先輩でもあり，就職先での先輩でもあるのでこれからもお世話になります．どうぞよろしくおねがいいたします．

次に同期の加藤千裕さん，小矢島諒君，谷川祐一君，石渡祥之佑君，金洪善さんに感謝いたします．同期の方々とはいつもお互いの研究内容について議論を行う合間にテニス，卓球，ソシャゲなど息抜きを一緒にする仲間でした．小矢島君の家には何度もお邪魔させていただきました．ありがとうございます．卒業してからもぜひお邪魔させてください．

加藤さん，小矢島君，石渡君とは特に修論提出前にはお互いに煽り，足を引っ張りあって楽しく修論を書くことができました．間違いなく自分一人では書き上げられなかったと思います．ありがとうございます．そしてごめんなさい．

後輩の岩成達哉君，小泉実加さん，佐藤翔悦君に感謝いたします．M1 の早い時期から研究を行っている優秀な後輩は刺激的でした．みんな多趣味で積極的だったので研究室内の雰囲気が明るくなって楽しい一年を過ごせました．ありがとう．

最後に，北海道から見守っていてくれた父と，いつもご飯を作ってくれた母，社会人で忙しい中影から支えてくれた妹に感謝の意を捧げます．

2016 年 2 月 4 日

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　マイクロブログにおける情報カスケードとその社会的影響力

近年，Twitter や Facebook といったマイクロブログが出現し，その上で友人関係をバーチャルに表すソーシャルネットワークが大規模化している．このソーシャルネットワークでは友人間でのコミュニケーションが行われるが，単なる友人間でのやりとりにとどまらず，友人から受け取った情報をさらに他の友人へと発信することが日常的に行われる．このような情報共有が連鎖することによって引き起こされる情報拡散を情報カスケードという [1]．ソーシャルネットワーク上での情報発信はどのようなユーザでも手軽に行うことができるため，ソーシャルネットワークには多種多様な大量の情報が存在するが，そのような情報の中でカスケードとなって広く拡散する情報はごくわずかであり，情報が拡散するかどうかを予測することは重要である．例えばユーザーはオンラインの雑多な情報から有用な情報のみを抽出することができれば，効率的に時間を使うことができ，企業は，効率的に広く拡散する情報を見つけることができれば広告の最適化などにも役に立つといえる．そのため，情報が拡散するかどうかを予測する研究近年盛んに行われている．

ここで「どのような情報が拡散するのか」ということに注目すると，近年では，商品を批判する意見が拡散し，商品の回収につながったり，事故や災害への対処方法がマイクロブログを通してひろがったりするなど情報が拡散することによって，実社会に影響を及ぼすような社会的影響力を持つ情報カスケードも観測されている．特にマイクロブログにおける情報カスケードには，マイクロブログのリアルタイム性の高さによって急速に広がる特徴があるため，政府，マスメディア，企業などに

とって社会的影響力の高い情報カスケードを早期に発見することは風評被害に対する未然の対処，世論動向，報道，商品に対するフィードバックとして重要であり，例えば，既にマスメディアであるテレビ局の NHK では人力でツイートの監視を行うことで報道題材を発見したり，番組内で Twitter で盛り上がった単語をまとめることで世間の関心のある出来事を可視化したりしている．

以上のような,「社会的影響力を持つ情報カスケード」は特に早期に対応，あるいは認知するべき情報である．しかし一方で，広く拡散した情報が全て社会的影響力を持つか，というとそうではなく，アフェリエイトリンクへと誘導するスパムやジョーク，有名人の日常のつぶやき，広告など社会的影響力の少ない情報も多い．

注目すべき情報カスケードを検知するための研究としては，将来的に広く拡散する可能性のあるカスケードを検知する手法 [2] や，スパムツイートを検知する手法 [3] があるが，前者は広く拡散する情報カスケードが必ずしも社会的影響力を持つとは限らない点で，また後者はスパム以外にも社会的影響力のない情報カスケードが存在するという点で，社会的影響力を持つ情報カスケードを検知する上では不十分である．

1.2 本論文の概要

そこで本研究では，マイクロブログの 1 つである Twitter を解析することによって，社会的影響力を持つ情報を早期発見することを目指す．

しかしながら，そもそも社会的影響力を持つ情報カスケードがどのようなものか自明ではなく，このような情報カスケードの早期検知に有効な特徴は明らかでないため，本研究では大きく分け (1) 社会的影響力の有無に基づいたアノテーションを行う (2) 社会的影響力を持つ情報カスケードの早期検知 の 2 つのタスクに取り組んだ．

本研究の貢献は以下の通りである．

- 今まで注目されていなかった情報カスケードが持つ社会的影響力に注目し，分析を行った．

- 社会的影響力を持つ情報カスケードの早期検知に有効であると考えられる特徴量を設計した.
- 作成したデータセットに基いて評価実験を行い，提案手法の有効性を示した.

1.3 本論文の構成

本論文の構成は以下の通りである.

第 2 章 マイクロブログにおける情報カスケードに関する先行研究についてまとめ，その中でも特に，マイクロブログにおける情報カスケードの拡散を予測する研究，マイクロブログにおける情報カスケードの分類を行う研究，マイクロブログにおけるスパム検知を行う研究についてまとめ，本論文が提案するタスクとの関連性，差異について述べる.

第 3 章 本研究が提案する社会的影響力について詳細を明らかにし，それに基いて行った情報カスケードへのアノテーションについて述べ，その結果に基いて行った分析の結果をまとめる.

第 4 章 社会的影響力を持つマイクロブログにおける情報カスケードの早期検知というタスクに対して有効であると考え，提案する特徴について詳細を述べる.

第 5 章 twitter データセットを用いて社会的影響力を持つ情報カスケードを早期検知する評価実験を行い，提案手法の有効性を確認する.

第 6 章 全体のまとめと今後の課題について述べる.

第2章　背景知識

本章では，本論文で研究対象としているマイクロブログにおける情報カスケードについて説明する．

2.1　マイクロブログの機能

本論文で扱うTwitterは，マイクロブログと呼ばれるWebサービスの一つである．マイクロブログにはTwitter以外にもFacebookやweiboなど様々な種類が存在するが，その主な機能としては，ユーザ間で友人関係を形成する機能と短いテキストや画像，動画などのメディアをリアルタイムにユーザが投稿する機能が挙げられ，これらを組み合わせることでマイクロブログ上でユーザはコミュニケーションを取ることができるようになる．

2.1.1　Twitterにおける友人関係

Twitterにおけるユーザ間の友人関係は，フォロー (Follow) という行為によって形成される．他のユーザをフォローすることでそのユーザの投稿を購読することができるようになるため，ユーザは自分が興味のあるユーザや，実世界で関わりのあるユーザをフォローする．このフォロー関係が様々なユーザ間に張り巡らされることで，Twitterにおける友人関係は構築されている．

2.1.2 Twitter におけるインタラクション

Twitter には，主にリプライとリツイートという 2 種類のユーザ間のインタラクションの方法が存在する．

メンション (Mt) @マークにユーザ ID を続けることでそのユーザに言及する投稿のことをメンション (Mt) という．メンションを受け取ったユーザは通常のタイムラインとは別に通知を受け取る．ため主にユーザ間のコミュニケーションに用いられる．また，特定のツイートに対してメンションを付けることが可能であり，ツイートに対して意見や感想を述べたり，逆にツイートに対する反応を見るために用いられる [4].

リツイート (RT) 他のユーザの投稿を自らの投稿として再投稿・拡散する機能のことをリツイート (retweet, RT) という．リツイートはユーザが興味をもった話題や意見を自身をフォローするユーザへ転送する目的や，投稿主に対する対話，意思表示として用いられる [5][6].

また，リツイートには Twitter が公式に用意した API を用いて行う公式リツイートの他に，ユーザが自身の意見などを書き加えて再投稿を行う非公式リツイートが存在する．

2.2 マイクロブログにおける情報カスケード

マイクロブログでは友人間での情報の共有が連鎖することで爆発的に情報が広まることがある．これを情報カスケードと呼ぶ.

特に Twitter では RT が連鎖することによる情報カスケードがしばしば観測される．Twitter における情報カスケードの例としては著名人のツイートが多数 RT されることによる情報カスケードから，一個人が目撃した竜巻を写真に撮ってツイートしたものが多数 RT されることによって起こる情報カスケードまで様々なものが存在する．

このようにマイクロブログにおいて情報カスケードが起こる主な要因として挙げられるのは速報性，リアルタイム性のある実世界に関する情報が発信されていることであり，この特徴により，マイクロブログから実世界で起こっている出来事を検出し，速報することを目指すマイクロブログをを Social sensor として扱う技術が研究されている．[7]

第3章　関連研究

本章ではマイクロブログからの社会的影響力を持つ情報カスケードの早期検知に関係する研究を紹介する．

本研究では社会的影響力を持つ情報カスケードの早期検知を行うが，我々の知る限り，社会的影響力を持つ情報カスケードの検知を対象とした研究は存在しない．しかしながら，注目すべき情報を早期に抽出するという観点で本研究の目的に近い研究としては，Web コンテンツの人気を予測する研究が行われており，その中でマイクロブログにおける情報カスケードは盛んに研究が行われている．また，興味のある情報を抽出するという観点で本研究に関連する研究として，情報カスケードやツイートを自動で分類する研究が行われている．

また，一方で情報カスケードに関する研究ではないが，社会的影響力という観点から関連する研究としてはマイクロブログにおけるスパム検出などが行われている．以下で，これらのタスクについて，我々が提案するタスクとの関連性を明らかにするとともに，情報カスケードの性質の分析に用いられている特徴量を紹介する．

3.1　カスケード予測に関する研究

カスケード予測に関する研究としては，カスケードの成長 [2] を予測する研究に限らず，実際につぶやくユーザを予測する研究 [8] や，ユーザの影響力を定量化する研究 [9] など幅広く存在する．また，対象とするカスケードもミームを Twitter のハッシュタグ [10] や URL[8][9] とする研究や，ツイートをクラスタリングした上で，それを情報カスケードとみなし，情報伝播の研究を行う研究 [11][12]，あるいは Facebook の投稿拡散機能であるシェアによる情報カスケードを扱う研究 [2] など多岐に渡る．

カスケードの成長を予測する研究 [13][14][15][2] にも様々な研究が存在し，例えばChengらは，時間に関する特徴量，ユーザに関する特徴量，構造に関する特徴量[16][17]，コンテンツに関する特徴など様々な特徴量を用いることで，社会の中で広く拡散される投稿を当てる問題を解いており，ある意味では社会的に影響力の高い投稿を当てているともいえる．しかしながら，1 節で述べたように，広く拡散される投稿だからといって，必ずしも社会的影響力を持つ投稿とは限らない．

また，どのような情報が早く拡散するのかを調査した研究 [18] も存在するが，早く拡散する情報が社会的影響力を持つかというと必ずしもそうではない．

3.2 カスケード分類に関する研究

マイクロブログにおけるカスケード分類の研究にはトピックによる意味的分類やグラフパターンによる構造的分類 [19] の他にツイートの信頼性判定を行うものなどがある．Sriram らはユーザが読むツイートを絞るための手助けとしてツイートを News, Opinions, Deals, Events, Private Messages の 5 つに分類する手法を提案している [20]．その際，分類の特徴量として Bag of words を用いている．また，Ren らは各ツイートに対し Web Forum や質問応答システムなどで一般的なラベル付けである複数の階層的な意味ラベル付けを行うことを提案している [21]．しかし，これらのタスクではツイートの持つ情報の社会的影響力については考慮されていない．

また，Castillo らはあるトピックのツイート集合において News クラス，Chat クラス，判断できない，のどのクラスに属するかを判定した後，自動でそのトピックの信頼性を判定している [22]．その際の特徴量としてはユーザの特徴量，トピックの特徴量，リツイートの特徴量を用いている．この研究においては，ツイートの信頼性判定に重点が置かれており，前段階の News クラス分類においても，特定の出来事に関するニュースかどうかという観点で判定が行われている．そのため，本研究で考える，情報が社会的に影響力を持つかどうかという判断基準とは異なる．

話題の早期検知を行う研究も存在する．このような研究は Topic Detection and Tracking と呼ばれ [23]，ユーザ毎のトピックや話題の時間変化をモデル化する研究

や [24]，近年ではSNS 上でのイベント検知を行う研究 [25] が行われている．これらの研究で対象とする話題やイベント自体の抽出は，情報カスケード自体の抽出という点で関連するが，どの研究も情報の社会的影響力は考慮していない．重要な話題の早期検知という観点ではSNS 上で長期間流行する話題の早期検出を行う研究 [26]，Twitter における政治的なトピックの発生を早期発見する研究 [27] なども存在するが，これらと本研究とは注目する情報カスケードの性質が異なるといえる．

3.3 マイクロブログにおけるスパム検出に関する研究

近年マイクロブログでは悪意のある投稿や，それを自動で行うスパムユーザが増加しており，これらを自動検出することが広く求められており，研究が行われている．Chen らはインタラクションの構造に着目し，クラスタリング係数や推移性等の指標が有効であるとしている [3]．Gao らは投稿に注目し，投稿毎にスパム判定をするシステムを提案している [28]．

しかし，1 章で述べた通り，社会的影響力の有無を判定するに際しては，カスケードがスパムでないと判定するだけでは不十分である．また，スパム検出は基本的にユーザや投稿に対して行っておりカスケードは対象としていない．

第4章　情報カスケードにおける社会的影響力

1章で述べた通り，社会的影響力という観点から情報カスケードを分類した研究は存在しないため，情報カスケードにどれだけ社会的影響力を持つ情報カスケードが存在するか，またどのような種類が存在するのかは明らかになっていない．以降，4.1節では，社会的影響力の有無をアノテーションした情報カスケードのデータセットの構築方法について述べる．続いて4.2節で，構築されたデータセットに対してどのような社会的影響力を持つ情報カスケードが存在するのか分析を行い，その結果について述べる．

4.1　情報カスケードへのアノテーション

アノテーション対象の情報カスケードとして，まず，Twitter APIによるツイートの収集を行い，リツイート数に基づいて情報カスケードの抽出を行う．このようにして得られた情報カスケードについて，被験者により社会的影響力の有無をアノテーションし，評価用のデータセットを得る．その後，社会的影響力を持つ情報カスケードの内容を分析する．

4.1.1　Twitterデータセット

情報カスケードの分析，評価を行うためのデータセットとしては，著者らの研究室において2011年3月より継続的に収集しているTwitterのデータセットを用いた．本データセットは，150万人程度の公開ユーザからタイムラインを継続的に収集した

表 4.1: 情報カスケードの統計量

	1月	2月
カスケード数 (50RT 以上)	31,479	16,817
カスケード数 (600RT 以上)	1,130	475
カスケードへの参加総ユーザ数	407,034	338,640

もので，2015 年 8 月時点で約 250 億のツイートが蓄積されている．収集対象のユーザは，2011 年 3 月に 30 名程度の著名な日本人ユーザを選択し，それらのユーザに対してメンション (Mt) やリツイート (RT) を行ったユーザをさらに収集対象として順次拡大していったものである．この中から，2012 年 1 月から 2013 年 2 月の間につぶやかれたツイートを用いて評価用データセットの構築を行った．

4.1.2 情報カスケードの抽出

本稿における情報カスケードは Twitter API による公式 RT によって拡散されたツイート (元ツイート) とその公式 RT の集合とする．分析対象となる情報カスケードは，次で述べるインタラクショングラフに含まれるユーザを観測対象のユーザセットと限定した上で 2013 年 1 月，2 月のツイートそれぞれから 50 回以上 RT が観測された日本語を含む元ツイートとその RT を抽出することによって作成した．また，非公式 RT が起点となって RT され拡散した情報カスケードは今回分析対象から外した．結果抽出された情報カスケードの統計量は表 4.1，図 4.1 の通りである．

また，社会的影響力の有無を判断する情報カスケードは 600 回以上 RT が観測されたものでフィルタし，さらに，情報カスケードの社会的影響力の有無を判断する際には（ツイート収集時に取得していなかった）リンクされている画像等も参照する必要があったため，リンク先を復元できなかった 29 の情報カスケードは分析対象から外した．

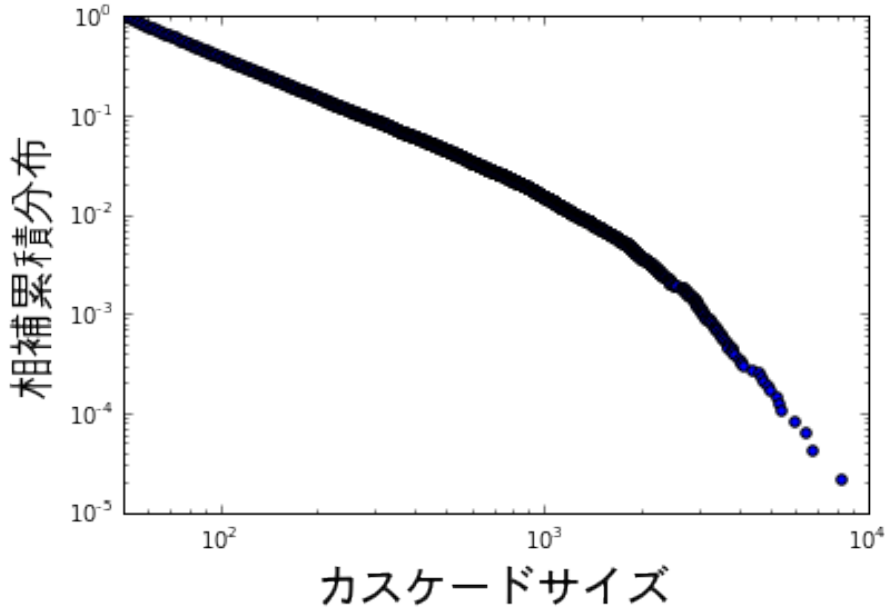

図 4.1: 1 月,2 月のカスケードサイズの分布

表 4.2: インタラクショングラフ

ユーザ数		1,066,870
エッジ数	Mt	58,627,341
	RT	114,848,093
	Mt or RT	153,711,945

インタラクショングラフの構築　情報カスケードを抽出する期間以前の 2012 年 1 月から 12 月のユーザ間の投稿のやり取りを元にユーザ間の関係を表す有向グラフ（インタラクショングラフ）G を作成し，情報カスケードの経路を推定する．なお，推定した経路は次節で述べる提案手法で分類の手がかりの一つとして用いる．

RT と Mt はどちらもユーザ間の情報のやり取りを表しており，このようにして得られるユーザ間のつながりはカスケードの情報伝播の主要な経路となると考えられる．そこで各ユーザをノードとして（過去に）情報が流れる方向と同方向となるよう，RT に関しては RT 元から RT したユーザへ情報が流れるため同方向のエッジ

表 4.3: 1月，2月毎のカスケードの社会的影響力の有無

	1月	2月
社会的影響力有	188	106
社会的影響力無	942	369
合計	1130	475

を，Mt に関しては Mt を送る際は送り先のユーザの投稿を見て送ったと考えられるため Mt の方向とは逆向きのエッジを追加しインタラクショングラフ G を得る．今回作成したインタラクショングラフの統計量を表 4.2 に示す．今回分析の対象としたユーザはこの期間に一度はインタラクション元となっているユーザであり，およそ 100 万ユーザ存在した．

4.1.3 被験者による社会的影響力の有無の注釈付け

4.1.2 節で得られた情報カスケードに対し，(著者を含まない) 3 人の被験者により，「ツイートに書かれた情報を知ったり，その情報を不特定多数に知られたりすることで，直接的あるいは間接的に行動や意思決定に影響を受けるか人がいるか」という観点で社会的影響力の有無を注釈付けした．なお，このアノテーションは実際に RT したユーザが読む/見ると考えられる，元ツイートの本文，画像，元ツイートに含まれる URL のリンク先の情報を基に行ってもらった．これによって得られたラベルにおける三人の被験者間一致度 [29] は 0.69 となり，[30] によれば相当な一致となり，社会的影響力を持つ情報に関してかなりの共通認識が得られていることが確認された．

最終的にラベルの不一致は多数決により解消した．その結果を表 4.3 に示す．表 4.3 から分かる通り，社会的影響力を持つ情報カスケードは全体のおよそ 20%弱と少ないことが確認された．

表 4.4: 意見の対象の事例

	明確	曖昧
個人	A 教授, B ちゃん, C 知事	渉外部 10 人のオヤジ，スポーツ記者，D 社社員
組織	E 社, F 新聞, G 番組	ある大手企業，テレビ，学校，朝の情報番組
業界	飲食業，マスコミ，ゲーム業界，日本の IT 企業	
社会	高校大学生，中高年，韓国，世論，素人，同性愛者	

4.2 社会的に影響力のある情報カスケードの分析

前節で抽出した社会的に影響力をもつ情報カスケードに対し，社会的影響力を持つカスケードにどのようなものが存在するかを明らかにするため，拡散された情報の内容について分類を行った．

分析の結果，社会的影響力のある情報カスケードは 1) 個人，組織，業界，社会への意見に対する共感・反感と 2) 影響力のある出来事 (事実) の周知，として大別されることが分かった．個人，組織，業界，社会への意見に対する共感・反感は，その意見に対し個人，組織，業界が対応する必要があるため社会的影響力を持ち，社会に対する意見では，そのカスケードが世論を反映していると考えられるため社会的影響力を持つ．一方，影響力のある出来事 (事実) の周知には，事件・事故の速報や知られていない問題の周知が含まれ，それぞれ，世論動向や注意喚起，啓蒙やデマ訂正，問題提起として重要である．

なお，以後引用する事例では，特定の人名や企業名などが含まれるため，ツイートの内容が特定の個人・企業の誹謗・中傷に繋がると考えられる場合は，著者の判断で＜人名＞や＜企業名＞のように匿名化を行った．

4.2.1 個人，組織，業界，社会への意見

個人，組織，業界，社会への意見に対する共感・反感に対しては，その意見の対象が何なのかを分類する．また対象が個人，組織のレベルである場合，対象の明確性によって情報カスケードに対する対応が変化すると考えられるため，対象が明確かどうかも分類した．まず表 4.4 にそれぞれの対象の事例を列挙する．次からそれぞれの対象の粒度毎に，元ツイートの例と社会的影響力があると判断される理由を分析する．

明確な個人 人名やスクリーンネームなど明確な個人に対して名指しで意見を述べている元ツイートが該当する．特に人数は一人とは限らず，グループで活動している場合はグループが対象にもなりうる．

> きょうのシンポジウムの質疑応答で「自身の発言をどう考えているのか」という質問が＜人名＞に対して出て、匿名質問には答えないとつっぱねた。質問者は名乗り、それでも＜人名＞は答えず。きいていて思わず泣いてしまった。

この例では，本文に人名が出現しており，明確な個人の粒度であると判断される．その上で批判をされているため，謝罪や釈明が必要になると判断されるため社会的影響力を持つ．

不明確な個人 意見の対象が個人であると判断できるものの，具体的な人物の特定まではできない元ツイート．

> ＜人名＞が授賞式に行くのにご本人の飛行機はエコノミー、ホテルは一泊 12000 円だったのに、必要もないのについてった英語 もしゃべれない渉外部 10 人のオヤジらの飛行機はビジネス、ホテルは一泊 5 万円だったんだと！(´ Д ‘) 言いふらしてやってくれと言われたので、言いふらすわ！

この例では，意見の対象は渉外部 10 人（不明確な個人）であるが，この行動

に対して対応するのはこれらの人物が所属する組織であると考えられるため，その規模で悪評が広まることが問題となる．

明確な組織 特定の企業や政府組織などが名指しで意見を述べられているもの．

> ＜URL＞ 2月3日の記事からご覧ください。 全国チェーン、＜社名＞の対応です。私は許すことは出来ませんし、今後、関わることは一生ないでしょう。 こんな会社が拡大することを望みませんし、全国のライダーにこの事実が届くことを切に願う。

この例では対象は＜社名＞であり，業務中の不手際を指摘されたものである．この情報が拡散し，社会的批判になると営業に支障がでる可能性があるため社会的影響力を持つ．

不明確な組織 明確な企業名などは明らかにされていないが，意見の対象が組織であるもの．

> 1985年の日本航空123便墜落事故で、事故からしばらくたった時、生存者の当時12歳の女の子がテレビの取材で「これから望むことは?」みたいなことを聞かれ、泣きながら「もうテレビが取材に来ないでほしい」と言ったのはいまだによく覚えています。

この例では，意見の対象はテレビ局であり，具体的な局名はわからないため区分としては不明確な組織である．この意見はテレビ局全体の報道業界として対応する必要があると考えられるため社会的影響力を持つ．

業界 同じ産業，商売などに携わる組織，個人の集合が意見，批判の対象となっているもの．およびそこで共有される価値観や体制が意見，批判の対象となっているもの．業界全体として対処する必要のある問題である場合，社会的影響力を持つといえる．

> ＜業界名 A ＞に転職したら人間として扱われるようになったのでカルチャーショックに驚いてる。仕事前に上司や先輩が怪我するなよって声かけてくる。やばい。熱でたら心配される。やばい。2 時間睡眠で動ける体になれって言われない。やばい。＜業界名 B ＞勤務時代と全然違う。やばい。

この例では，本文で業界の勤務体制が批判されているため，この区分に分類される．また，対象の業界全体として勤務体制の改善の必要があるため社会的影響力を持つ．

社会 国や年代などでくくられた人の集団や，その間で共有されている価値観や体制が意見，批判の対象となっているもの．およびそこで共有される価値観や体制が意見，批判の対象となっているもの．

> 「ゲームでなら人殺しができる！」という人と、「ゲームの中でも人殺しはちょっと……」という人では、明らかに後者の方が「ゲームと現実の区別がついていない」のだが、世間的には何故か前者こそが「ゲームと現実の区別がつかなくなって殺人を犯す！」と囃し立てられる。

この例では，ゲームによって犯罪が助長されるという風潮が批判されているため，社会の粒度に分類される．この問題は今後ゲーム規制などに発展する可能性があり，ゲーム業界にとって世論として重要である．

4.2.2 影響力のある出来事 (事実) の周知

(2) の影響力のある出来事 (事実) の周知に分類されるカスケードに対しては，周知している出来事の事件性，その周知している出来事がどのような人，集団に対して重要なのかをもとに，速報，注意喚起・デマ訂正，啓蒙のいずれかに分類された．

速報 周知している出来事の事件性が高いもの．また，社会全体に対して大きな影響を及ぼすと考えられる出来事の周知．

> 【北朝鮮核実験情報】本日１１時１９分頃、気象庁が北朝鮮を震源とする地震波を観測、自然地震ではない可能性があります。北朝鮮による核実験の可能性もあるので、官邸対策室を設置しました。今後も随時情報をお知らせします。

この例は北朝鮮が核実験をした可能性があるという事件性の高い出来事を周知しており，重要である．

> 湘南新宿ライン、６日午前の運転取りやめ http://t.co/CEV7gUL9

この例は翌日の鉄道の運行予測を周知するものだが，この運行予測によって影響が大きく出ると考えられるため社会的影響力を持つ．

注意喚起・デマ訂正 周知している出来事が特定のユーザに対しての警告や対処法であったり，大きな影響を与える誤った情報の訂正である元ツイート．

> 【重要】昨日辺りから流行ってる「日頃の行いおみくじ」、どうやら勝手にフォローを行ったりツイートをする権限も解放してしまうスパムらしい。使ってしまった人は承認アプリリストから速やかに削除した方がいい。削除ページは公式サイトのココ→＜ URL ＞

この例は「日頃の行いおみくじ」というスパムに対する対処法を拡散するもので，対処法を知らないユーザに対して広めることは重要であるため，社会的影響力を持つ．

> 本日、18 歳以下の方が＜サービス名＞を利用できなくなるというデマが出ていますが、そのような事実はありません。18 歳以下の方も引き続きご利用いただけますのでご安心下さい。公式情報はこのアカウントや公式ブログでご提供致します。＜ URL ＞ ＜ハッシュタグ＞

この例は＜サービス名＞の利用制限が始まるというデマを訂正する元ツイー

トであり，18 歳以下のユーザが当該サービスを使い続けるかどうかに影響を持つ．

啓蒙 事件性のあまりない出来事，問題であるが，社会にあまり知られていない事実の周知．社会に対して有益な情報が拡散される場合や，問題の周知自体が目的のものが存在する．

> タバコの煙は、いま話題の PM2.5 の一種に含まれ、喫茶店の喫煙席の PM2.5 濃度は、中国の大気とほぼ同等。友人が参加した産業医の講習会で開会早々にあった衝撃的な話。

この例は対してタバコの煙が PM2.5 の一種であるという社会にあまり知られていない事実を広める元ツイートであるため，啓蒙としての社会的影響力を持つ．

> 文章力を語る人は大抵まず語彙や知識に着目するけど、リズム感に関してはなかなか触れる人が少ない。語彙や知識が骨だとしたら肉となるのは間違いなくリズム感だと思う。リズムの生きた文章は滑るように読み進められる傍ら、リズムの死んだ文章は幾ら中身があっても疲れるだけで全く頭に入ってこない。

この例は文章力をつけるに当たって，リズム感という新しい観点からの評価の必要を訴えている．文章力の教育や，文章力をつけようと思っているユーザへ有用であり，社会的影響力を持つ．

3.3.2 節の分析を踏まえて著者が社会的影響力の種類を分類した結果，個人，組織，業界，社会への意見に対する共感・反感の個数の合計は 195，影響力のある出来事 (事実) の周知の個数の合計は 99，社会的影響力の無いカスケードの数の合計は 1311 となった．詳細な分類の結果は表 4.5, 4.6 に示す．また，意見の項目では組織や社会に対する意見が多いことが分かった．特に世代などへの意見や，特定の企業に対する批判が多く見られた．被験者による社会的影響力のアノテーションの有無の不

表 4.5: 個人，組織，業界，社会への意見に対する共感・反感の分布

	明確	不明確
個人	23	33
組織	44	18
業界	32	
社会	45	

表 4.6: 影響力のある出来事 (事実) の周知の分布

速報	36
注意喚起・デマ訂正	37
啓蒙	26

一致は啓蒙の項目に多く見られた．啓蒙では，社会的影響力の有無を判断する際に被験者自身の価値観に依存することになり，その部分のゆれが原因であると考えられる．

第5章　社会的影響力の有無に基づくマイクロブログにおける情報カスケードの分類手法

本章では，社会的影響力の有無による情報カスケードの分類に対して有効な手がかりとして働くと考えられる特徴量を提案し，それをまとめる．具体的には，社会的影響力の有無の判断に際しては，どのような内容の情報を，どのような過程で，誰が広めているかが手がかりとなると考えられる．そこで，これらをそれぞれカスケードした元ツイートから抽出したテキスト統計量，インタラクショングラフを利用して抽出したカスケードのグラフ特徴量，及び拡散に参加したユーザに関するユーザ特徴量で捉えることを考えた．以下で，順に説明する．

5.1　テキスト特徴量 (Text)

テキスト特徴量としては大きく分けて元ツイートの本文特徴量と情報カスケードに付随する Mt の本文特徴量を考える．

元ツイートの本文特徴量　元ツイートの本文特徴量としては，(1) 出現する単語 (BoW) (2) 本文の長さ ($\mathrm{len}_{\mathrm{Text}}$) (3) 固有表現が含まれるかどうか (has_ne) (4)URL が含まれるかどうか (has_url) の 4 種類を利用した．

まず，本文に出現する単語の特徴量としては，Bag of words を用い，具体的には，元ツイートの本文から URL，ユーザネームを除き，w, W の連続を一つの文字列と

する正規化を行った後 MeCab[1] で mecab-ipadic-NEologd[2] を辞書に用い形態素解析し，自立語の動詞，名詞 [3]，形容詞を用いた．本文の長さとしては，@から始まるユーザネーム，URL を除いた単語数を用い，長さが 70 以上であるかどうかで特徴量とした．

また，固有表現が含まれる場合，4.2 節で分析したように，拡散される情報に社会的影響力を持つ可能性が高いと考えられる．そこで，固有表現抽出を行い，ツイート中の固有表現の有無を特徴量として用いた．固有表現抽出は CaboCha[4] を用いて行い,，組織名 (ORGANIZATION)， 人名 (PERSON)， 地名 (LOCATION)，固有物名 (ARTIFACT) のいずれかの固有表現が含まれているかどうかで特徴量とした．

Mt の本文特徴量　Mt の本文特徴量としては，(1) 出現する単語 ($\mathrm{BoW}_{\mathrm{Mt}}$) (2)Mt の長さ ($\mathrm{len}_{\mathrm{Mt}}$) の 2 種類を用いた．

それぞれの特徴量は元ツイートの本文特徴量と同様にして求めたが，テキストの正則化の際に，非公式 RT や引用の際に広く用いられる”RT”, ”Rt”, ”rt”, ”QT”, ”Qt”, ”qt”から始まる元ツイートの引用，元ツイートと全く同様の文字列は削除することで，情報カスケードに対する反応のみを抽出した．また，本文の長さの閾値は，訓練データの分布を参考にし，長さが 0 であるか (引用やユーザ名のみの Mt)，長さが 22 より小さい，長さが 23 以上の 3 種類のどれであるかで特徴量とした．

5.2 グラフ特徴量 ($Graph$)

カスケードの伝播経路構造を特徴量として捉えることを目的にしてカスケードのグラフ構造を作成する．カスケードのグラフ構造はインタラクショングラフ G の RT したユーザの集合による部分グラフから作成されるが，親密なユーザ間でのやりとりに注目すること，仮想的な伝播経路を定めることを目的とし，エッジの残し方を

[1] http://taku910.github.io/mecab/
[2] https://github.com/neologd/mecab-ipadic-neologd/
[3] ただし接尾辞，数は除く
[4] http://taku910.github.io/cabocha/

表 5.1: カスケード構造グラフ

	双方向エッジのみ	片方向エッジ
全ユーザ	$G'(R)$	$\hat{G}(R)$
直前ユーザ	$G'_2(R)$	$\hat{G}_2(R)$

変えることで 4 種類のグラフ構造を得る．まずエッジの残し方を 1) 以前に RT したユーザで，インタラクショングラフ上でエッジがあるユーザ全てからのエッジを用いる場合 2) インタラクショングラフ上でエッジがあるユーザのうち直近に RT したユーザからのエッジのみを用いる場合の二種類を用いることでインタラクショングラフ G の情報をそのまま残したグラフ $\hat{G}$ と仮想的に伝播経路を定めたグラフ $\hat{G}_2$ を作成する．さらにそれぞれに対し双方向のエッジのみを残すことで得られるグラフ G',G'_2 を用いることでユーザ間の親密性を捉えたグラフを作成した．以上の 4 種類のグラフをまとめると表 5.1 のようになる．

今回用いたグラフ特徴量は表 5.2 に示す．大きく分けて，ルートユーザに関する特徴量 (Graph_root)，RT したユーザに関する特徴量 (Graph_RT)，グラフ構造に関する特徴量 (Graph_Structure) を提案する．前者 2 つは，ユーザ自身の特徴とインタラクショングラフ上で直接接続しているユーザとの関係を表し，グラフ構造に関する特徴量はカスケードの伝播の特徴を捉えることを目的としている．

RT したユーザに関する特徴量としては，インタラクショングラフ上での次数の分布を用いた．この分布は次数の逆累積度数分布を次数の軸において対数軸で 10 個の bin に分け，特徴ベクトルの各次元に対応させるという手法で特徴量とした．また，グラフ構造に関する特徴量としては連結成分，総エッジ数についてはどれだけ密なグラフであるかという指標として，深さはどれだけルートユーザから遠くまで伝播したかという指標として用いた．

表 5.2: グラフに関する特徴量

(a) ルートユーザに関する特徴量 (Graph_root)

$\hat{G}(V_0), \hat{G}_2(V_0)$ の outdegree(outdeg($\hat{\mathrm{G}}$(V_0)), outdeg($\hat{\mathrm{G}}_2$(V_0)))
$G'(V_0), G'_2(V_0)$ の degree (degree(G′(V_0)), degree(G′$_2$(V_0)))

(b) RT したユーザに関する特徴量 (Graph_RT)

$\hat{G}, \hat{G}_2$ の outdegree の分布 (outdeg($\hat{\mathrm{G}}$), outdeg($\hat{\mathrm{G}}_2$))
G', G'_2 の degree の分布 (degree(G′), degree(G′$_2$))

(c) グラフ構造に関する特徴量 (Graph_Structure)

$G', \hat{G}$ の最大の連結成分の大きさ (size_cc(G′),size_cc($\hat{\mathrm{G}}$))
$G', G'_2, \hat{G}, \hat{G}_2$ の総エッジ数 (#edges(G′), #edges(G′$_2$), #edges($\hat{\mathrm{G}}$), #edges($\hat{\mathrm{G}}_2$))
G' の深さの平均 (ave_depth)
G' の深さの分布 (dist_depth)
G' のクラスタリング係数の平均 (clustering)

5.3 ユーザ特徴量 (User)

ユーザ特徴量もテキスト特徴量と同様に，RT したユーザの特徴量 ($\mathrm{user}_{\mathrm{RT}}$) と，情報カスケードに付随する Mt のユーザ特徴量 ($\mathrm{user}_{\mathrm{Mt}}$) を考える．マイクロブログではユーザごとに RT するツイートの内容に偏りがあり，社会的影響力のあるカスケードばかりを RT するユーザや，ネタなど社会的影響力のないツイートばかりを RT するユーザが投稿を拡散しているかどうかが分類の手がかりとなる．そこで，テキストに対する Bag of Words を参考に，カスケードに参加するユーザを，いわば Bag of Users として特徴量にした．

このユーザの特徴量の次元は，それぞれカスケードサイズ 600 の学習データに含まれるユーザに対応しており，次元数はカスケードサイズ 600 のもので 137,167 次元，50 のもので 33,234 次元となった．そして，あるカスケードのユーザ特徴量を作る際にはそのカスケードをつぶやいたユーザの対応する次元を 1, それ以外の次元を 0 にすることでユーザの特徴ベクトルとする．

Mt を行ったユーザに対しても，情報カスケードのユーザ特徴量と同様にして，Bag of Users として特徴量とした.

第6章　社会的影響力を持つマイクロブログにおける情報カスケードの早期検知

本章では5章で述べた提案手法を用いて，社会的影響力を持つマイクロブログにおける情報カスケードの早期検知を行う実験を行う．4節で作成したデータセットを用い実験を行い，観測されたRTの数を変化させることで提案手法を用いることでどれだけ早期の段階で検知することができるかを確かめた．

6.1　社会的影響力を持つマイクロブログにおける情報カスケードの早期検知実験

本節では,50回以上RTが観測された情報カスケードに対して，2013年1月のカスケードによって分類器を学習し，2月のカスケードを自動で分類することによって，提案した分類手法でどれだけ正確に分類することができるかを評価する実験について述べる．

表 6.1: 正例，負例の分布

		負例						
	正例	50	100	200	300	400	500	600
1月	188	31291	12088	4749	2758	1815	1298	939
2月	103	16714	6014	2203	1195	784	520	364

6.1.1 実験手順

本研究の目的は早期の，つまりカスケードが広がらない段階でのカスケードの社会的影響力の有無の判断であるため，学習・評価の各カスケードは n 回以上 RT が観測されたカスケードとし，分類器の学習・評価を行った (n=50, 100, 200, 300, 400, 500, 600)．n が小さければ小さいほど候補となる情報カスケード（負例）の数が多くなり，正例と負例のバランスが偏ることから難しい問題となる．表 6.1 に，今回の学習セット，評価セットの正例負例の分布を示す．

分類器の学習の際，本文の長さの特徴量，グラフ特徴量は各次元が実数値を取るため，0 から 1 の間の値へ正規化を行い，分類器としては LIBLINEAR[1] 線形カーネルの SVM を学習した．また，分類のラベルに偏りがあるため，学習の際に正例側に重みをかけることで対応した．なお，パラメータチューニングは学習データにおいて 5 分割交差検定を用いて最大の F_1 値を取るパラメータを用いた．

6.1.2 実験結果

実験を行った際の検知結果を表 6.2 に示す．参考のため，各被験者と正解ラベル (多数決) との一致度もともに示す．どれも，全てのカスケードが社会的影響力を持つとした場合 (Baseline) の F_1 値と比較して改善していることが分かる．また，カスケードサイズが小さい (早期である) ほど F_1 値が顕著に小さくなっており，早期の段階で分類するのは難しいことが確認できる．この原因はカスケードサイズの分布が図 4.1 で示した通り，べき乗分布に従っているためであると考えられる．

次に表 6.2 のテキスト特徴量，ユーザ特徴量，グラフ特徴量それぞれのみで分類した場合の結果を参照すると，カスケードサイズが 300 以下であれば最も効果的な特徴量はグラフ特徴量であるため，早期検知という観点ではグラフ特徴量が有効であると考えられる．さらに，この結果からは，全ての特徴量を同時に使うよりもグラフ特徴量を単独で使う場合の方が良い性能を示しているということが分かる．

[1] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 6.2: 同時に解いた場合の特徴量による精度変化

	カスケードサイズ						
	50	100	200	300	400	500	600
ALL	0.142	0.189	0.289	0.338	**0.461**	0.587	**0.762**
Text	0.087	0.117	0.190	0.255	0.382	0.489	0.580
User	0.029	0.103	0.232	0.329	0.460	**0.606**	0.752
Graph	**0.186**	**0.300**	**0.350**	**0.373**	0.355	0.380	0.513
Baseline	0.012	0.033	0.086	0.147	0.208	0.284	0.361

一方で，カスケードサイズが大きくなった段階ではユーザ特徴量が有効に働いているということが分かる．

また，提案手法のカスケードサイズ 600 時での F_1 値は最も正解と一致しなかった被験者の F_1 値とはまだ差がある．この原因として考えられるのは．提案手法では被験者が参照していないユーザやグラフに関する特徴量を参照している点においては被験者より有利であるが，URL のリンク先（特に画像）を情報として利用できていないため，これによる差であると考えられる．

6.2 考察

6.2.1 分類結果の分析

各カスケードを SVM で分類する際の判断基準である，分離超平面からのマージンを 0 から動かすことで，分類器の精度と再現率のトレードオフを調査した．特に，二段階分類による手法では，一段階目は最大の F_1 値を取る閾値を用い，二段階目の分類平面のマージンを動かすことでトレードオフを調査した．図 6.1 にそれぞれのカスケードサイズで分類を行った際の適合率と再現率のトレードオフを示す．

適合率と再現率の関係は，カスケードサイズが 600 未満の時点においてはどの時点においても再現率をどこまで小さくした場合においても適合率の値を増加させることが難しいことが分かる．この原因として考えられるのは，本研究の対象とした

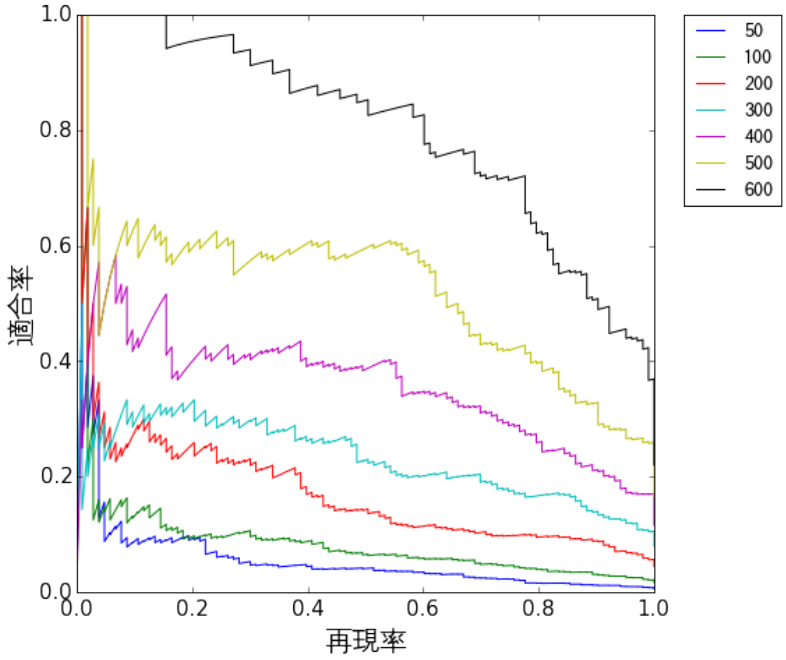

図 6.1: 同時分類のカスケードサイズによる精度変化

社会的影響力を持つ情報カスケードがカスケードサイズ 600 以上の情報カスケードであるため，そのサイズに満たなかった情報カスケードを誤検知してしまった可能性がある．

6.2.2 特徴量の有効性の分析

6.1.2 節で述べた通り，カスケードの大きさの段階によって有効に働く特徴量は変化している．この原因を探るために，6.4.1 節と同様に SVM の超分離平面を動かすことで分類のトレードオフを調査した．図 6.2 図 6.8 にそれぞれのカスケードサイズで分類を行った際の適合率と再現率のトレードオフを示す．カスケードサイズが 300 以下の時点においてはグラフ特徴量がもっともよい精度を示しているが，この際の適合率と再現率の関係は，適合率に比べて再現率を高く保てているということが分かる．この原因として考えられるのは,「カスケードサイズが 600 以上に成長するかどうか」という問題に対してグラフ特徴量は有効に働いている一方で，テキス

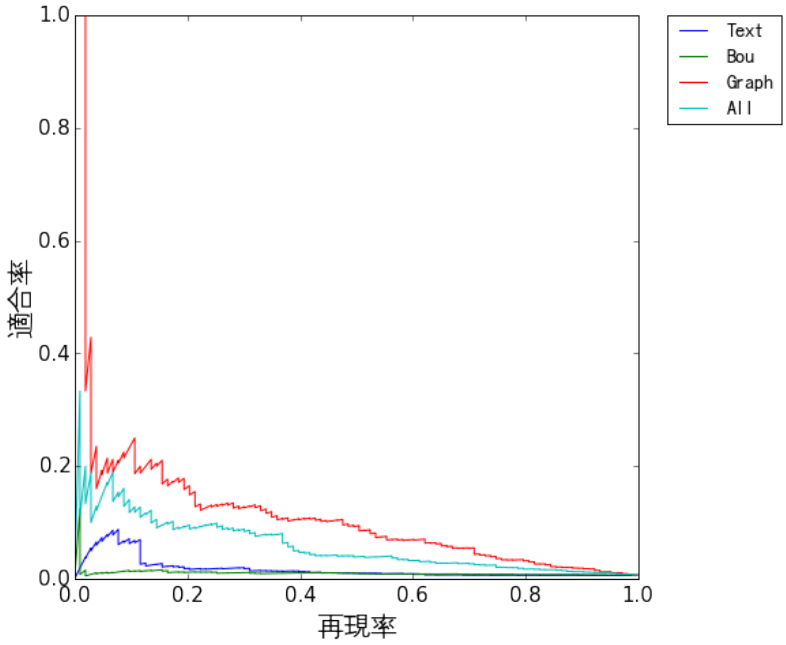

図 6.2: カスケードサイズ 50 時点のトレードオフ

ト特徴量，ユーザ特徴量は「成長するのか」という問題ではなく,「拡散する情報が社会的影響力を持つか」という基準に対して判断が行われているため，適合率が低くなってしまっている可能性がある．

次に，提案した特徴量の有効性を詳しく調査するために，提案した全ての素性から 1 つずつ素性を抜いた状態で学習，テストを行うことでアブレーションテストを実行した．その結果を表 6.3 にまとめる．この結果より，n=50 の時点で除くことで F 値が最も大きく下がる，つまり最も効果的に働いている特徴量は Graph であり，逆に除くことで最も F 値が大きく増加する，つまり分類に悪影響を与えている特徴量は BoU であることが分かる．一方で，n=600 の時点で最も効果的に働いている特徴量は BoU であり，悪影響を与えている特徴量は Mt の POS，グラフ特徴量であることが分かる．これらの結果から，どれだけ早期の段階で社会的影響力の有無を分類するかによって有効な素性が変化するということが判明した．具体的には，初期の段階であれば，カスケードが成長するかどうかという分類の判断に Graph 特徴が有効に働く一方で，カスケードが拡散した状態においては拡散経路の情報は社会的

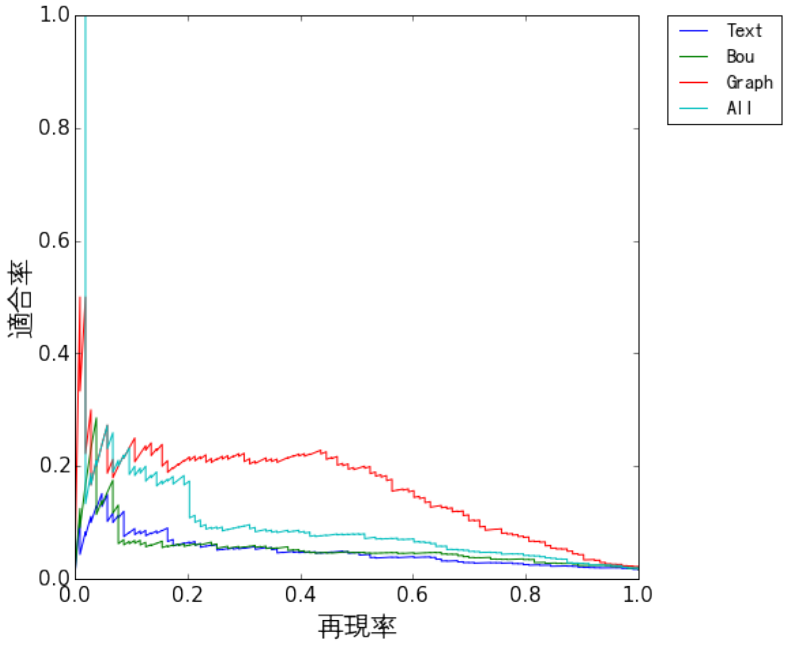

図 6.3: カスケードサイズ 100 時点のトレードオフ

影響力の有無の判断には有効ではないということが原因として考えられる．一方でユーザ特徴量は初期の段階では，初期の段階では判断基準となるユーザが十分でないため，手がかりとして不十分である一方で，カスケードサイズがある程度大きくなった段階においては，どのようなユーザが RT しているかということを見るだけで多くのカスケードの社会的影響力の有無を判断することができるようになるということが分かった．このことから，ユーザ毎に社会的影響力を持つ情報カスケードを RT するかどうかの傾向があるということも確かめられた．

この分析によりカスケードサイズによって社会的影響力の有無に基づく分類に有効な素性に変化が見られることが分かったため，次節では n=600 以上となる情報カスケードのみを用いて，分類実験を行うことで，情報カスケードの成長問題と分離して社会的影響力の有無のみの分類性能を調査する．

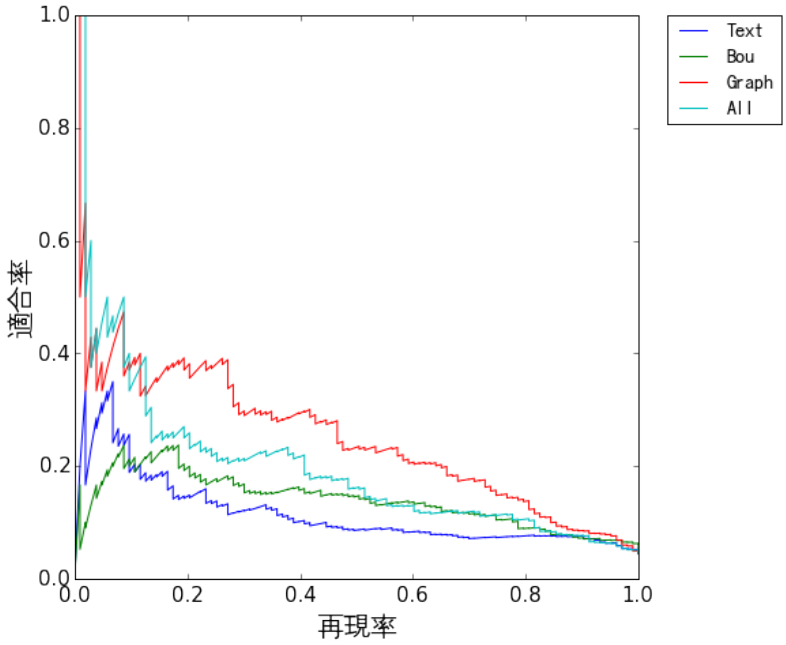

図 6.4: カスケードサイズ 200 時点のトレードオフ

表 6.3: アブレーションテストによる F 値の増減

除いた特徴量	50	100	200	300	400	500	600
ALL(基準)	0.142	0.189	0.289	0.338	0.461	0.587	0.762
Text	0.025	-0.032	-0.023	-0.003	0.009	-0.012	-0.021
bow_{text}	-0.003	-0.036	-0.038	-0.010	0.008	0.011	0.002
bow_{Mt}	-0.005	0.010	-0.020	0.022	0.016	0.021	-0.034
Graph	**-0.071**	**-0.051**	-0.024	0.007	0.025	0.028	-0.009
Graph_root	-0.018	-0.012	-0.016	0.008	0.002	0	-0.002
Graph_RT	-0.040	-0.035	-0.013	0.013	0.026	0.034	-0.009
Graph_Structure	-0.020	-0.026	-0.012	0.001	0.009	0.005	-0.002
user_{RT}	0.006	0.018	0.016	**-0.050**	**-0.075**	**-0.098**	**-0.152**
user_{Mt}	-0.007	0.010	**-0.051**	0.003	-0.004	0.018	-0.006

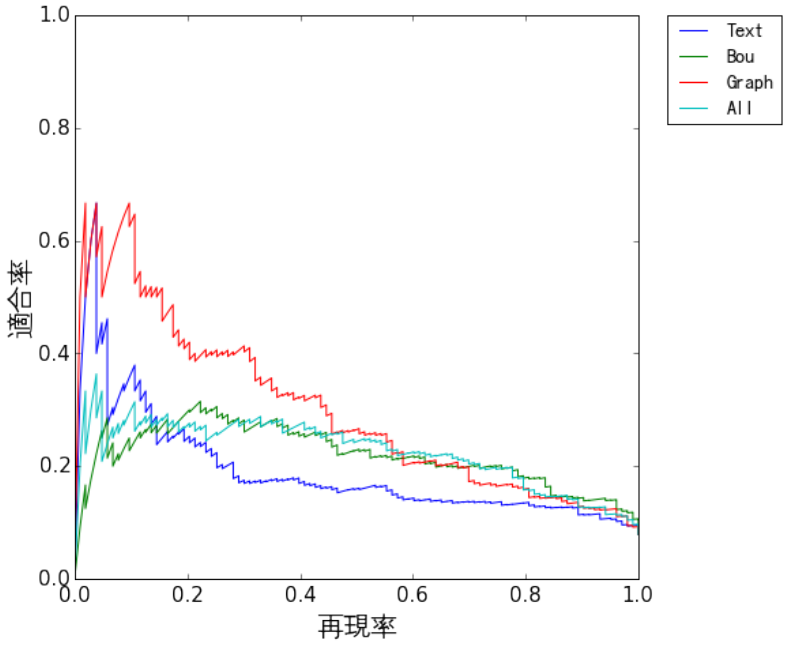

図 6.5: カスケードサイズ 300 時点のトレードオフ

6.2.3 600 回以上 RT が観測されたカスケードのみについての分類実験

本節では，600 回以上 RT が観測されたカスケードのみを用いて，2013 年 1 月のカスケードによって分類器を学習し 2 月のカスケードを自動で分類することによって，提案した特徴量でどれだけ正確に分類することができるかを評価する実験について述べる．これによって，カスケードサイズが 600 以上にまでカスケードが成長するかどうかという問題と切り離して，社会的影響力の有無のみを分類する問題とすることで，6.1 節で議論した素性の有効性がカスケードサイズに応じて変化することの原因について詳しく調査を行う．

6.2.3.1 実験手順

学習・評価の各カスケードには，6.1 節の実験と同様，4 章でアノテーションを行った情報カスケードの 1 月，2 月の情報カスケードをそれぞれ用いる．また，本章で

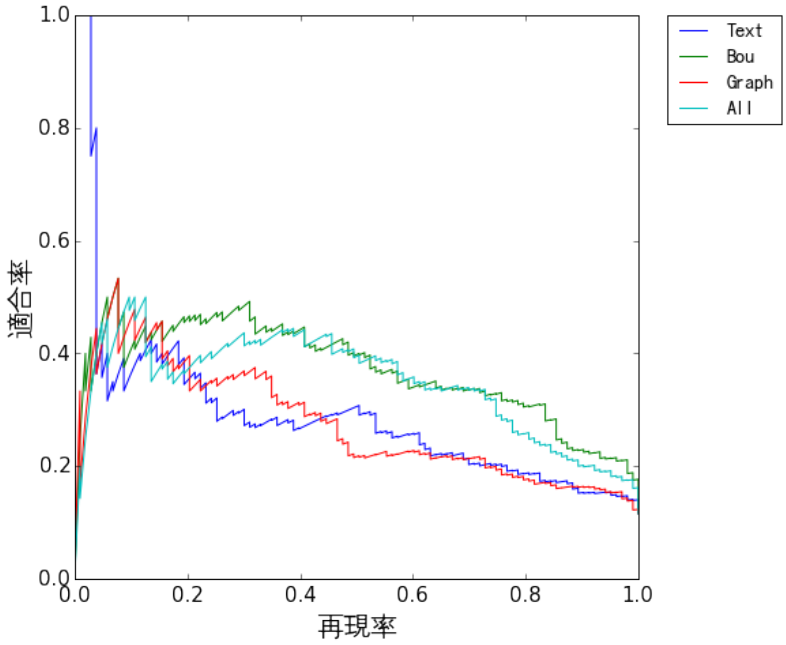

図 6.6: カスケードサイズ 400 時点のトレードオフ

の実験では，カスケードの成長問題から切り離して社会的影響力の有無の分類のみを考えるため，どちらも 600 回以上 RT が観測された情報カスケードのみを用いた．

しかし，早期検知というタスクに対して素性の有効性を調査を行う場合，情報カスケードを観測する段階に応じて，ユーザ特徴量や，グラフ特徴量は用いることができる情報は少なくなる．これを表現するために，初期の RT を先頭から n 件を取り出し，これを初期カスケードとみなして分類器の学習・評価を行う (n=50, 200, 400, 600)．これによって，初期カスケードの段階で,「将来 600 以上に成長するかどうか」，という成長予測問題が解けていた場合，その予測問題と組み合わせることで社会的影響力の有無の早期検知を実現することができる分類問題ということができる．

また，分類器としては 6.1 節と同様のものを用いた．

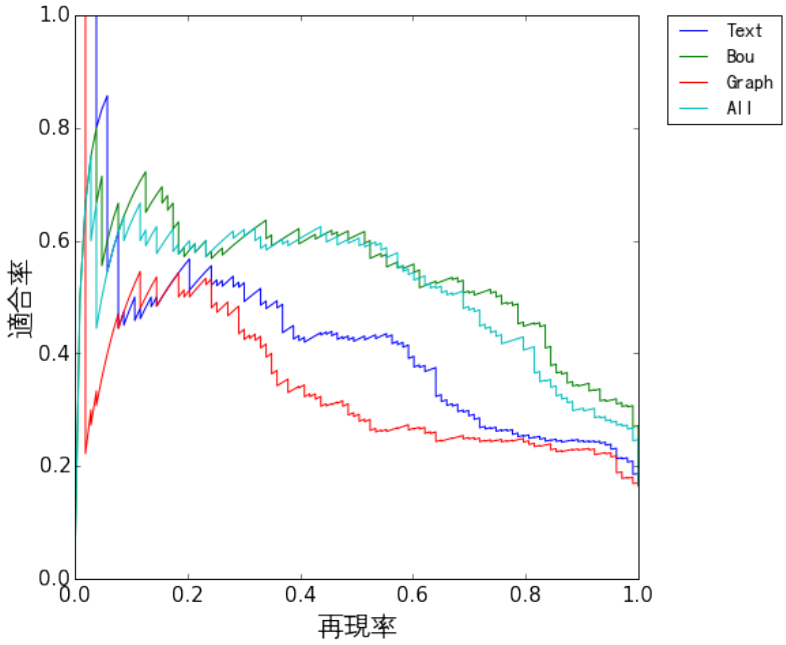

図 6.7: カスケードサイズ 500 時点のトレードオフ

表 6.4: カスケードサイズの変化に伴う F 値の変動

	カスケードサイズ						
特徴量	50	100	200	300	400	500	600
ALL	0.571	0.593	0.667	0.667	0.690	0.726	**0.762**
Text	0.580	0.580	0.580	0.580	0.580	0.580	0.580
User	0.601	0.650	0.694	0.711	0.721	0.730	0.752
Graph	0.484	0.482	0.507	0.541	0.508	0.510	0.513
Baseline	0.361	0.361	0.361	0.361	0.361	0.361	0.361
被験者 (A)	0.907	0.907	0.907	0.907	0.907	0.907	0.907
被験者 (B)	0.818	0.818	0.818	0.818	0.818	0.818	0.818
被験者 (C)	0.900	0.900	0.900	0.900	0.900	0.900	0.900

6.2.3.2 実験結果

特徴量を変えて実験を行った際の検知結果を結果を表 6.4 に示す．参考のため，各被験者と正解ラベル (多数決) との一致度もともに示す．この表から分かる通り，提

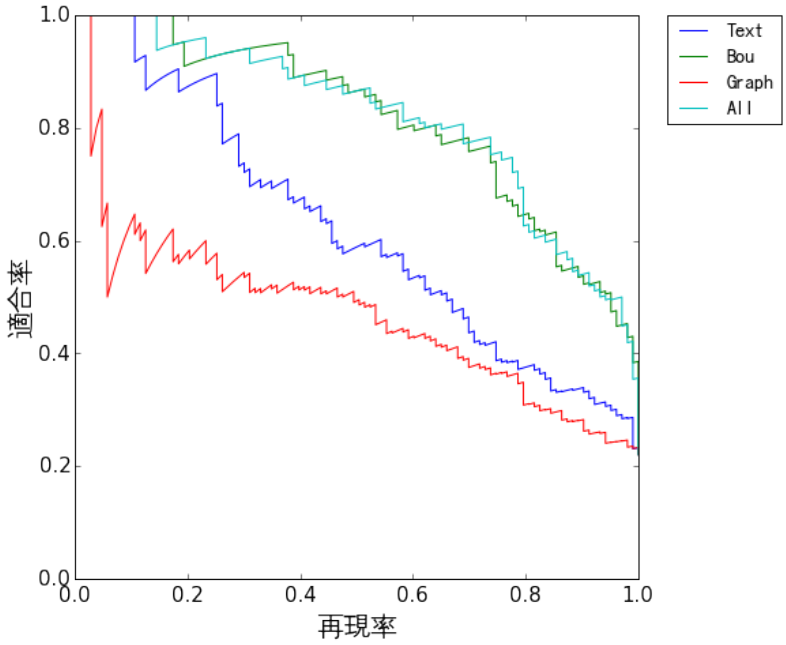

図 6.8: カスケードサイズ 600 時点のトレードオフ

案手法はカスケードサイズが全ての時点で全てのカスケードが社会的影響力を持つとした場合 (Baseline) の F 値 0.361 と比較して良い性能を示すということが分かる．

また，提案手法においてカスケードサイズに応じて F_1 値の変化が起こるのは，6.1 節の実験とは異なり，正例と負例の数の変化はないため，純粋な手がかりの数の変化が原因である．つまり，カスケードが成長していない段階ではユーザ特徴量，グラフ特徴量，Mt に基づく特徴量の効果は薄れ，カスケードのサイズに異存しない本文の特徴量の効果が大きくなると考えられる．

また，提案手法は，最も F 値の高いカスケードサイズ 600 のときでも 0.762 であり，最も正解と一致しなかった被験者の F 値 0.818 に対してはまだかなりの開きがある．提案手法では被験者が参照していないユーザやグラフに関する特徴量を参照している点において被験者より有利であるが，URL のリンク先（特に画像）を情報として利用できていないため，この点で差が出たと考えられる．6.2.3.5 節でこの点について考察を行う．

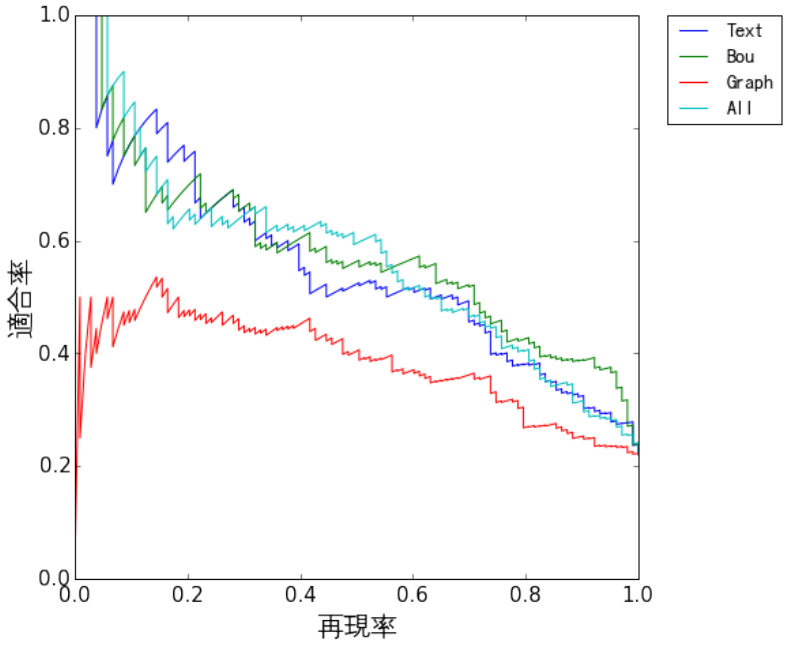

図 6.9: カスケードサイズ 50 時点のトレードオフ

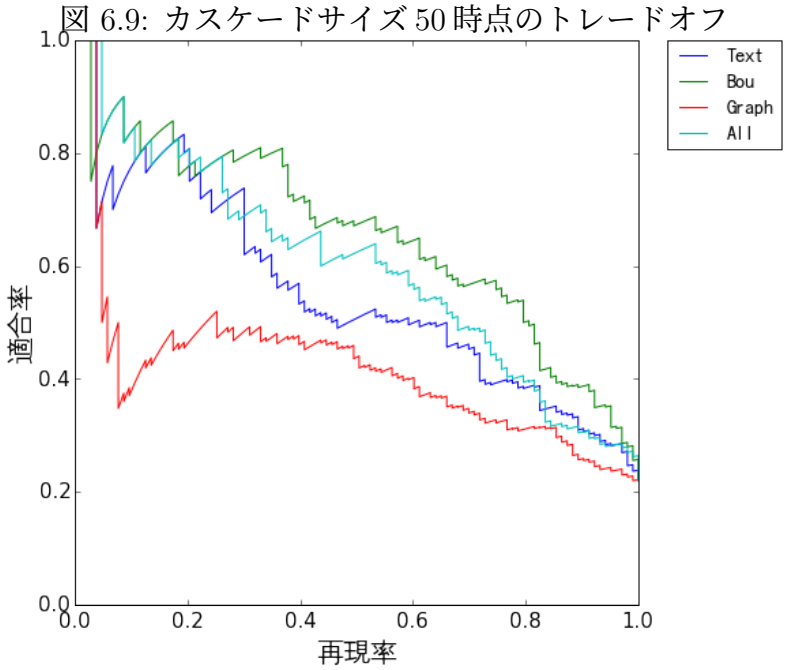

図 6.10: カスケードサイズ 100 時点のトレードオフ

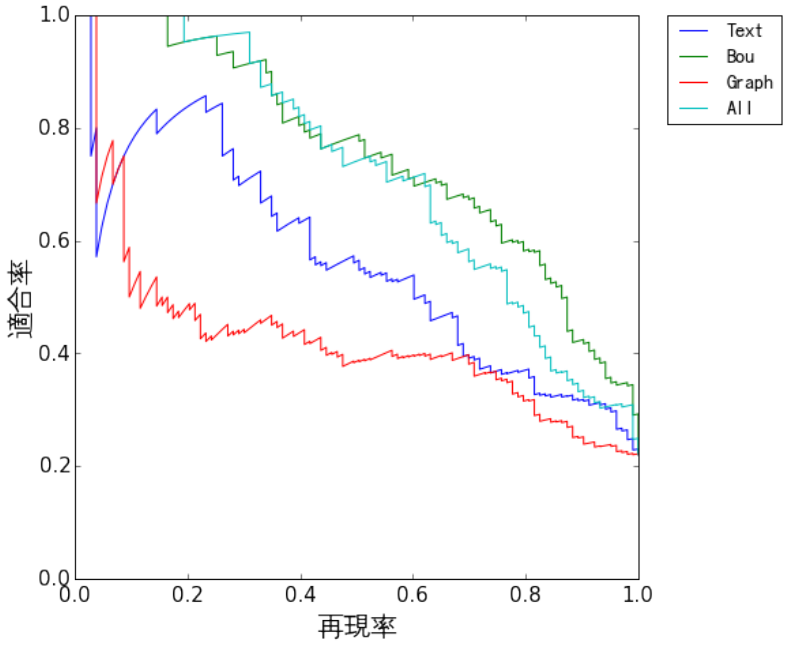

図 6.11: カスケードサイズ 200 時点のトレードオフ

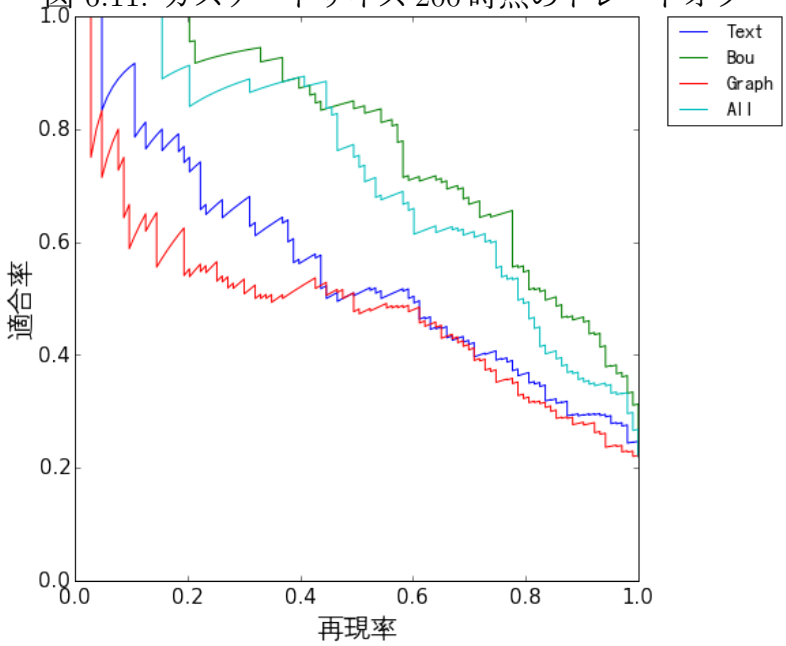

図 6.12: カスケードサイズ 300 時点のトレードオフ

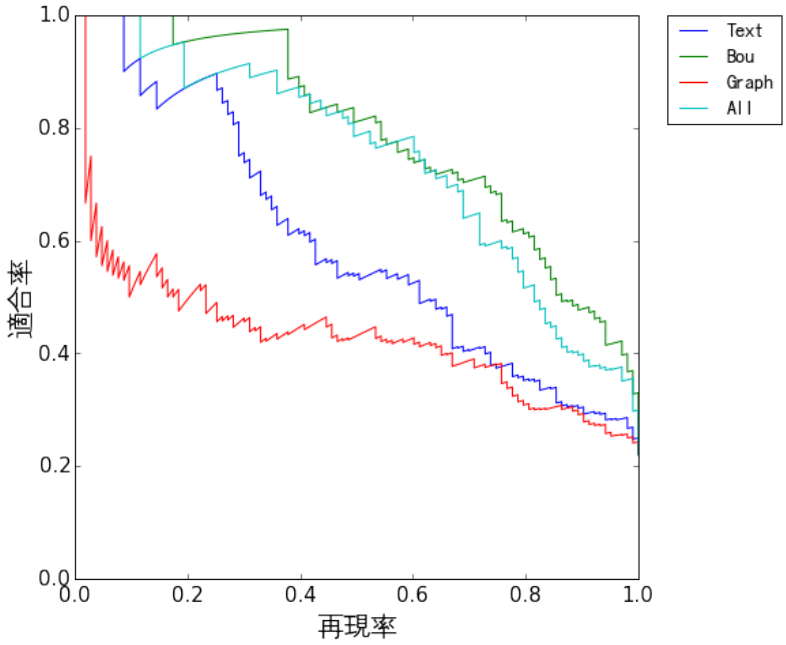

図 6.13: カスケードサイズ 400 時点のトレードオフ

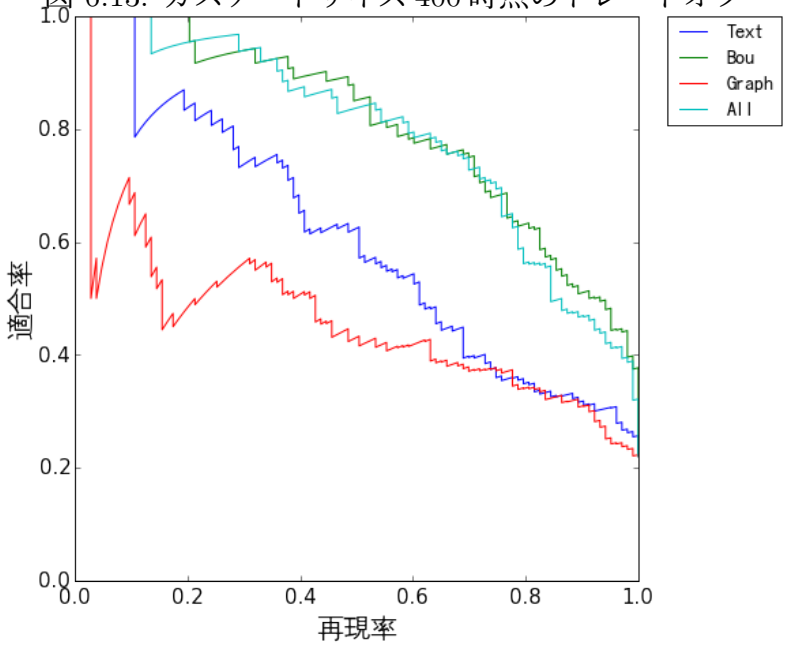

図 6.14: カスケードサイズ 500 時点のトレードオフ

表 6.5: アブレーションテストによる F 値の増減

除いた特徴量	50	100	200	300	400	500	600
ALL	0.571	0.593	0.667	0.667	0.690	0.726	0.762
Text	0.025	-0.032	-0.023	-0.003	0.001	-0.012	-0.021
bow_{text}	-0.003	-0.036	-0.038	-0.010	0.008	0.011	0.002
bow_{Mt}	-0.005	0.020	-0.003	0.015	0.020	0.007	
Graph	0.025	0.034	0.003	0.012	0.028	0.005	
Graph_root	0.003	0.008	0	-0.006	0	0.001	
Graph_RT	0.025	0.019	-0.007	0.003	0.032	0.009	
Graph_Structure	0.006	0.014	0.003	0	0.015	0.001	
user_{RT}	0.012	-0.025	-0.097	-0.093	-0.111	-0.153	
user_{Mt}	-0.006	-0.024	-0.011	-0.041	-0.020	0.001	

6.2.3.3 効果的な特徴量

6.1 節で行った実験と同様に，素性の有効性を調査するために，提案した全ての素性から 1 つずつ素性を抜いた状態で学習，テストを行うことでアブレーションテストを実行した．6.1 節とは異なり，本実験では「カスケードの成長」という要素は除かれた実験であるが，カスケードのサイズによる手がかりの多さは素性の有効性に影響する．

その結果を表 6.5 にまとめる．この結果より，カスケードサイズが 50 の時点で除くことで F 値が最も大きく下がる，つまり最も効果的に働いている特徴量は BoW であるということが分かる．それにつづく特徴量も has_url，len_text であり，本文特徴量が効果的に働いているということが分かる．一方で，早期予測に有効であった Graph は効果的に働いていないということが分かる．このことから，Graph 特徴量が有効に働いていたのは，カスケードが成長するかどうかという点についてであったということが確認できた．

6.2.3.4 カスケードの早期検知

次に，各カスケードを SVM で分類する際の判断基準である，分離超平面からのマージンを 0 から動かすことで，分類器の精度と再現率のトレードオフを調査した．図 6.9〜6.14 にカスケードサイズがそれぞれの時の適合率と再現率のトレードオフを示す．特にカスケードが小さい状態で有効であると考えられるテキスト特徴量は，図 6.9 のカスケードサイズ 50 の問題において，Recall が 0.4 以下と低い段階で，ユーザ特徴量と比べて良い性能を示している一方で，Recall が 0.4 程度の時点で性能に大幅な低下が見られる．ユーザ特徴量はユーザ特徴量と比べて Recall が上がった状態での性能の低下が少なく，これがを組み合さった結果，分類器の性能が向上しているということができる．

6.2.3.5 URL を含まないカスケードについての分析

図 6.9 のグラフから分かる通り，テキスト特徴量の性能は Recall が 0.4 程度の時点で性能の大幅な低下が見られる．この原因を探るためカスケード本文の URL の有無について分析を行った．2 月のデータで本文に URL が含まれるカスケードは全体 475 件中 252 件と，全体の半数以上であり，その多くは画像を添付したものであった．テキスト特徴量の性能低下の原因として，カスケードに添付された画像が投稿の社会的影響力の有無を判断する上で重要な手がかりとなっている例が多かったと考えられるため，URL を含まない情報カスケードについて提案手法の性能を分析する．図 6.15 に URL を含まないカスケードについてカスケードサイズが 50 の時の適合率と再現率のトレードオフを示す．図から分かる通り，図 6.9 と比較してテキストのみを手がかりに用いた場合の分類性能が大きく改善していることが分かる．つまり，テキスト特徴量のみを用いた分類器の分類性能低下の原因は，リンクされた情報（特に画像) によって社会的影響力を持つカスケードが存在するためだといえる．そのため，今後性能を改善させるためにはツイートに含まれるリンク先の画像の特徴量など新しい特徴量を設計する必要がある．

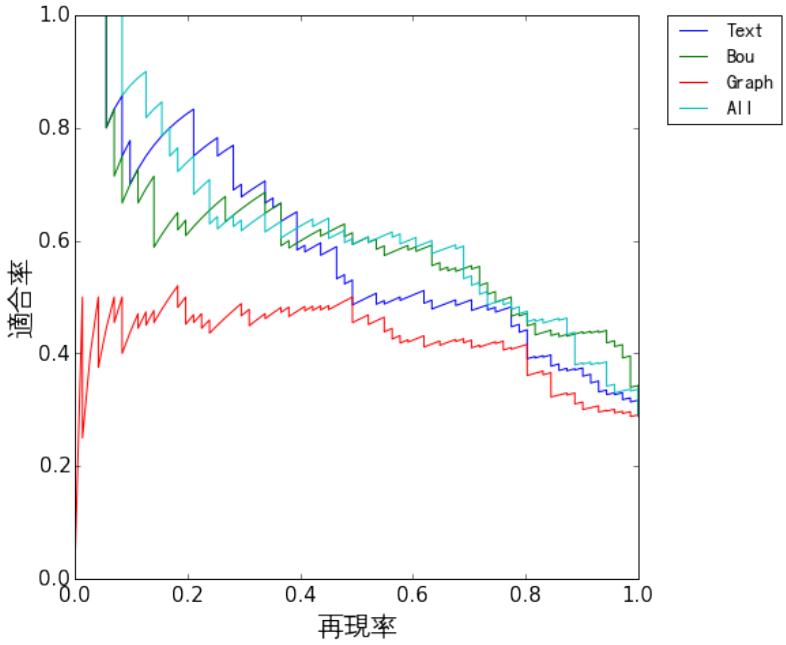

図 6.15: URL を含まないカスケードについての先頭 50 ユーザによる分類結果

6.3 成長予測と社会的影響力の有無の二段階で早期検知を行う手法

6.1,6.2.3 節の議論を踏まえると，テキスト特徴量，ユーザ特徴量は社会的影響力の有無の分類に有効であり，グラフ特徴量はカスケードが成長するかどうかの分類に有効であると考えられる．

本研究が目指す社会的影響力を持つ情報カスケードは「カスケードが成長」し,「社会的影響力を持つ情報」である必要があるため，早期検知のためには，カスケードが小さい段階での分類が求められる．また 6.2.3 節の手法は結果として成長しないカスケードが考慮されていないという点で現実には用いることができなかった．そこで，本節では，カスケードが成長するかどうかをグラフ特徴量を用いて予測し，その後，成長すると予測された情報カスケードに対し，6.2.3 節で作成した分類器を用い，社会的影響力の有無の予測を行うという手順を踏むことで社会的影響力を持つ情報カスケードの早期検知を目指す．これにより，テキスト特徴量，ユーザ特徴量，

表 6.6: 正例，負例の分布

		負例					
	正例	50	100	200	300	400	500
1 月	1127	31291	12088	4749	2758	1815	1298
2 月	467	16714	6014	2203	1195	784	520

グラフ特徴量それぞれの強みを有効に活かすことで精度よく分類できることが期待できる．

6.3.1 実験手順

学習，評価のカスケードは 4 でアノテーションを行った情報カスケードの 1,2 月の情報カスケードをそれぞれ用いる．

カスケードの成長予測 まず一段階目の分類問題は，単純な情報カスケードの成長予測問題であり，6.1 節の実験と同様，カスケードサイズが 50, 100, 200, 400, 500 の時点のそれぞれのスナップショットを用い，その情報カスケードの最終的な観測サイズが 600 に達するかどうかという問題を 1 月のデータで学習し，2 月のデータでテストを行う．つまり，正例は観測されたカスケードサイズが 600 以上のもので，不例は判断するカスケードサイズが小さい段階であればあるほど多くなるということになる．この分布は表 6.6 の通りとなる．分類の際に使う特徴量は 6.1 節 6.2.3 と同様，分類器も同様の LIBLINEAR を用いて実験を行う．

社会的影響力の有無の分類 一段階目のカスケードの成長予測の自動分類によって成長が予測された情報カスケードに対し，予測したカスケードサイズの段階に対応した 6.2.3 節で学習した分類器を用い，分類を行う．

表 6.7: カスケードの成長予測の F_1 値の変化

	カスケードサイズ					
	50	100	200	300	400	500
ALL	0.290	0.398	0.505	0.602	0.711	0.857
Text	0.102	0.208	0.375	0.545	0.695	0.854
User	0.150	0.297	0.449	0.580	0.701	0.855
Graph	**0.406**	**0.576**	**0.732**	**0.819**	**0.869**	**0.923**
Baseline	0.054	0.142	0.336	0.528	0.688	0.855

6.3.2 実験結果

カスケードの成長予測 特徴量を変えて実験を行った際の検知結果を表 6.7 に示す. Baseline は全てがカスケードサイズ 600 まで成長すると分類した場合である．この表から分かる通り，グラフ特徴量を単独で用いた場合が最も良い F_1 値を示していることが分かる．この結果より 6.1 節，6.2.3 節で議論した Graph 特徴量はカスケードの成長予測に大きく貢献しているという予想が確かめられた．また，この実験により，Graph を単独で用いる場合が最も正確に分類することができることが分かったため，第一段階のカスケードの成長予測の出力としては，訓練データの交差検定で最も良い F_1 値を示したパラメーターを用いて学習した Graph を単独で用いた分類器を用いた.

カスケードの社会的影響力の有無の分類 第一段階により出力されたカスケードにはテストデータ内に存在する社会的影響力を持つ情報カスケード全てが含まれるとは限らず，二段階の手順で分類する方法

表 6.8: 同時に解く場合と二段階で解く場合の F_1 値の比較

	カスケードサイズ						
	50	100	200	300	400	500	600
提案手法 (同時)	0.131	0.158	0.275	0.342	0.463	0.592	0.748
提案手法 (二段階)	0.131	0.158	0.275	0.342	0.463	0.592	0.748
Baseline	0.012	0.033	0.086	0.147	0.208	0.284	0.361
被験者 A	n/a	n/a	n/a	n/a	n/a	n/a	0.907
被験者 B	n/a	n/a	n/a	n/a	n/a	n/a	0.818
被験者 C	n/a	n/a	n/a	n/a	n/a	n/a	0.900

6.4 考察

6.4.1 分類結果の分析

各カスケードを SVM で分類する際の判断基準である，分離超平面からのマージンを 0 から動かすことで，分類器の精度と再現率のトレードオフを調査した．一段階目は最大の F_1 値を取る閾値を用い，二段階目の分類平面のマージンを動かすことでトレードオフを調査した．図 6.16 にそれぞれのカスケードサイズで分類を行った際の適合率と再現率のトレードオフを示す．

図 6.16 から分かる通り，直接解いた場合に比べ再現率が小さくなっていることが分かる．この原因は第一段階でカスケードの成長予測を行った段階で正例となる社会的影響力を持つ情報カスケードが落ちてしまっていることである．これを防ぐ方法として，第一段階の成長予測の時点で社会的影響力を持つ情報カスケードに対してはより大きな重みを付けて学習を行うことなどが考えられる．

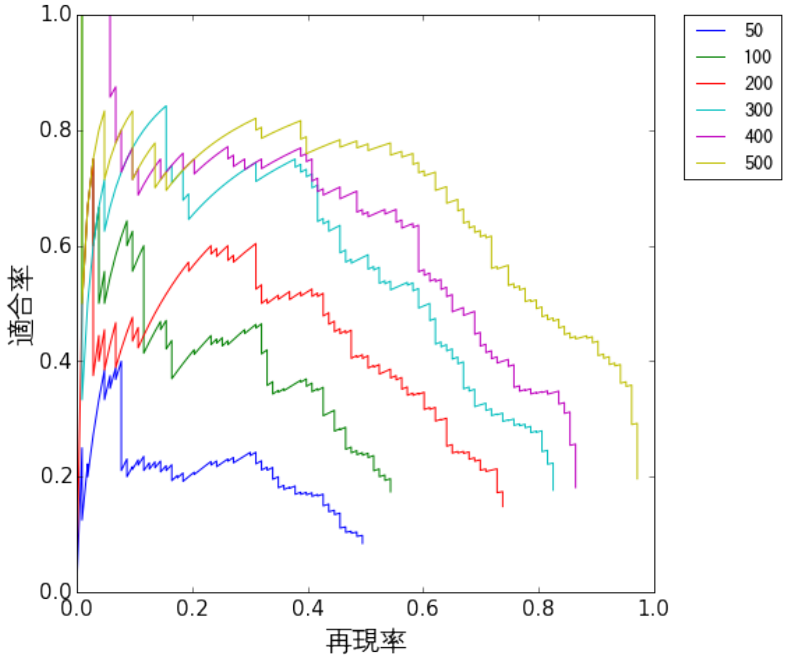

図 6.16: 二段階分類による分類結果

表 6.9: 二段階で解いた場合の特徴量による精度変化

	カスケードサイズ						
	50	100	200	300	400	500	600
ALL	0.268	0.378	0.467	0.545	0.585	0.670	0.748
Text	0.243	0.372	0.417	0.445	0.498	0.533	0.580
User	0.227	0.382	0.475	0.571	0.615	0.673	0.752
Graph	0.199	0.303	0.359	0.435	0.415	0.472	0.513
Baseline	0.012	0.033	0.086	0.147	0.208	0.284	0.361

第7章　おわりに

本研究では社会的影響力を持つ情報カスケードの早期検知という新しいタスクを提案し，これを教師あり学習に基づく分類器を用いて解く手法を提案した．

社会的影響力を持つ情報カスケードの性質を明らかにするため，被験者実験により収集した社会的影響力のある情報カスケードについて，その内容を分類した．この知見に従い，本文に含まれる単語情報やカスケードのグラフ構造の特徴量，カスケード毎のユーザ特徴量を特徴量とした分類手法を提案した．実験では，社会的影響力を持つ情報カスケードの自動検知を行い，カスケードサイズ (RT 数) が 50 の段階で，F 値として 0.576，さらに 600 の段階で F 値 0.748 で検知が可能であることが示された．これは全てのカスケードが影響力有りだとした場合 (Baseline) の F 値 0.39 と比較して顕著な改善である．

さらにテキスト特徴量の性能について詳細な分析を行った結果，テキスト特徴量は URL を含まないカスケードについてはよい分類性能を示していることがわかり，URL の中でも多くを占める画像によって社会的影響力をもつ情報カスケードによって性能が低下していることが分かった．

今後の課題としては，(1) 訓練データの拡充，(2) 小さな情報カスケードに対するアノテーション，(3)URL を含む情報カスケードに対する新しい特徴量の考案の三点が挙げられる．以下，これらの点に関して詳細を述べる．

訓練データの拡充　訓練データの拡充については今回アノテーションを行ったデータセットは 2 月分のデータに限られており，訓練セット 1127, テストセット 467 とかなり小さいものであったため，数年分のデータを用いるなど，よりデータセットを大きくすることが求められる．しかし，データのアノテーションを行う限り大規

模化することは難しいという問題点がある．そのためこれを解決するために自動で訓練データを取得する方法を検討したい．

小さな情報カスケードに対するアノテーション　二点目の小さなカスケードに対するアノテーションについては，本研究ではマイクロブログ中で大きく拡散することで社会的影響力を持つと考え，サイズが 600 以上となった情報カスケードに対してアノテーションを行った．しかし，広まらずに収束するような批判を早期に発見するというニーズなど，それほど拡散していない情報カスケードについて社会的影響力の有無を分類することもまた重要な問題であると考えられる．そのため，小さな情報カスケードに対するアノテーション，自動検知の実験も重要な問題である．

URL を含む情報カスケードに対する新しい特徴量　6.2.3.5 節で考察した通り，URL を含まない情報カスケードに比べて，URL を含む情報カスケードを多く間違えているということがわかっており，精度向上のためには URL を含む情報カスケードを分類するのに効果的な特徴量を導入することが必要である．特に多く存在する画像へのリンクに対しては画像特徴量の導入，外部記事へのリンクであればその記事のテキスト特徴量などが考えられる．

参考文献

[1] Sushil Bikhchandani, Ivo Welch, and David A. Hirshleifer. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Political Economy*, pages 992–1026, 1992.

[2] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of International World Wide Web Conference*, pages 925–936, 2014.

[3] Pei-Chi Chen, Hahn-Ming Lee, Hsiao-Rong Tyan, Jain-Shing Wu, and Te-En Wei. Detecting spam on twitter via message-passing based on retweet-relation. In *Proceedings of Technologies and Applications of Artificial Intelligence*, pages 56–65, 2014.

[4] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, 2011.

[5] Danah. Boyd, Scott. Golder, and Gilad. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of Hawaii International Conference on System Sciences*, pages 1–10, 2010.

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of International World Wide Web Conference*, pages 591–600, 2010.

[7] 奥村 学. マイクロブログマイニングの現在 (第3回集合知シンポジウム). **電子情報通信学会技術研究報告**. *IBISML,* **情報論的学習理論と機械学習**, 111(480):77–84, 2012.

[8] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. In *Proceedings of Workshop on Online Social Networks*, 2010.

[9] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of ACM International Conference on Web Search and Data Mining*, pages 65–74, 2011.

[10] Oren Tsur and Ari Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of ACM International Conference on Web Search and Data Mining*, pages 643–652, 2012.

[11] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506, 2009.

[12] 鳥海不二夫 and 榊剛史. STAP 問題に対するソーシャルメディアにおける反応の分析. In *Proceedings of WebDBForum*, 2014.

[13] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *Proceedings of The International AAAI Conference on Web and Social Media*, pages 586–589, 2011.

[14] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. Analyzing and predicting viral tweets. In *Proceedings of International World Wide Web Conference*, pages 657–664, 2013.

[15] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Proceedings of The International AAAI Conference on Web and Social Media*, pages 355–358, 2010.

[16] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of International World Wide Web Conference*, pages 695–704, 2011.

[17] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3(2522), 2013.

[18] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of ACM Web Science conference WebSci*, pages 8:1–8:7, 2011.

[19] Masaru Kitsuregawa Geerajit Rattanaritnont, Masashi Toyoda. Characterizing Topic-Specic Hashtag Cascade in Twitter Based on Distributions of User Influence. In *Proceedings of Asia-Pacific Web Conference*, pages 735–742, 2012.

[20] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of Special Interest Group on Information Retrieval*, pages 841–842, 2010.

[21] Zhaochun Ren, Maria-hendrike Peetz, Shangsong Liang, Willemijn V. Dolen, and Maarten De Rijke. Hierarchical Multi-Label Classification of Social Text Streams. In *Proceedings of Special Interest Group on Information Retrieval*, pages 213–222, 2014.

[22] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of International World Wide Web Conference*, pages 675–684, 2011.

[23] James Allan, Jaime Carbonell, George Doddington, et al. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[24] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of International Conference on Machine Learning*, pages 113–120, 2006.

[25] Xiangmin Zhou and Lei Chen. Event detection over twitter social media streams. *The International Journal on Very Large Data Bases*, 23(3):381–400, 2013.

[26] 斎藤 翔太，冨岡 亮太，山西 健司. ソーシャルネットワークにおける長期間流行する話題の早期検出. **電子情報通信学会技術研究報告**. *IBISML,* **情報論的学習理論と機械学習**, 111(480):77–84, 2012.

[27] Sven Rill, Dirk Reinel, Jörg Scheidt, and Roberto V. Zicari. PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69(1):24–33, 2014.

[28] Hongyu Gao, Yan Chen, Kathy Lee, et al. Towards online spam filtering in social networks. In *Proceedings of The Network and Distributed System Security Symposium*, 2012.

[29] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, (5):378–382, 1971.

[30] Richard J. Landis and Gray G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

発表文献

論文誌

1. 川本 貴史，豊田 正史，吉永 直樹，マイクロブログからの社会的影響力を持つ情報カスケードの検知手法．情報処理学会論文誌データベース (TOD)，2016.06.(投稿中)

査読付き国内会議

1. 川本 貴史，豊田 正史，吉永 直樹，マイクロブログにおける社会的影響力を持つ情報カスケードの早期検知に向けて．第 8 回 Web とデータベースに関するフォーラム論文集, pp.48-55 (WebDB Forum 2015)，東京，2015.

査読なし国内会議

1. 川本 貴史，豊田 正史，社会問題に関する情報カスケード検出．第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2015)，福島，2015.03.

2. 川本 貴史，豊田 正史，吉永 直樹，マイクロブログからの社会的影響力を持つ情報カスケードの早期検知．第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2016)，福岡，2016.03. (発表予定)

ポスター発表

1. 川本 貴史, 豊田 正史, 吉永 直樹, マイクロブログにおける情報カスケードの社会的影響力に基づく分類手法に関する検討. 言語処理若手の会 第 10 回シンポジウム (YANS2015), 石川, 2015.

付録A 情報カスケードで拡散するツイートに対する社会的影響力の有無のアノテーション ver 1.2.1

概要

皆さんにはツイッター上でRTにより拡散した投稿（ツイート）を一例ずつ読んでもらい、**そのツイートに含まれる情報が広まることで行動や意思決定に影響を受ける人が存在するかという観点で、社会的影響力をもつツイート（情報カスケード）となるかを判断してもらいます。**

データ形式

データは1行1ツイートの形式でファイルに保存されているので、ファイルを直接編集して影響力を持つと判断した場合は行頭に1を、持たないと判断した場合は0を注釈付けしてください。なお、行頭に付与されたタブは置換・削除せず、影響力の有無のみを追記する形で編集してください。

アノテーション基準

ツイートに書かれた情報が社会的影響力を有するか判断するにあたっては、主観性を極力排するため、ツイートに書かれた情報を知ったり、その情報を不特定多数

に知られたりすることで、直接的あるいは間接的に行動や意思決定に影響を受けるか人がいるかを客観的に想像して、影響力の有無を付与してください。

社会的影響力を有する情報（意見・事実）の例としては，具体的には例えば、

- 常識的に考えて、他人の反感を生む意見であれば、炎上して発言者が釈明する可能性がありますし、共感を生む意見であれば、共感した人々の行動に影響を与える場合があります。また個人・組織・業界・社会を対象とする意見であれば、その意見が広がることで、意見の対象に属する人が、意見に回答する必要が生じる場合があります。

  例)

  - 同性愛者を差別する発言が炎上して政治家が釈明する
  - 憲法 9 条にノーベル平和賞を、という主張が広まり実際に署名活動が行われる
  - 食品の異物混入に関する苦情を受けて、食品会社が謝罪する

- 投稿時点において周知されていない出来事（速報性のある事件）、大衆の認識と異なる事実（注意喚起・デマ訂正）、あるいは人々の知らない啓蒙的知識であれば、事実に関与する人や情報を知った人の行動や意思決定に影響が及ぶことがあります。

  例)

  - 竜巻の発生を報告する投稿を見て、マスコミが投稿者に取材する
  - 地震で負傷したというデマの訂正投稿を読み、救助に向かうのを止める
  - デング熱ではアスピリンを飲むと危険であると知り、飲む薬の種類を変える

があります。なお、ツイートに URL が含まれる場合は、リンク先も参照して影響力の有無を判断してください。

なお、注釈付けに関して判断に迷う場合もあると思いますが、上記の基準を元に直感で判断を行い、一旦注釈付けを行ったら、前に遡って修正を行わないでください。

注釈例

以下に、注釈付けの例を列挙します。

意見の例

意見に否定的な反響があった例

> 最近のマスコミの報道は倫理観に欠けている、何でも珍しいいことがあれば良いネタのようにして報道する、報道したことでその人物はなおさら優越感が出るのだ、一例が同性愛とやらだ！生物の根底を変える異常動物だということをしっかり考えろ！マスコミで取上げる影響を考えろ！まじめ人間が馬鹿を見る
>
> https://twitter.com/turusashi_masum/status/670638601424740352

炎上した結果，発言の取り消し，謝罪を行うことになった．

意見に肯定的な反響があった例

> やっと見つけた。日本語バージョンのポスター!! これは大切〜!! どんどん広めたい。 学ぼうとしない無知な飼い主に腹を立てるより、ここから学んでもらえるかも〜。 ドンドン拡散お願いします!!! <図>
>
> https://twitter.com/giselleai/status/322677702899417089

Yellow dog project に対し大切であり，広める必要があるという意見を述べて，それに賛同した人々によって広められている例である．

意見の対象が個人・組織・業界 意見の対象が何らかの対応が必要になる可能性があると考えられる場合重要

これはひどい。加工されていたのは人質映像じゃなく、専門家のコメントの方だった！／映像屋のざれごと：共同通信社さん配信の報道に関して。

https://twitter.com/hirorin0015/status/558813529826791425

共同通信社が報道が婉曲されていることを批判されており，謝罪報道などが必要になる可能性があるため重要である．

これだけ世間が「ミスをしたら徹底的に叩く」ような空気になってるのに、子供たちや若者たちに「ミスを恐るな」とか「失敗は成功のもと」だとか言っても、何の説得力もないと思うんですけどね。

https://twitter.com/isa_kent/status/553377512164564992

意見の対象は明確ではないが,「ミスを恐れるな」とか「失敗は成功のもと」とか言う指導者を批判しており，日本の教育指導方針などへの意見として重要である．

意見の対象が社会 世論を反映している場合重要である．

スポーツなどで国際的に活躍すると、「同じ日本人」として思いっきり共感するのに、紛争地帯で拘束されたりすると、いきなり「自己責任」と言って突き放してしまう冷たさは何なのか。

https://twitter.com/hiranok/status/558111770318217217

日本における自己責任という風潮に対する批判であり，世論に対して疑問を投げかけているという意味で重要である．

事実の例

速報 事件・事故など緊急性が高く，社会に対して大きな影響を及ぼすと考えられる出来事の周知が該当する

注意喚起・デマ訂正 ユーザに対する警告，情報の訂正など緊急性が高い情報の周知が該当する

> ここ最近、アカウント主が不在中に「アプリ連携」を使ってアカウントを乗っ取りスパムをツイートする業者が流行ってるようなので、「アプリ連携のチェック＆解除の手順」を改めてツイートします。スパムを呟いている方に解除を勧める際ご利用ください。　<図>，<図>
>
> https://twitter.com/tarareba722/status/559893718216359936

スパムへの対処法という緊急性の高い情報の周知という意味で重要．

啓蒙 緊急性はないが，知られていない事実の周知，社会に対する情報を拡散するもの

> イスラム国の日本人殺害予告 身代金要求ビデオの人質の影が左右違うのでクロマキー合成の可能性があります。
>
> https://twitter.com/motoshiromizu/status/557439171216691200

「合成の可能性がある」という知られていない事実を周知するものであるため重要である．

ボーダーケースの例

ネタや社会批判などが同時に含まれる例

> そう教えてくれた世代が特攻してくる　<図>
>
> https://twitter.com/TsuyoshiWood/status/556389151012970496

この例では，画像がネタと年代，社会批判を同時に含んでいる．このように批判的な要素とネタの要素が同時に含まれる場合も社会的影響力が有るとする．

一部の人に対して有用な情報の例

> ここ最近、時折キッチンの方から「ドンッ！」という異音が聞こえるオカルト現象に悩まされていたのですが、なんと先ほどその霊の正体と対面！知らないまま生きて行きたかった。。。（´－ι_－｀）
>
> #炭酸凍らせたら爆発するで <図>
>
> https://twitter.com/chiyomaru5pb/status/556386686758682624

この例は，炭酸を凍らせたら爆発するという事実を知らない人としては重要なカスケードと言えるため社会的影響力があるとする．

格言的なもの

> そしてさっき、水木しげる先生のありがたい言葉に心を打たれていたのだ。「しないではいられないことをし続けなさい」とか、「努力は人を裏切る」などすごい。　<図>
>
> https://twitter.com/aki_u_ench/status/552224737615556608

格言的なものについては，知識的なものではなく思想について語るものは社会的影響力のないものとする．