卒業論文

# テキスト分析に基づくソーシャルメディア上での ニュースの影響度予測に関する研究

03-130937　**上子 優香**

**指導教員　森 純一郎 特任講師**

2015 **年** 2 **月**

**東京大学工学部システム創成学科知能社会システムコース**

## 概要

ソーシャルメディアの普及により、これまで報道者から大衆への一方向でしかなかった情報の動きが、双方向へと移り変わり、大衆がジャーナリズムにおいて果たす役割が大きくなっている。

本研究はジャーナリズムやソーシャルメディアに注目を集める研究者間で高まるソーシャルニュースへの関心と、キュレーションメディアの台頭、企業や政府官公庁に存在する効果的な広報戦略に対するニーズに着目し、従来研究においてニュースコンテンツに含まれる言語的特徴が大衆の反応へ及ぼす影響は未だ解明できていない為、手法の提案と実験を行うことにより言語的特徴の持つ影響力を明らかにすることを目的とする。

提案手法は、ニュース記事に対するコメントの感情度とコメント数を大衆に対する影響の指標と捉え、それらをニュースコンテンツの言語的特徴に重点を置くことにより予測する、というものである。データセットとしては、Yahoo!ニュースのトピックスに取り上げられた記事を対象として、Twitter の関連ツイートと Yahoo!ニュースのコメント欄のコメントを収集し、関連ツイートからは投稿者本人が書き足したコメント部分を抽出した後に、サポートベクターマシンを用いてニュースコンテンツの言語的特徴を含む記事の特徴から、コメントの感情度の回帰とコメント数の分類を行い評価をする。

実験においては上記の提案手法を実施し、Twitter と Yahoo!ニュースのコメント欄に投稿されたコメントの感情度をそれぞれ回帰すると共に、各メディアにおける上位 50%、上位 10% となる境界値によるコメント数の分類を行い、さらにカテゴリ別の分類にも取り組んだ。

結果として最も評価値が高かった素性の組み合わせにおいて、コメントの感情度はコメントツイートで決定係数 0.449、Yahoo!ニュース内コメントで決定係数 0.693、コメント数の分類では上位 50% を境界としたコメントツイートの分類で F 値 0.675、Yahoo!内コメントで F 値 0.709、上位 10% を境界としたコメントツイートの分類で F 値 0.707、Yahoo!内コメントで F 値 0.741 となり、ニュースコンテンツの言語的特徴が大衆の投稿するコメントに対して少なからず影響を及ぼしていることがわかった。

# 目次

# 図目次

# 表目次

# 第 1 章

# 序論

## 1.1 研究背景

ソーシャルメディアの発達により、社会における従来の情報「受信者」であった大衆の役割が大きく変わりつつある。従来新聞やテレビを媒介として報道に携わる「発信者」が提供したニュースコンテンツを「受信」するだけであった人々は、ソーシャルメディアを通じて情報を自由に「発信」することが可能になり、「発信者」としての側面を強めていくこととなった。以前より社会的影響力を持つ著名人の中で従来の「発信者」を介することなく直接意見を「発信」するによりその影響力を強める者もいれば、また無名であった大衆の中からも、ソーシャルメディアという場で「発信」することにより新たに社会的影響力を持つ者も現れ始めた。

一方で、ソーシャルメディアの発達により「発信者」であった報道メディアは大衆がソーシャルメディア上で「発信」したコンテンツを受け取る「受信者」としての側面を得ることになった。大衆が「発信」する反応を自身の「発信」コンテンツの向上に繋げるなど、その有用性を理解するにつれて報道メディアにとって大衆がソーシャルメディア上で「発信」する情報は見逃せないものとなっている。

このように、これまでは報道メディアから大衆への一方向の情報伝達であったものが、ソーシャルメディアの発達により報道者と大衆との双方向の情報伝達へと変化を遂げた。

## 1.2 研究目的

本研究の目的は「発信」されたニュースコンテンツが、「受信者」である人々の反応にどのように影響を及ぼしているかを明らかにすることである。

ソーシャルメディアによる報道者と人々との交流の深まりだけでなく、オンラインニュースの普及に端を発し今なお続く、ジャーナリズム産業の体制を一変させる動きに対して、研究者は高い関心を持っている。報道者ではなく一般の人々の手によってニュースが作成され, 議論が交わされる“Digg”などに代表されるニュースサイトの登場から“social news”、“data-driven journalism”という言葉も生まれ、ソーシャルメディア上のニュース関連コメントの抽出が研究トピックとして取り上げられる (SNOW Workshop:http://www.snow-workshop.org) など、新しいジャーナリズムは今脚光を浴びている。

ジャーナリズム産業に注目しているのは、研究者だけではない。“Gunosy”や“Smartnews”といったニュース量の過多に伴う人々の取捨選択需要の高まりに着目した、個人の嗜好を考慮したニュース

提供アプリの登場は人々のニュースへの接し方にも影響を与えており、スマートフォンの普及と相まって現在急成長を遂げているキュレーションメディア業界に、企業も注目している。

キュレーションメディアでは、消費者に対してどのようなレイアウトで、いかに消費者を満足させられる記事を提供できるかが人気を左右するため、人々の反応に関する新たな知見が求められている。また従来から企業や政府官公庁は、ニュースとして「発信」された際の人々の反応を考慮した、報道発表の効果的な配信方法に対する情報を求めており、ニュースに対する人々の反応に纏わる研究への需要は高い。

従来の研究 [1][2][3] ではニュースの表層的・部分的な特徴から代表的なソーシャルメディア“Twitter”または各国の主要ニュースサイトを対象とした、単一メディアでの投稿数、リツイートやコメントのされ易さの予測が行われていたが、上述の“social news”の裏にある背景と社会に存在するニーズを考慮し、本研究ではニュースから得られる網羅的な言語特性に重点を置き、ソーシャルサイト“Twitter”と日本最大級のポータルサイト Yahoo! JAPAN 内のニュースサイト“Yahoo! ニュース”という複数のメディアを対象として、ユーザの反応に対するニュースコンテンツの影響度をコメントの感情度とコメント数の伸びによって捉えることで、ニュースコンテンツのユーザへの影響を分析・考察する。

具体的には Yahoo!ニュースのトピックスに掲載された記事を対象として、Twitter におけるニュースコンテンツに対するコメント発言（関連ツイート）と Yahoo!ニュース内に設置されたコメント欄のコメントを収集し、ツイートに関してはコメント部分の抽出作業を行った後、ニュース記事に含まれる表層的・言語的・時間的・環境的特徴からそれぞれのプラットフォームのコメント感情度と数の伸びをサポートベクターマシンにより回帰・分類することで、ニュースコンテンツからユーザの反応がどれだけ予測できるかを示す。

以上を通じてニュースコンテンツの言語特性がユーザへ与える影響力を明らかにすることで、将来的な各言語特性の影響研究に役立つ基礎的な知見の導出することを本研究の目的とする。

# 第 2 章

# 関連研究

本章ではこれまでの研究の潮流を 2.1 節で、特に本研究に類似した予測研究について 2.2 節で紹介する。

## 2.1 ソーシャルメディアと影響度分析

2000 年代後半にソーシャル・ネットワーキング・サービスの認知が爆発的に広がると共に、その果たす役割について研究がなされるようになり、2007 年にはそれらが個々人の日常の出来事や情報を共有するために用いられているということが明らかになった [4]。2010 年頃からはソーシャル・ネットワーキング・サービスの情報を共有するソーシャルメディアとしての側面にもスポットライトが当てられ、ソーシャルメディアとニュースとを題材とした研究が多数行われている。

### 2.1.1 ソーシャルメディアとしての Twitter

ソーシャルメディアの中でも Twitter は利用者数、投稿数の面で大きな存在感を示している。Twitter がニュースを拡散させるプラットフォームとして、どのように働いているかについて明らかにしたものとしては、H. Kwak ら [5] が Twitter のトポロジカルな特徴と新しい情報共有メディアとしての潜在的能力を示した研究や、I. Subašić ら [6] が Twitter を利用してニュースを伝える「一般人ジャーナリスト」はニュースを制作しているのか、或いは既にあるニュースを再報道しているのか、という点に着目して行った研究があり、この論文では「一般人ジャーナリスト」の一番の役割は既に報道されているニュースにコメントをつけて投稿することでニュースの拡散を促進することである、と結論付けられている。

他にもニュースワイヤーよりも幅広いニュースを Twitter がカバーしている [7] ことや、Facebook, Twitter, Google Plus というソーシャルメディアの中でも Twitter がニュース速報を最も早く伝えやすいプラットフォームである [8] ことがこれまでに既に示されており、ソーシャルメディアの中でも Twitter が優位性を持っていることがわかっている。

### 2.1.2 ニュースとソーシャルメディアのコミュニティ

ソーシャルメディア上でのニュースの広がり方をコミュニティに焦点を当てて解明しようとする研究もここ数年非常に多く行われている。ソーシャルメディア上でニュースを共有する人々のグループ「ニュースクラウド」に着目したもの [9][10] や、ニュースを収集し拡散させる「ニュースキュレー

ター」となる人に注目したもの [11] が 2013 年には発表されている。中でも D. Saez-Trumper ら [12] は同年オンラインニュースソースやソーシャルメディアコミュニティを取り巻くバイアスについて研究しており、ある人物について言及しているニュース記事と、その人物について言葉を交わしているコミュニティに属するメンバーによるツイートの持つ感情を比較・分析し、ソーシャルメディアでは従来のメディアよりもより主張を伴った文面になり、またよりネガティブな表現になりやすいという結果を示している。

### 2.1.3　コメントの感情推定手法

テキストに含有された感情を推定するという課題は英語では広く研究されてきたが、日本語での研究は多数行われてはいるものの、確立された手法や公開資源の乏しさといった点で英語に比べ遅れをとっている。熊本ら [15] は 2005 年に Web ニュース記事の感情（喜・怒・哀・楽）の程度を「喜ぶ⇔怒る」・「悲しい⇔嬉しい」という２つの感情尺度により評価するシステムを提案している。2010 年にはツイートのデータを元に感情コーパスを自動で作成し、感情分析をする手法 [14] が提案されており、その手法が英語以外の言語でも利用できる可能性を筆者らは示唆しているが、日本語でそれを利用した例はまだ示されていない。

近年では高野ら [16] は感情関連語を用いて記事に含まれる感情（喜・好・安・怒・怖・嫌・悲・恥）を推定法する手法を 2012 年に提案しており、その推定方法により記事に対して付与された感情の約 80% が正解とは限らないものの、間違っていない、という結果を得ている。また内藤ら [17] は同年、感情表現辞典 [18] を用いて 10 の感情（喜・怒・哀・怖・恥・好・嫌・昂・安・驚）について扱った感情語辞典を作成して実験を行い、あるイベントについて取り扱ったニュース記事のみを提示した場合と、ニュース記事と共に感情語を含むコメントを提示した場合に、読み手の持つイベントの印象に違いがあること、特に提示されたコメントがポジティブな場合には回答者の印象もポジティブに、コメントがネガティブな場合には、回答者の印象もネガティブになりやすいことを明らかにしている。

### 2.1.4　ニュース記事に関連するコメントの抽出手法

ソーシャルメディアの影響力が無視できなくなった昨今、従来メディアとソーシャルメディアを繋ぎ、ニュース記事と関連するツイートを収集しようと試みる研究が取り組まれてきた [23][24]。日本でも邱ら [25] がニュース記事とツイートとの内容的類似性・時間的類似性の２つに注目することによって関連コメントを収集する研究を行っている。ニュース記事に関連するツイートには、ニュース記事の見出しや内容を要約したものが含まれている場合が多く、読み手の反応を調べる課題においては記事を読んだユーザが自ら付与した「コメント」を含むツイートを抽出できることが望ましい。A.Kothari ら [26] はツイートが「コメント」付きのツイートであるか否かをサポートベクターマシン (SVM) により分類する研究を行っており、その中で言語的特徴として、一人称や顔文字だけでなく、感情語の有無を用いている。

### 2.1.5　感情を含む言語的特徴を用いた機械学習

ツイートの感情分析を用いた有名な予測課題としては、世論調査 [19] や株価予測 [20] があり、これらは特徴量としての感情の有用性を示している。AM.Popescu らはツイートの品詞（名詞・動詞）の割合や感情を特徴として含め、論争となっている出来事を機械学習により発見する課題 [21] やセ

レブリティに関連する出来事とそれに対する Twitter のユーザのコメントとの関係性を明らかにする課題 [22] に取り組んでいる。

他にも、ニュースに関連するツイートの内、約半数は同じ内容であるとして、興味深いツイートセットを作り出す課題に取り組み、その中で各ツイートの持つ特徴として感情を利用した研究 [27] などがあり、2014 年には過去のツイートに使用されている言葉の特徴を心理言語分析し、インフルエンサーになりそうな人を見つけるという研究 [29] も行われており、言語的特徴が与える影響に注目した研究は 2010 年以降増えつつある。

## 2.2 ソーシャルニュースの分析

M.Tsagkias ら [30] は、Surface, Cumulative, Textual, Semantic, Real-world という 5 つの特徴を用いて、オランダの 7 つのニュース配信社、1 つのニュース情報サイトで配信されたニュースに対して、ニュース配信前にコメントが有るまたは無い、コメントが多いまたは少ないという予測を Random Forest で行っている。Twitter 以外の自国のメディアを対象としているという点で、Yahoo!ニュースを対象に含めている本研究と類似しており興味深いが、言語的特徴としては TF-IDF を用いているだけである。

A.Tatar ら [3] はニュース記事の人気を配信後間もない段階でのコメント数を用いて予測しているが、配信前の段階で予測が可能であることが、記事の書き手にとっては必要不可欠である。ニュースの配信前に、ニュース配信元、カテゴリー、客観性、固有表現（主に人名・地名など）を用いてニュースの人気を予測しようと試みた研究としては R. Bandari ら [1] が行った研究があり、ツイート数の回帰を線形回帰、サポートベクター回帰で、記事に対するツイート数が多い・中程度・少ないという３つに分類するという予測課題を Bagging、決定木、サポートベクターマシン、ナイーブベイズの４種類の手法で行っている。

書き手の工夫次第で変更できる特徴を用いたものとしては、Y. Artzi らによる研究 [2] が挙げられる。ユーザのこれまでの投稿やフォロー・フォロワー数、投稿時刻・曜日に加え、言語的特徴としてバイグラム、ハッシュタグ、感情を用いて返信やリツイートのされ易さを予測している。

言語的特徴を用いた人気予測として、タイトルやニュースの内容が言葉の言い回しが人気に与える影響を、異なる層のコミュニティが集まる掲示板ごとに調査した研究 [28] では、言語的特徴としては品詞の割合や感情、タイトルの長さ等が含まれており、独自に作成したモデルでコメント数を回帰している。

従来の研究においては、コメント数を回帰または分類する人気度予測モデル構築の際には不変的な特徴が主に使用されており、可変的な言語特徴を包括的に使用してきたものは少なかった。更に単一プラットフォームによる実験である場合が多く、影響が見える範囲も限定的であった。またコメントの感情度をニュース配信前に機械学習による回帰で予測しようとする試みはこれまで行われてきておらず、大規模なニュース記事に対するコメントの感情度の調査の例は少ない。

以上より本研究は、

1. 言語的特徴を網羅的に使用する
2. 複数メディアのコメントを研究対象とする
3. コメントの感情度分析を機械学習で行う

という三つの独自性を生かし、ニュース記事の表現がいかに人々の反応に影響を与えるかについて、包括的かつ詳細な分析を通して明らかにすることを試みる。

# 第 3 章

# ソーシャルメディアにおけるニュースの影響度予測

本章ではニュースコンテンツがソーシャルメディア上のユーザの反応へ及ぼす影響度を予測する手法の提案を行う。

## 3.1 テキスト分析に基づくニュースの影響度予測手法

第 1 章で既に述べたように、近年ニュースに対する人々の反応に関わる何らかの知見を求める声は多く、社会的にもそれらの研究の必要性は高い。従来の人気度予測研究からはニュースのカテゴリーや時間といった特徴のみならず、ニュースコンテンツに含まれる固有表現がニュースの人気度に影響を及ぼしていること [1] がわかっており、その他にも感情的な言葉があるほど、読み手に共感を呼ぶこと [17] が研究成果として発表されている。

上記の研究結果を踏まえ、本研究では容易に変更が可能であるという利点を有する言語特性に焦点を当て、固有表現や感情語以外にもニュース記事の文字数や、記号の有無、品詞の割合など言語的な要素を包括的に検討し、これらのコンテンツがユーザの反応に与える影響を分析する。

ユーザの反応分析にはポータルサイト“Yahoo!JAPAN”のニュースページとソーシャルサイト‘Twitter”の 2 種類のメディアから集めたコメントを使用し、サイト間によるコメントの違いという新しい知見を、ジャーナリズムの今後に注目する研究者たちへ提供することを目標とする。

なおニュースコンテンツのユーザの反応への影響度を捉える基準としては、ユーザのコメントの感情度と、ユーザのコメント数の伸びという 2 つを用意した。これらを選択した理由としては感情的な言葉による感情の誘引が既に示されていること [17]、これまでの人気度予測ではツイート数やコメント数が広く用いられていたこと、が挙げられる。

手法のフレームワークは図 3.1 の通りである。

## 3.2 ニュースとコメントの収集および抽出手法

### 3.2.1 ニュース記事の収集

本研究では Yahoo!ニュースにおいてトピックスとして取り上げられた記事を対象とする。

Yahoo!ニュースのトピックスとして扱われる記事はヤフーの編集者によって選択され、その上で新たに記事の内容に見合うおよそ 13 文字の見出しが付与され、Yahoo!ニュースの画面で表示される

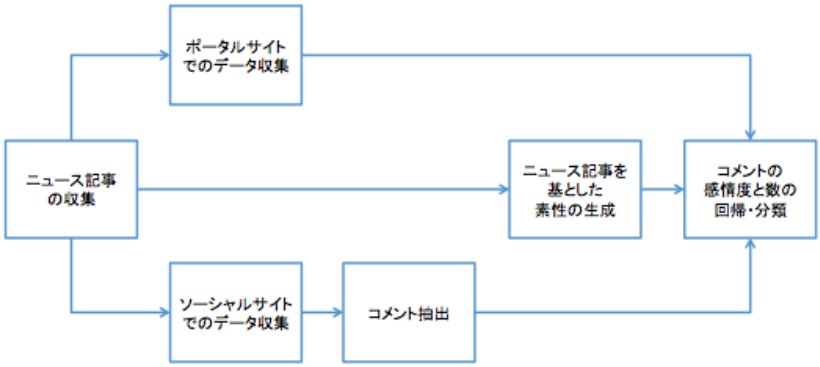



**図** 3.1. 提案手法のフレームワーク

[35] 。Yahoo!ニュースのトップ画面からヤフー側で付けられた見出しをクリックすると、まず元々の記事の見出しと文章の始めの一段落、ヤフーの編集者が作成した関連リンクが表示された画面（図 3.2[1]）へと移る。そしてその画面から [記事全文] というボタンをクリックすることで、また新たなページへと移り、全文が読めるという構造になっている。（図 3.2[2]）以後、これらを明確に区別するために前者のページ（図 3.2[1]）を「トピックス画面」、後者のページ（図 3.2[2]）を「記事全文画面」と呼ぶ。以上の事柄を考慮し、記事のデータとしては、ヤフー側で付与された見出し、関連リンクの見出し、関連リンク名、記事配信社の付与した見出しと記事全文の収集を行う。


[1] トピックス画面


[2] 記事全文画面

**図** 3.2. Yahoo! ニュース画面

ニュース記事のクローリングには Yahoo!ニュースでカテゴリー別に配信されている RSS を利用し、配信された RSS に記述された URL から、該当記事ページの見出し、記事、関連リンクの内容が書かれているコード部分を抜き出すこととする。

### 3.2.2　ポータルサイトでのコメント収集

ポータルサイト Yahoo!JAPAN に設置されている Yahoo!ニュースサイト内で記述されたコメントの取得を該当 URL のコードから抜き出すことによって行う。ただし、Yahoo!ニュースサイト内のコメント欄（図 3.3）は記事全文画面へのリンク先や記事の配信元によっては、コメント欄が設置されない場合があり、すべての記事に対してポータルサイトにおけるコメントが取得できる訳ではない。


**図** 3.3. Yahoo!ニュース コメント欄

### 3.2.3　ソーシャルサイトでのコメント収集

ソーシャルサイト Twitter におけるニュース関連ツイートの収集には、Twitter Search API を使用する。3.2.1 節で述べたように、1 つのニュース記事あたりトピックス画面と記事全文画面の 2 ページがあり、どちらのページ経由でツイートすることも可能であるため、対象とするニュース記事のヤフー作成の見出しまたはトピックス画面へのリンクが含まれているツイート、ニュース記事の配信元が付与した見出しまたは記事全文画面へのリンクが含まれているツイートをそれぞれ関連ツイートとして取得する。2.1.4 節で述べたように、ニュース記事に関連するツイートにおいては記事の内容が反復されている場合や、記事の見出しが含まれている場合が多いため、ユーザが独自に書いた「コメント」部分を抽出することが、ユーザの意見を調べる上で望ましいが、日本語での確立された抽出手法はない。そのため今回は次のようなアルゴリズムを考案する。


[1] 関連ツイートに含まれるコメント例 1


[2] 関連ツイートに含まれるコメント例 2

**図** 3.4. 関連ツイートに含まれるコメント例

**コメント抽出アルゴリズム**

1 ：RT の削除

Twitter Search API で取得したツイートデータにはリツイートデータが含まれており、それらのデータは”RT：” という文字で始まる。よって取得データの先頭の文字列が”RT:” であるものは削除する。

2 ：ニュース配信元の確認

Yahoo!ニュースで収集したニュース記事の配信元以外の、別の配信元の名前が入っているツイートを削除する。ただし配信元のリストは収集したニュース記事から作成する。

3 ：記事の反復部分を削除

ツイートをスペースごとに区切った上で、各部分が記事の一部とマッチするかを確認。マッチした場合は、その部分を削除する。

4 ：全角・半角の統一

ツイートと記事の見出しに含まれる数字・カタカナ・記号の全角・半角をすべて統一する。

5 ：記事の見出し・リンクを削除

記事の見出しやリンク部分はツイートから削除する。

6 ：同一のツイートを削除

スパムを排除するため、一言一句同一のツイートは最初のツイートを除き削除する。

7：複数のツイートで出現する語の削除

ある記事に対する関連ツイートセットの内 10% 以上のツイートで出現した語は削除する。

8：2次メディアの名称やその他ノイズの削除

2次メディアの名称とそれらに関連するワード（Y!ニュース, ヤフートピックス, Yahoo!ニュース, goo ニュース, スマートニュース, d メニュー, Google ニュース検索:, 楽天, 芸能ニュース速報！, ログ速 :)、そして頻繁に出現するノイズ（【拡散】, 【政治】, 【社会】, 【スポーツ】, [経済]）を消去する。

## 3.3　コメントの感情度と数によるニュースの影響度予測手法

### 3.3.1　素性

これまでのニュース記事の人気度予測において R. Bandari ら [1] がカテゴリー・配信元・固有表現・客観性を、M. Tsagkias ら [30] がテキスト中に含まれるリンク数・テキスト（TF-IDF）・時間・同時刻に出版された記事数・ 天候などを使用してきたことを参考に、本研究ではニュースの表層的素性、言語的素性、時間的素性、環境的素性の大きく分けて 4 種類の素性を用いる。

**表層的素性**

ニュース記事の持つ表層的素性として、Yahoo!ニュースのトピックスに掲載された際に付与されるカテゴリーと、ニュースの配信元の名称、そして独自の素性としてニュース記事が掲載されているリンク先の種類（Yahoo!ヘッドライン、 Yahoo!ニュース BUSINESS、 Yahoo!個人、 Yahoo!政治、 Yahoo!選挙、 Yahoo!ニュース:新着雑誌記事、 スポーツナビ、 ネタりか）を用いる。

**言語的素性**

ニュース記事、ニュース見出し、ニュース関連リンクからそれぞれ素性を生成する。ニュース記事では記事のバイグラム（最低文書頻度=2)、トリグラム（最低文書頻度=2)、TF-IDF、文字数、品詞の占める割合、固有表現、客観性、感情語、ポジティブ・ネガティブ・ニュートラルの各感情語の数、感情度を素性として用いた。ニュース見出しからは、Yahoo!トピックス独自の見出しと、記事の見出しの両方から、文字数、記号と数字の出現頻度、ひらがなとカタカナのそれぞれの割合、固有表現、感情語、ポジティブ・ネガティブ・ニュートラルの各感情語の数を、最後に両見出しの類似度を計算し、素性として加える。ニュースの関連リンクからは、リンク先につけられた名前とヤフー編集者が付与したリンク先の要約に用いられた言葉を基に、感情語とポジティブ・ネガティブ・ニュートラルの各感情語の数の素性を生成する。

具体的な素性の計算方法は以下の通りである。

TF-IDF

TF は Term Frequency（単語の出現頻度)、DF は Document Frequency（文書頻度)、 IDF は Inverse Document Frequency(逆文書頻度) の略で、単語 i の文書 d における TF-IDF は以下のよう

に計算される。(ただし D を総文書数とする)

$$TF(i,d) = \frac{n_{i,d}}{\sum_{k \in D} n_{i,k}}$$

$$DF(i) = \sum_{k \in D} \{k : k \ni i\}$$

$$IDF(i) = \log \frac{D}{DF(i)}$$

$$TF \cdot IDF(i,d) = TF(i,d) \times IDF(i)$$

素性生成の際には、最低文書頻度を 2 として計算を行う。

品詞の割合

品詞の判定には形態素解析ソフト MeCab[*1]を使用し、名詞・動詞・形容詞・形容動詞・助詞・助動詞・副詞・接続詞・接頭詞・感動詞・記号・フィラー・その他の 13 種類の品詞が、全形態素の内占める割合をそれぞれ計算する。

固有表現

固有表現の抽出には、日本経済新聞で 2012 年、 2013 年に出現した人名・組織名の固有表現のデータを用いたマッチングを行う。マッチングにより得られた表現を素性、表現の出現頻度を素性値として加える。

客観性

日本語評価極性辞書(用言・体言)[*2]では、主観的・客観的、ポジティブ・ネガティブ・ニュートラルという 2 種類の評価軸で言葉が評価されている。そのため、文書に対してこの辞書による言葉のマッチングを行い、主観語・客観語の数を数える。そして客観性を次のように数値化し、素性値とする。

$$\text{客観性} = \frac{\text{客観語の数}}{\text{主観語の数} + \text{客観語の数}}$$

感情語および感情語の数、感情度

上述の日本語評価極性辞書を用いて、用言・体言の感情語のマッチングを行う。

感情語という素性は、これらの感情表現そのものを素性、その出現頻度を素性値としたものである。それに対し、感情語の数という素性は、文書内に出現したポジティブ・ネガティブ・ニュートラル表現の回数をそれぞれ素性値として使用している。

感情度という素性は、

$$\text{感情度} = \frac{\text{ポジティブ語の数} \times 1.0 + \text{ニュートラル語の数} \times 0.5 + \text{ネガティブ語の数} \times 0}{\text{ポジティブ語の数} + \text{ニュートラル語の数} + \text{ネガティブ語の数}}$$

と定義し、素性値とする。なお上記のようにネガティブ語に対する重みを 0、ポジティブ語に対する重みを 1、ニュートラル語に対する重みは両者の中央値の 0.5 とした重み付けを行うことで、感情が

[*1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[*2] http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources

ネガティブなほど 0 に近く、ポジティブなほど 1 に近い値となるよう数値化している。ただし分類の際には感情度の代わりとして、感情度がニュートラルからどれだけ離れているかの指標である、感情の非中立性を用いることとする。

$$感情の非中立性 = |感情度 - 0.5|$$

類似度計算手法

見出し語の類似度計算には Jaccard 係数を用いる。係数の計算式は以下の通りである。

$$Jaccard\ 係数 = \frac{|X \cap Y|}{X \cup Y}$$

**時間的素性**

ニュース記事が Yahoo!のトピックスとして掲載された時刻、曜日、トップニュースとして扱われた時間、トピックスとして扱われた時間を素性として用いる。曜日に関しては国民の祝日の他に年末年始 (12/29 - 1/3) を祝日として扱うこととする。

**環境的素性**

札幌・東京・名古屋・大阪・福岡の 5 都市における各日の昼の天気概況（06 時〜18 時)、夜の天気概況（18 時〜翌日 06 時)、最高気温、最低気温の情報を気象庁発表の過去気象データから取得し、ニュースが配信された日付における最高気温と最低気温、配信時刻の天候を素性とする。天気は「晴れ」「曇り」「雨」「雪・みぞれ・あられ」の 4 つを素性とし、天気概況の文中に各素性名が含まれる場合、素性値を 1、含まれない場合、素性値を 0 とする。気温についてはデータから得られた日最高気温・日最低気温をそのまま素性値として使用する。

また本研究独自の環境的素性として日経平均株価の記事配信日における終値、終値ー始値、最高値ー最安値の値を素性として採用する。

**表 3.1.** 素性リスト

| 素性 | 定義 | 値 |
|---|---|---|
| 表層的素性 (SF) | | |
| カテゴリー | { 国内, 国際, 地域, 経済, エンターテイメント, スポーツ, コンピューター, サイエンス } | 0-1 |
| 記事の配信元 | ニュース記事の配信元の名称 | 0-1 |
| 記事のリンク先 | {Yahoo!(ヘッドライン, ニュース BUSINESS, 個人, 政治, 選挙, ニュース:新着雑誌記事), スポーツナビ, ネタりか } | 0-1 |
| 言語的素性 (LI) | | |
| 記事 (LI-AR) | | |
| N-gram(Character) | $DF \geq 2$ のバイグラム, トリグラム | 0-1 |
| TF-IDF(Character) | $DF \geq 2$ の TF-IDF 値 | Float |
| 文字数 (Character) | 記事の文字数 | Int |
| 品詞 (Character) | 記事中の各品詞の割合 | Float |
| 固有表現 (NamedEntity) | 記事に含まれる固有表現 (人名・組織名) | 0-1 |
| 客観性 (Sentiment) | 記事が主観的であるか客観的であるかを表した 0 から 1 までの値 | Float |
| 感情語 (Sentiment) | 記事に含まれる感情語 | 0-1 |
| 感情語の数 (Sentiment) | 記事に含まれるポジティブ・ネガティブ・ニュートラルの各感情語の数 | Int |
| 感情度 (Sentiment) | 記事がどれだけ感情的であるかを表した 0 から 1 までの値 | Float |
| 見出し (LI-HE)（Yahoo!トピックスの見出し (LI-HE-y)・記事の見出し (LI-HE-n)） | | |
| 文字数 (Character) | 見出しの文字数 | Int |
| 記号・数字 (Character) | 記号 {!,?, 「, 【, &, =,、, 。, 〈, 『,”, % }、数字の出現回数 | Int |
| ひらがな・カタカナ (Character) | 見出しに含まれるひらがな、カタカナそれぞれの割合 | Float |
| 固有表現 (NamedEntity) | 見出しに含まれる固有表現 (人名・組織名) | 0-1 |
| 感情語 (Sentiment) | 記事に含まれる感情語 | 0-1 |
| 感情語の数 (Sentiment) | 記事に含まれるポジティブ・ネガティブ・ニュートラルの各感情語の数 | Int |
| 見出しの類似度 (-) | Yahoo!トピックスの見出しと記事の見出しの Jaccard 類似度 | Float |
| 関連リンク (LI-RL) | | |
| 感情語 (Sentiment) | 記事に含まれる感情語 | 0-1 |
| 感情語の数 (Sentiment) | 記事に含まれるポジティブ・ネガティブ・ニュートラルの各感情語の数 | Int |
| 時間的素性 (TM) | | |
| 曜日 | Yahoo!ニュースで配信された曜日 { 月, 火, 水, 木, 金, 土, 日, 祝 } | 0-1 |
| 時刻 | Yahoo!ニュースで配信された時刻 {0 - 23} | 0-1 |
| 配信時間 | Yahoo!ニュースのトピックスとして、トップニュースとして扱われた時間 | Int |
| 環境的素性 (EV) | | |
| 天気 | 配信日昼/夜の札幌・東京・名古屋・大阪・福岡の天気 { 晴れ, 曇り, 雨, 雪・みぞれ・あられ } | 0-1 |
| 気温 | 配信日の札幌・東京・名古屋・大阪・福岡の最高気温・最低気温 | Int |
| 日経平均株価 | 配信日の終値, 終値-始値, 最高値-最安値 | Float |

# 第 4 章

# 実験

本章では、先の提案手法の有効性を確認し、各特徴の影響度の大きさを明らかにするため実施した実験の設定について説明する。本章ではその実験について説明する。

## 4.1 実験方法

前章で提案した手法の手順に則り、ニュースデータの収集、ポータルサイトでのコメント収集、ソーシャルサイトでのコメント収集、コメントの抽出、特徴ベクトルの作成、コメントの感情度と数の回帰・分類を行う。

### 4.1.1 実験環境

データ収集から特徴ベクトルの生成までの過程はすべて Python によるプログラムを作成・実行することにより行い、TF-IDF の計算には Python の機械学習ライブラリである scikit-learn[*1]を使用した。コメントの感情度回帰とコメント数の分類に用いるサポートベクターマシンの実装には LIBSVM[*2]を使い、[0,1] の範囲でスケーリングをし、線形カーネルを用いた学習を行った。

### 4.1.2 データ

#### ニュース記事のデータ

本実験においては、ニュース記事として 2014 年 12 月 7 日から 2015 年 1 月 15 日までの 40 日間に、Yahoo!ニュースのトピックスとして扱われた記事 4078 点を収集した。記事の取得漏れを防ぐため、記事データのクローリングは 10 分おきに行った。クローリングの際の条件は次の通りである。

条件 1： 記事全文へのリンク先が SportsNavi の速報画面である場合は、試合の実況を行っており、ユーザのアクセス時間により表示されている文章に大きな差が出てしまう為、記事データをクロールしない。
条件 2： 地震速報についても、記事全文へのリンク先には震源地と各地の震度が示された地図が表示されている画面であり、言語特性が取れないためクロール対象外とする。

[*1] http://scikit-learn.org
[*2] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

### 3.3.2　回帰と分類

事前研究の結果、感情度の回帰手法およびコメント数の伸びの分類手法としてはサポートベクターマシンを用いることとした。サポートベクターマシンは、識別平面と最も近い訓練サンプルとの距離“マージン”を最大化するように、識別平面を設定し、データを２分類するパターン認識手法である。訓練サンプル集合が線形分離不可能である場合には、誤分類をした場合にコストを課すこととし、N 個の訓練サンプルから制約条件

$$\sum_{i=1}^{N} \alpha_i t_i = 0$$

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, ..., N$$

の下で、目的関数

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j t_i t_j x_i^{\mathrm{T}} x_j$$

（$\alpha_i$：ラグランジュ乗数　$x_i$：特徴ベクトル　$t_i$：正解ラベル　$C$：コストパラメータ）

を最大にする２次計画問題を解くことによって、識別平面を決定する。

本研究においては分類だけでなく回帰でもこのサポートベクターマシンを用いることとし、ニュース記事の素性から特徴ベクトルを作成し、Twitter と Yahoo!ニュースのそれぞれプラットフォームに対して、学習を行う。

条件 3：12 月の上旬には選挙に関連して記事全文画面として”Yahoo!みんなの政治” にリンクが貼られた、Yahoo!がこれまでに配信された記事や情報をまとめたページがトピックスが何度か取り上げられたが、これらについても記事データのクロールの対象から外す。

**ポータルサイトのコメントデータ**

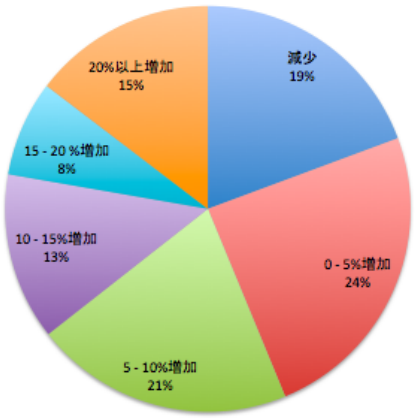
従来研究 [3] において、フランスの日刊紙“20Minutes”のサイトにおいてはニュース記事のコメントはほとんどの場合 1 日で付き終わると示されており、Yahoo!ニュースにおいてもおよそその傾向は見られた。そのため Yahoo!ニュースのコメント欄からコメントが消去されることは稀であることを考慮し、Yahoo!ニュースのトピックスとして記事が掲載されてから 3 日後にはすべてのコメントが付き終わると予想し、このタイミングでコメント欄に記載されている全コメントの取得を行った。

**ソーシャルサイトのコメントデータ**

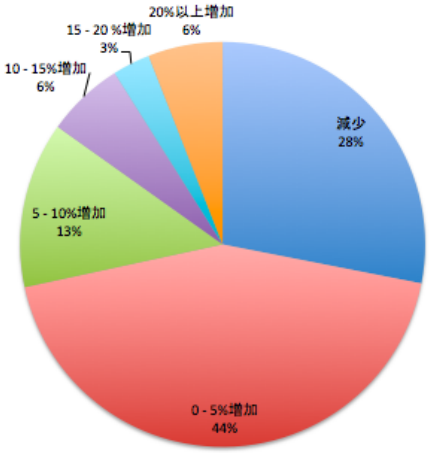
Twitter Search API を使ってニュース記事の配信時刻の 1, 2, 3, 4, 5, 6, 8, 10, 12, 24, 48, 72 時間後のツイートを収集した。12 - 24 時間後、 24 - 48 時間後、48 - 72 時間後のツイート数の増加率を図 4.1 に示す。図 4.1[2] より、48 時間後には 24 時間後に比べ 5% 以上ツイート数が増大するのは記事全体の 28% に過ぎず、それと同じだけの割合の記事でツイート数が減少している。そのため、本実験においては 24 時間後のデータを使用することに決定し、3.2.3 節の通り、コメント抽出を行った。

コメント抽出前のツイートデータと抽出後のデータの例をそれぞれ表 4.1、表 4.2 に示す。

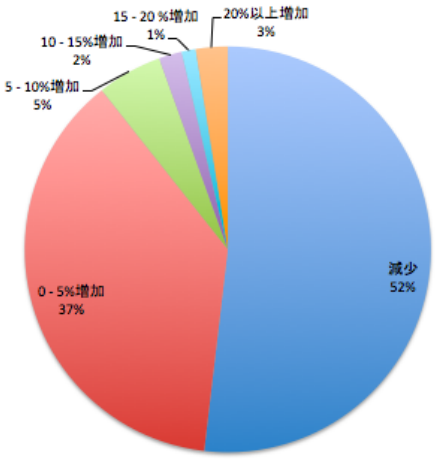
[*3] ただし開始時間のツイート数が 0 である場合は、多くの場合スパムが発生している為データから外した




[1]12 - 24 時間後のツイート増加率




[2]24 - 48 時間後のツイート増加率




[3]48 - 72 時間後のツイート増加率

**図 4.1.** 各時間帯におけるツイートの増加率[*3]

表 4.1. 関連ツイートからのコメント抽出前データ

・東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷 http://t.co/DozX8Mf7TO ようやく販売のめどがついたみたいですね。わたしも記念に 1 枚買おうかな〜。

・東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷 - ITmedia ニュース http://t.co/OuKKLgv1Pi ゼッタイ欲しいわ

・東京駅記念 Suica1 月 30 日から販売受け付け 詳細発表年度内に 10 万枚増刷 http://t.co/NGwwZNokFd 1 月 28 日に駅で申し込み用紙もらって、1 月 30 日に郵送して 払い込み用紙が来たら払って、 3 月 20 日以降に着くのを待つ。 申し込みは一人一回。

・あっ！俺も買う。(笑) RT @kohkuma: 買いまーす w - 東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷（ITmedia ニュース） - Yahoo!ニュース http://t.co/yQbHXfEWfb #fb

・買いまーす w - 東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷（ITmedia ニュース） - Yahoo!ニュース http://t.co/f5rpEEi0Jj #fb

・もう随分前の印象だべ ^^; 東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷（ITmedia ニュース） - Yahoo!ニュース http://t.co/E5RD75zhUV

・【東京駅記念 Suica】《詳細発表》プレミアムは付かなくなりましたが、デザインは素敵ですよ！⇒東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷（ITmedia ニュース） - Yahoo!ニュース http://t.co/gr1YXJBqy9

・東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷 http://t.co/KJxWwyzUxl 所詮はデザインが違うだけのこと。先着だの限定だのと煽るから無用な混乱を招く。限定発行にするなら相応の態勢で準備するなり抽選にするとか工夫が必要だろう!!

・やっぱりミーハーの私としては申し込みだな！→東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷（ITmedia ニュース） - Yahoo!ニュース http://t.co/1w9wwtRcMZ

・あ、ネットでも申し込めるのか！どうにかして、たくさん買う人が出そうですね。 東京駅記念 Suica、1 月 30 日から販売受け付け　詳細発表、年度内に 10 万枚増刷（ITmedia ニュース） - Yahoo!ニュース http://t.co/jM8UR9m1cO

表 4.2. 関連ツイートからのコメント抽出後データ

・ ようやくのめどがついたみたいですね わたしもに 1 買おうかな

・ - ゼッタイ欲しいわ

・Suica130 から に 10 128 にで申し込み用紙もらって、130 に郵送して 払い込み用紙が来たら払って、 320 以降に着くのを待つ 申し込みは一人一回

・あっ!俺も買う (笑)

・買いまーす w - () - #fb

・もう随分前の印象だべ^^; () -

・【Suica】《》プレミアムは付かなくなりましたが、デザインは素敵ですよ!⇒ () -

・ 所詮はデザインが違うだけのこと 先着だの限定だのと煽るから無用な混乱を招く 限定発行にするなら相応の態勢で準備するなり抽選にするとか工夫が必要だろう!!

・やっぱりミーハーの私としては申し込みだな!→ () -

・あ、ネットでも申し込めるのか!どうにかして、たくさん買う人が出そうですね () -

表 4.2 からわかるようにこの例では、「販売」や「記念」という言葉がアルゴリズム 7 により除去対象となっているため、いくつかのコメントからはこれらの単語が消え、完全なコメントを抽出できている訳ではない。しかしながら、感情や意見を表すコメント部分は除去されることなく抽出できた。

### 4.1.3 素性の組み合わせ

以下の実験 1 - 実験 3 に登場する素性の組み合わせで特徴ベクトルを作成し実験を進める。

**実験 1**

まずニュースの記事の表層的特徴 (SF)、言語的特徴 LI)、時間的特徴 (TM)、環境的特徴 (EV) の影響を調べるため以下の実験 1 を行う。(省略文字の表す具体的な特徴については表 3.1 を参照)

Run 1-1: SF
Run 1-2: LI
Run 1-3: TM
Run 1-4: SF + LI
Run 1-5: SF + TM
Run 1-6: LI + TM
Run 1-7: SF + LI + TM
Run 1-8: SF + LI + TM + EV

**実験 2**

言語的特徴 (LI) に含まれる素性数が他の特徴より多いため、言語的特徴を記事 (LI-AR) と見出し (LI-HE)、関連リンク (LI-RL) という特徴セットに分け、以下の実験 2 を行う。(省略文字の表す具体的な特徴については表 3.1 を参照)

Run 2-1: LI-AR
Run 2-2: LI-HE-y
Run 2-3: LI-HE-n
Run 2-4: LI-HE
Run 2-5: LI-AR + LI-HE

Run 1-2*: LI-AR + LI-HE + LI-RL

**実験３**

言語的特徴（LI）に含まれる素性数が他の特徴より多いため、言語的特徴を文字（LI-Character)、固有表現（LI-NamedEntity)、 感情（LI-Sentiment）という特徴セットに分け、以下の実験 3 を行う。(省略文字の表す具体的な特徴については表 3.1 を参照)

Run 3-1: LI-Character
Run 3-2: LI-NamedEntity
Run 3-3: LI-Sentiment

各特徴セットに含まれる特徴数のオーダーは表 4.3 の通りである。

**表** 4.3. 特徴セットと特徴数のオーダー

| 特徴セット | オーダー |
|---|---|
| SF | 100 |
| LI-AR | 100,000 |
| LI-HE | 1,000 |
| LI-RL | 1,000 |
| LI-Character | 100,000 |
| LI-NamedEntity | 10,000 |
| LI-Sentiment | 10,000 |
| TM | 10 |
| EV | 10 |

### 4.1.4　回帰と分類

**コメントの感情度回帰**

ニュース記事全文の言語的素性を用いて特徴ベクトルを作成していることを考慮し、トピックス画面ではなく記事全文画面の見出しまたは URL を含むツイートを対象とし、学習を行った。コメントの感情度は各コメントについて日本語評価極性辞書とのマッチングを行い、コメント数 N のニュース記事における感情度を

$$\text{コメントの感情度} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{コメント } i \text{ 中のポジティブ語の数}\times 1.0 + \text{コメント } i \text{ 中のニュートラル語の数}\times 0.5}{\text{コメント } i \text{ 中の全感情語の数}}$$

と定義した。ただしコメントが非常に少ない場合には、辞書に感情語が十分にとれないため、コメントが 5 件以上あるものを対象として実験を行った。

**コメント数の分類**

コメント数の伸びに関する２値分類を行うにあたって、境界値の設定を行った。本実験で収集したコメントデータから、各メディアでコメント数が全記事のおよそ上位 50%、上位 10% になるように２つの値を決定した。なおコメント数の分類では、トピックス画面の見出しまたは URL を含むコメントツイートと記事全文画面の見出しまたは URL を含むコメントツイートの和をコメント数としている。結果としてコメントツイート数では、コメント数が 100 以上または 100 未満・300 以上または 300 未満、Yahoo!ニュース内コメント数では、コメント数が 200 以上または 200 未満・700 以上または 700 未満という、分類問題に取り組むことにした。

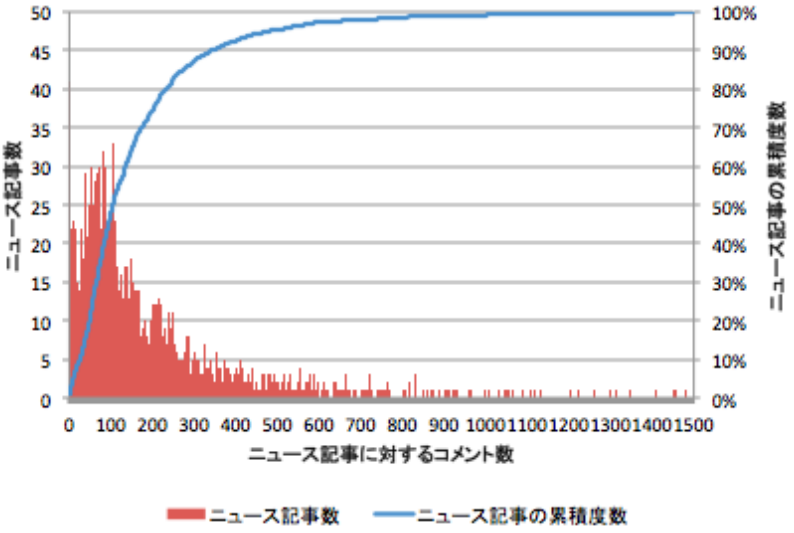



**図** 4.2. ニュース記事に対するコメントツイート数分布

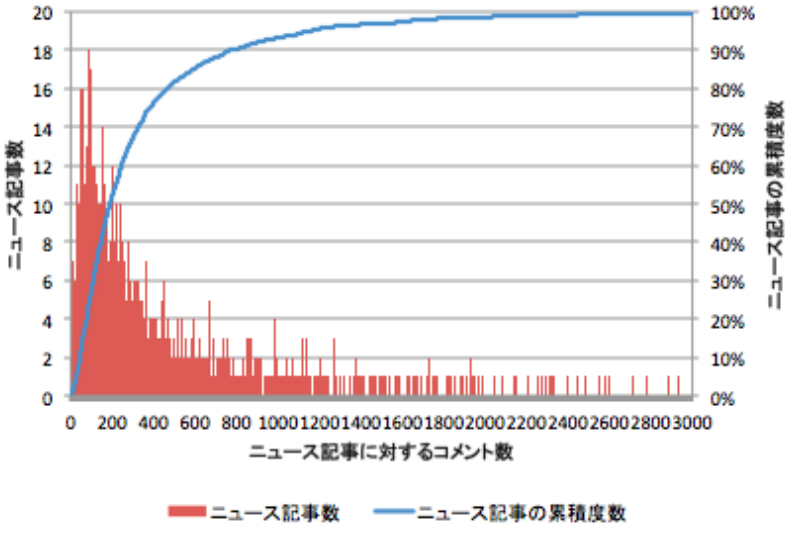



**図** 4.3. ニュース記事に対する Yahoo!ニュース内コメント数分布

## 4.2　評価方法

**K-分割交差検定**

K-分割交差検定は、データを K 個のグループに均等に分割し、K-1 個のグループを訓練用データ、残り 1 個のグループを検証用データとして K 回の学習を行い、それぞれの結果の評価値を平均することで予測の精度を図ろうとする検定方法である。本実験では、K = 10 の交差検定を行った。

**評価指標（回帰）**

回帰の評価指標としては、あてはまりの良さの尺度である決定係数（$R^2$）があり、この値は平均二乗誤差（MSE）と分散（VAR）から次のように計算される。

$$R^2 = 1 - \frac{MSE}{VAR}$$

**評価指標（分類）**

機械学習により 2 値分類を行った結果は表 4.4 のようにまとめることができる。正答率、 適合率、

**表 4.4.** 分類結果の分割表

| 予測結果＼真の結果 | 正 | 負 |
|---|---|---|
| 正 | 真陽性（TP:True Positive） | 偽陽性（FP:False Positive） |
| 負 | 偽陰性（FN:False Negative） | 真陰性（TN:True Negative） |

再現率は以下のように定義されており、評価基準として用いられる。

$$\text{正答率} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{適合率} = \frac{TP}{TP + FP}$$

$$\text{再現率} = \frac{TP}{TP + FN}$$

適合率と再現率は一般的に一方が高くなればもう一方が低くなる、という関係性にあるため、適合率と再現率を統合した指標 F 値がある。

$$F \text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

本実験においては分類学習器の評価には正答率と F 値を使用する。

# 第 5 章

# 結果と考察

本章では前章の実験から得られた結果とその結果に基づく考察を記す。

## 5.1 収集したニュースとコメント

### 5.1.1 表層的素性の統計量

収集した記事のカテゴリーと配信元の内訳を表 5.1、表 5.2 に示す。コンピューター・サイエンスの 2 つのカテゴリーは全体の 5% ずつと、他のカテゴリーよりも記事数が少ない。配信元に関しては、一般紙・スポーツ紙が比較的多く、地方紙は少ない。

**表 5.1.** ニュース記事内訳（カテゴリー）

| カテゴリー | ニュース記事数 | Yahoo!ニュース内にコメント欄が設置された記事数 |
|---|---|---|
| 国内 | 635 (16%) | 355 (13%) |
| 国際 | 566 (14%) | 412 (15%) |
| 地域 | 704 (17%) | 264 (10%) |
| 経済 | 459 (11%) | 263 (10%) |
| エンターテイメント | 602 (15%) | 546 (20%) |
| スポーツ | 714 (18%) | 615 (23%) |
| コンピューター | 193 ( 5%) | 134 ( 5%) |
| サイエンス | 205 ( 5%) | 107 ( 4%) |
| 総計 | 4078 (100%) | 2696 (100%) |

**表 5.2.** ニュース記事内訳（配信元）

| 配信元 | ニュース記事数 | Yahoo!ニュース内にコメント欄が設置された記事数 |
|---|---|---|
| 時事通信 | 506 (12%) | 483 (18%) |
| 毎日新聞 | 463 (11%) | 217 (8%) |
| 産経新聞 | 326 ( 8%) | 326 (12%) |
| スポニチアネックス | 282 ( 7%) | 282 (10%) |
| 朝日新聞デジタル | 224 ( 5%) | 0 ( 0%) |
| 読売新聞 | 184 ( 5%) | 0 ( 0%) |
| デイリースポーツ | 168 ( 4%) | 168 ( 6%) |
| ＡＦＰ＝時事 | 129 ( 3%) | 125 ( 5%) |
| オリコン | 117 ( 3%) | 117 ( 4%) |
| スポーツ報知 | 104 ( 3%) | 99 ( 4%) |
| TBS 系（JNN） | 92 ( 2%) | 0 ( 0%) |
| サンケイスポーツ | 84 ( 2%) | 84 ( 3%) |
| 日刊スポーツ | 82 ( 2%) | 82 ( 3%) |
| 日本テレビ系（NNN） | 73 ( 2%) | 0 ( 0%) |
| テレビ朝日系（ANN） | 72 ( 2%) | 0 ( 0%) |
| フジテレビ系（FNN） | 71 ( 2%) | 0 ( 0%) |
| THE PAGE | 64 ( 2%) | 60 ( 2%) |
| ロイター | 57 ( 1%) | 50 ( 2%) |
| ウェザーマップ | 48 ( 1%) | 48 ( 2%) |
| SankeiBiz | 37 ( 1%) | 18 ( 1%) |
| まんたんウェブ | 33 ( 1%) | 33 ( 1%) |
| 東スポ Web | 32 ( 1%) | 31 ( 1%) |
| 東洋経済オンライン | 31 ( 1%) | 0 ( 0%) |
| SOCCER KING | 28 ( 1%) | 28 ( 1%) |
| 神戸新聞 NEXT | 28 ( 1%) | 0 ( 0%) |
| CNN.co.jp | 26 ( 1%) | 26 ( 1%) |
| レスキューナウニュース | 25 ( 1%) | 25 ( 1%) |
| ITmedia ニュース | 22 ( 1%) | 22 ( 8%) |
| NEWS ポストセブン | 22 ( 1%) | 0 ( 0%) |
| Sportsnavi | 22 ( 1%) | 0 ( 0%) |
| ゲキサカ | 22 ( 1%) | 22 ( 8%) |
| フットボールチャンネル | 22 ( 1%) | 0 ( 0%) |
| CNET Japan | 21 ( 1%) | 15 ( 1%) |
| その他 | 561 (14%) | 331 (12%) |
| 総計 | 4078 (100%) | 2696 (100%) |

更にカテゴリー別のコメント数に対する記事の累積度数を図 5.1、図 5.2 に示す。これらの図からは、Twitter の場合も Yahoo!ニュースコメント欄の場合もエンターテイメントカテゴリーに属する記事のコメント数が他のカテゴリーよりも比較的伸び易いこと、国内カテゴリーの記事が特に Yahoo!ニュースコメント欄でコメントがつき易いこと、コンピューターカテゴリーは Twitter では全カテゴリーの中でコメント数が比較的多いが、Yahoo!ニュースコメント欄では比較的少なく特にコメント数が 400 を超える記事の割合は全カテゴリー中最も少ないことが読み取れる。

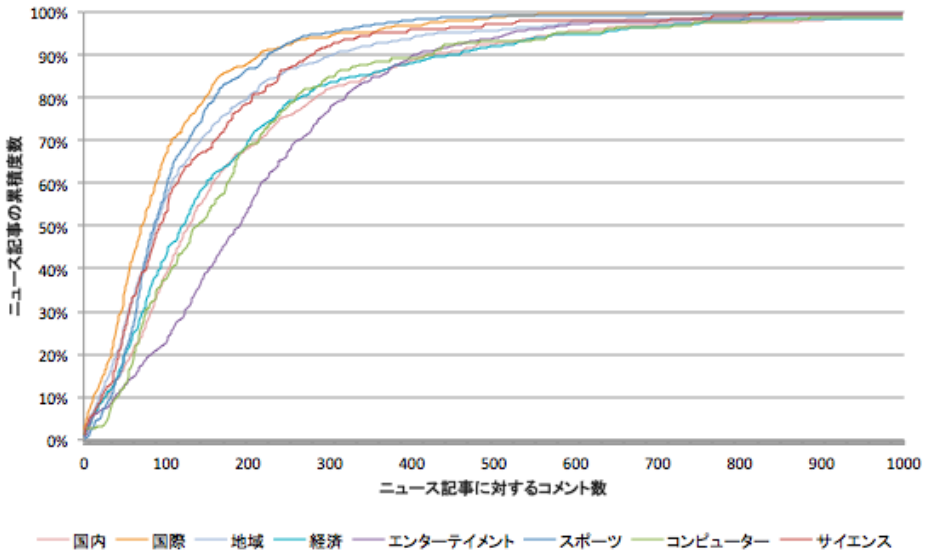



**図** 5.1. カテゴリー別コメントツイート数に対するニュース記事の累積度数

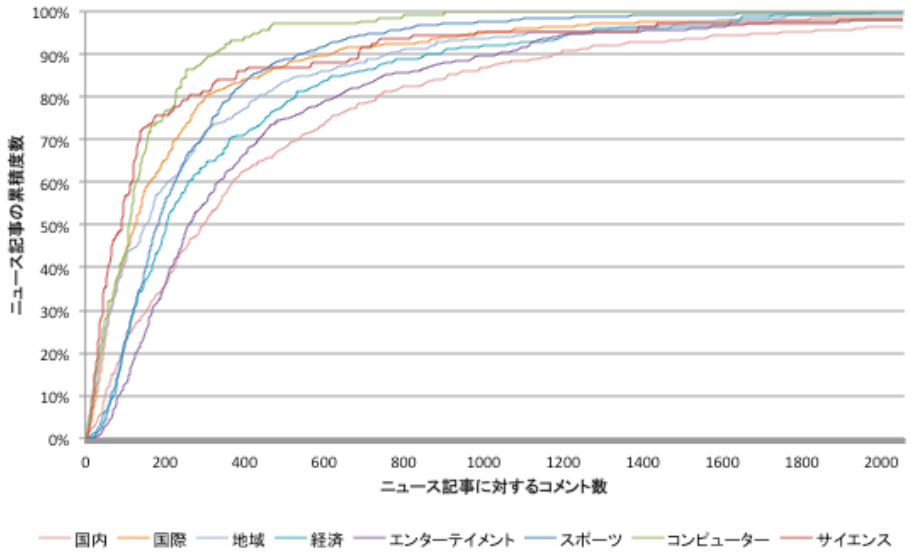



**図** 5.2. カテゴリー別 Yahoo!ニュース内コメント数に対するニュース記事の累積度数

### 5.1.2　言語的素性の統計量

ある感情度あたりのニュース記事数をまとめたヒストグラムを図 5.3 に示す。本実験ではニュース記事は感情度が 0.4 - 0.7 であるものが、全体の 69% を占め、特にニュートラルな感情（感情度：0.5）よりも少しポジティブな感情をもつ記事が Yahoo!ニュースのトピックスとして扱われる記事の中では多いことがわかる。

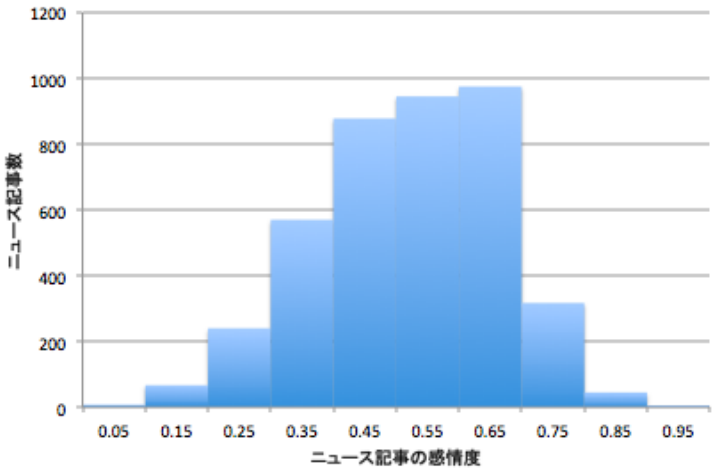



**図** 5.3. 感情度あたりのニュース記事数

次に、感情的な言葉がある程、読者が共感し易い [17] というこれまでの研究結果について、本実験でも同様の結果が見られるか調査を行った。その結果ニュース記事の感情度とコメントツイートの感情度との相関係数は 0.59 であり、ニュース記事の感情にコメントを付ける読者も同じような感情を持ち易い傾向があるようであった。（図 5.4）またニュース記事の感情度と Yahoo!ニュース内のコメントの感情度との相関を調べると、相関係数は 0.71 と、コメントツイートよりも更に相関が高く、Twitter にコメントを書き込む人よりも、Yahoo!ニュース内にコメントを書き込む人の方がニュース記事の感情により近い感情を抱き易いということがわかった（図 5.5）。

さらにニュース記事の感情度とコメント数との関係を見ると、Twitter・Yahoo!ニュース内のコメント欄共に、ニュース記事の感情が強いポジティブまたは強いネガティブを示すほどコメント数の伸びは悪く、ニュース記事の感情がニュートラルに近いほど、コメント数が伸び易いという傾向が見える (図 5.6、図 5.7)。

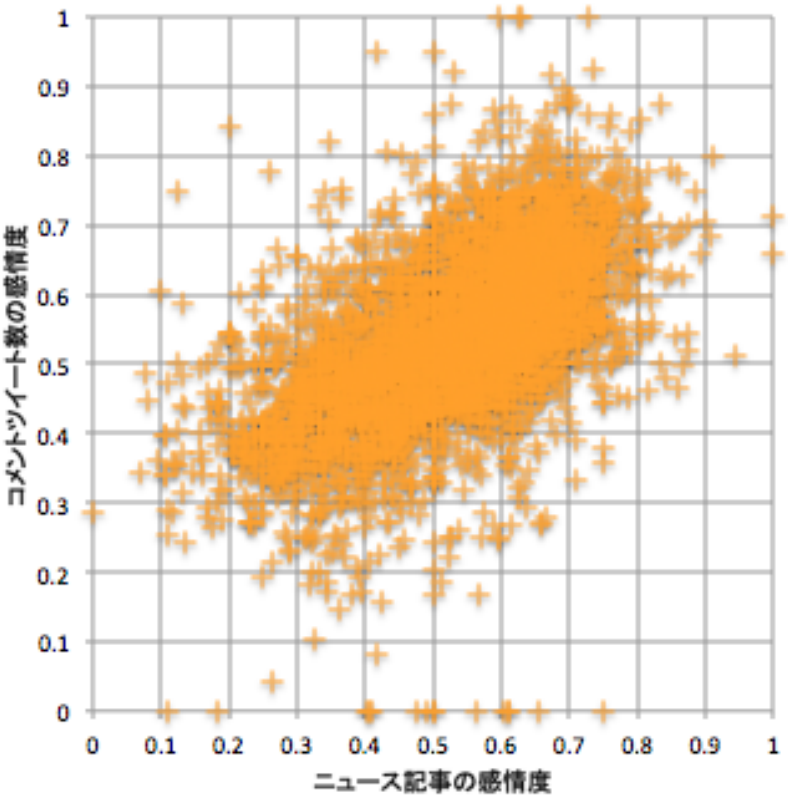



**図 5.4.** ニュース記事の感情度とコメントツイートの感情度の関係

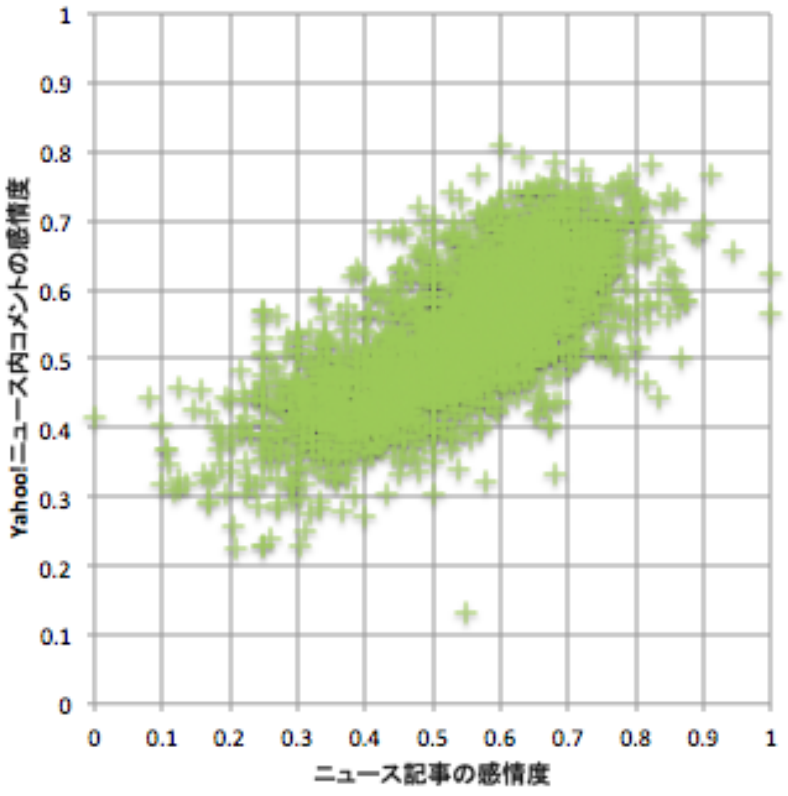



**図 5.5.** ニュース記事の感情度と Yahoo!ニュース内コメントの感情度の関係

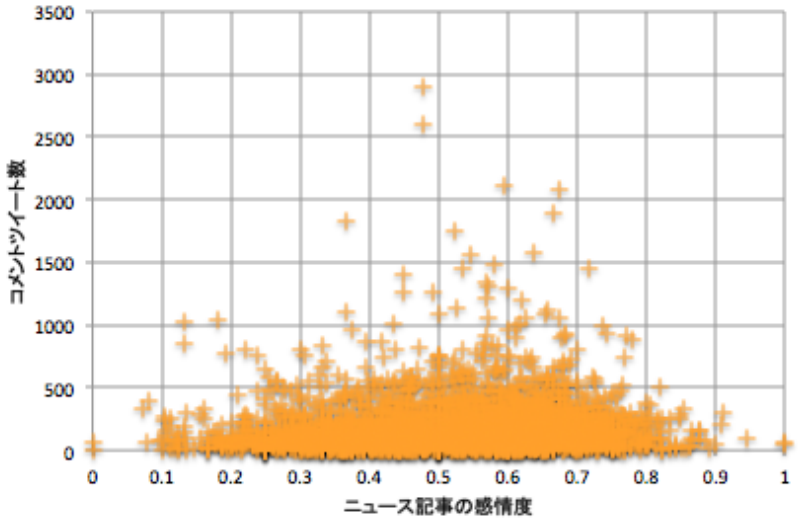



**図** 5.6. ニュース記事の感情度とコメントツイートの数の関係

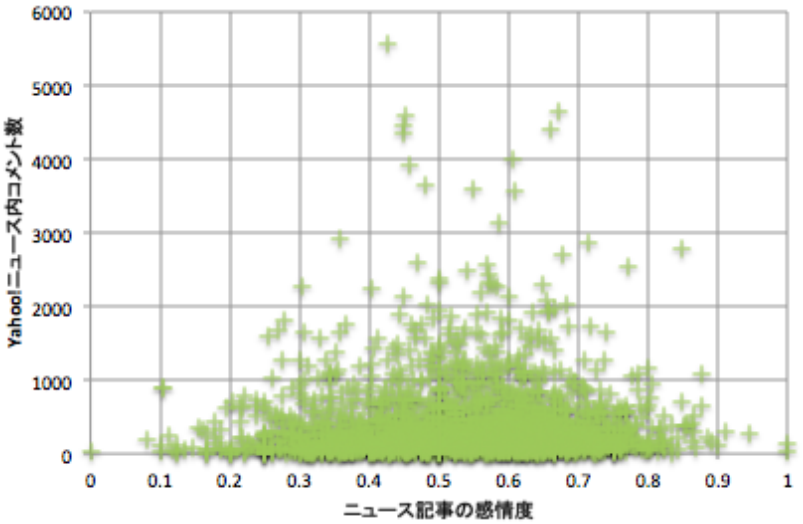



**図** 5.7. ニュースの感情度と Yahoo!ニュース内コメントの数の関係

### 5.1.3 時間的素性の統計量

従来の研究によりニュース記事の配信数は平日に比べ休日には少なくなることが示されている [1] 為、Yahoo!ニュースでトピックスとして扱われる記事数も同様の傾向があるかを調べた。その結果、休日は該当週の平日に比べて少なくなりやすいこと、そして年末年始と国民の祝日も休日同様に記事数が少なくなることが示された (図 5.8、図 5.9)。週明けよりも週末が近いほど記事が増え易いという傾向が少しあるようだが、特にはっきりとした傾向は平日の曜日別記事数には見られなかった (図 5.9)。

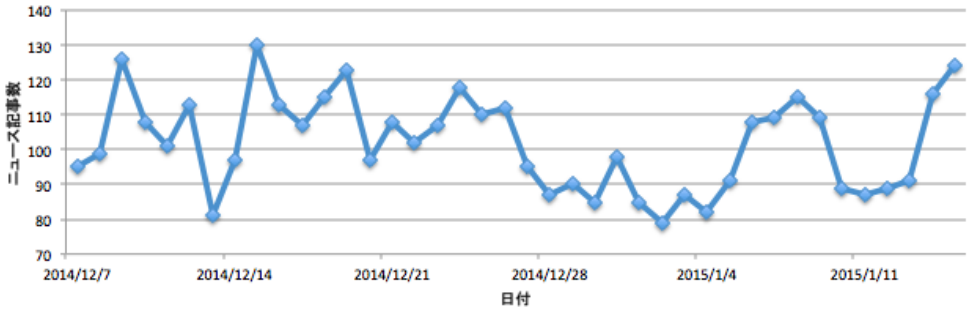



**図** 5.8. ニュース記事の数

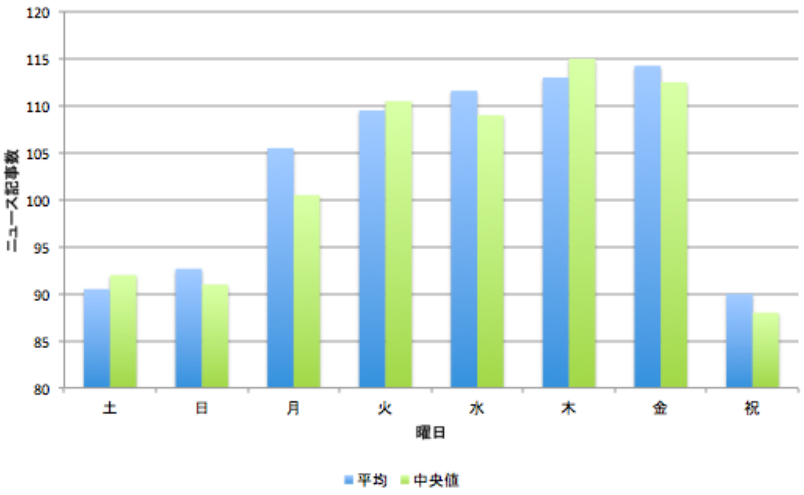



**図** 5.9. ニュース記事数の平均・中央値（曜日別）

## 5.2　感情度の回帰によるニュース影響度予測

### 5.2.1　関連コメントツイートの感情度回帰結果

まず表層的特徴 (SF)、言語的特徴 (LI)、時間的特徴 (TM)、環境的特徴 (EV) のいずれがコメントの感情度に影響を与えているかを調べ、結果として言語的特徴 (LI) がコメントツイートの感情度に与える影響が他の 3 特徴に比べ大きく、言語的特徴 (LI) のみで評価値は平均二乗誤差 0.0084、決定係数 0.448 となった。また時間的特徴 (TM) 及び環境的特徴 (EV) の影響は非常に小さいことがわかった (表 5.3)。

**表 5.3.** 関連コメントツイートの感情度回帰：実験 1

| 特徴 | 平均二乗誤差 (MSE) | 決定係数 ($R^2$) |
|---|---|---|
| SF | 0.0110 | 0.273 |
| LI | 0.0084 | 0.448 |
| TM | 0.0151 | 0.004 |
| SF + LI | 0.0083 | 0.449 |
| SF + TM | 0.0110 | 0.273 |
| LI + TM | 0.0084 | 0.447 |
| SF + LI + TM | 0.0083 | 0.448 |
| SF + LI + TM + EV | 0.0084 | 0.446 |

次に言語的特徴 (LI) の中でも記事 (LI-AR)、見出し (LI-HE)、関連リンク (LI-RL) のいずれの特徴に影響力があるのかを調べたところ、記事中の言語的特徴 (LI-AR) が与える影響が最も大きいが、記事より文字量が著しく少ない見出し (LI-HE) のみでも、平均二乗誤差 0.0105、決定係数 0.320 という結果を得られることがわかった。関連リンクについては、この特徴を加えることにより感情度の決定係数はやや上がるようである (表 5.4)。

**表 5.4.** 関連コメントツイートの感情度回帰：実験 2

| 特徴 | 平均二乗誤差 (MSE) | 決定係数 ($R^2$) |
|---|---|---|
| LI-AR | 0.0087 | 0.424 |
| LI-HE-y | 0.0114 | 0.249 |
| LI-HE-n | 0.0106 | 0.297 |
| LI-HE | 0.0105 | 0.320 |
| LI-AR + LI-HE | 0.0085 | 0.434 |
| LI-AR + LI-HE + LI-RL* | 0.0084 | 0.448 |

最後に言語的特徴 (LI) の中でも、文字（LI-Character)、 固有表現（LI-NamedEntity)、 感情 (LI-Sentiment）のいずれが重要であるかを調べ、やはりコメントツイートの感情度はニュース記事の感情特徴により回帰する場合が最も良い結果が得られることがわかった。しかしながら、文字特徴 (LI-Character) のみでも、平均二乗誤差では感情特徴 (Sentiment) を使った場合とほぼ同等の値が

得られている（表 5.5)。

**表 5.5**. 関連コメントツイートの感情度回帰：実験 3

| 特徴 | 平均二乗誤差 (MSE) | 決定係数 ($R^2$) |
|---|---|---|
| LI-Character | 0.0090 | 0.406 |
| LI-NamedEntity | 0.0108 | 0.285 |
| LI-Sentiment | 0.0089 | 0.421 |

### 5.2.2 Yahoo!ニュース内コメントの感情度回帰結果

Yahoo!ニュース内のコメントの感情度を上述のコメントツイートの感情度同様回帰した結果、コメントツイートの場合と同じく言語的特徴 (LI) が表層的特徴 (SF)、時間的特徴 (TM)、環境的特徴 (EV) に比べ大きくコメントの感情度に影響を与えていることがわかった。表層的特徴 (SF) も素性数の少なさを考慮すれば有効であるが、時間的特徴 (TM) と環境的特徴 (EV) の影響は小さいと言った点でコメントツイートの場合と傾向は非常に似ているが、どの特徴を使った場合においても決定係数はコメントツイートの場合よりも上がっている (表 5.6)。

**表 5.6**. Yahoo!ニュース内コメントの感情度回帰：実験 1

| 特徴 | 平均二乗誤差 (MSE) | 決定係数 ($R^2$) |
|---|---|---|
| SF | 0.0058 | 0.398 |
| LI | 0.0030 | 0.690 |
| TM | 0.0097 | 0.008 |
| SF + LI | 0.0030 | 0.693 |
| SF + TM | 0.0058 | 0.399 |
| LI + TM | 0.0030 | 0.689 |
| SF + LI + TM | 0.0030 | 0.692 |
| SF + LI + TM + EV | 0.0030 | 0.691 |

次に実験 2 によりニュースの記事と見出し、関連リンクのそれぞれの影響力を見ると、やはり記事の影響力が他の 2 つを大きく上回っているが、ニュースの見出しのみでも平均二乗誤差 0.0050、決定係数 0.486 という評価値で回帰を行うことができている（表 5.7)。

最後に言語的特徴 (LI) の中でも文字 (LI-Character) と固有表現 (LI-NamedEntity)、感情 (LI-Sentiment) のいずれがコメントの感情度に影響を強く与えているかを見ると、コメントツイートの場合と同じように感情 (LI-Sentiment) を特徴として用いた場合の評価値が最も高いが、文字 (LI-Character) 特徴のみでも、平均二乗誤差 0.0037、決定係数 0.616 という値を得ている（表 5.8)。

### 5.2.3 関連コメントの感情度回帰に関する考察

関連コメントの感情度回帰結果より、感情度に最も影響を与えている特徴は言語的特徴であり、記事の特徴が見出し・関連リンクの特徴に比べ影響が大きく、感情の特徴が文字・固有表現の特徴よ

**表 5.7.** Yahoo!ニュース内コメントの感情度回帰：実験 2

| 特徴 | 平均二乗誤差 (MSE) | 決定係数 ($R^2$) |
|---|---|---|
| LI-AR | 0.0034 | 0.646 |
| LI-HE-y | 0.0054 | 0.439 |
| LI-HE-n | 0.0050 | 0.486 |
| LI-HE | 0.0044 | 0.544 |
| LI-AR + LI-HE | 0.0031 | 0.684 |
| LI-AR + LI-HE + LI-RL* | 0.0030 | 0.690 |

**表 5.8.** Yahoo!ニュース内コメントの感情度回帰：実験 3

| 特徴 | 平均二乗誤差 (MSE) | 決定係数 ($R^2$) |
|---|---|---|
| LI-Character | 0.0037 | 0.616 |
| LI-NamedEntity | 0.0051 | 0.477 |
| LI-Sentiment | 0.0036 | 0.640 |

りも優れているが、文字特徴を使った場合でも感情特徴を用いた場合と比べ平均二乗誤差の違いは 0.0001 に留まることが示された。文字特徴のセット中に TF-IDF が入っていることにより、感情特徴同様に感情的な単語が特徴として含まれている為、それらが影響しこのような結果を生んでいる可能性がある。

Twitter と Yahoo!ニュース内のコメント欄に書き込まれたコメントを比較すると後者の感情度の方が予測し易いという結果になったが、5.1.2 節において、コメントツイートの感情度よりも Yahoo!ニュース内のコメントの感情度の方がニュース記事のもつ感情度との正の相関が高かったことも合わせて考えると、コメントを書き込むという行為自体が記事に対する強い共感を抱いている場合に起こり易い、という可能性も十分にあるため断言はできないが、Yahoo!ニュース内でコメントを書き込む場合の方がニュース記事に対する共感が起こり易い為に、感情度の予測が容易なのではないかと考えられる。

ただしコメントを書き込むという行為自体が記事に対する強い同意を抱いている場合に起こり易い、という可能性も十分にあるため、ニュース記事に共感する人が多いとは断言できない。

時間や天気・気温・株価などの特徴はコメントの感情度にはほとんど影響を与えないという結果となった為、実社会で置かれている状況よりも、読んだニュース記事に含まれている特徴から強く影響を受けコメントを書き込んでいると言える。

## 5.3　コメント数の分類によるニュース影響度予測

### 5.3.1　関連コメントツイート数の分類結果

関連コメントツイート数の分類を境界値 100 コメント、300 コメントの場合について行った結果、表層的特徴 (SF)・言語的特徴 (LI)・時間的特徴 (TM)・環境的特徴 (EV) の内、言語的特徴 (LI) が単独では最も影響力があるが、時間的特徴 (TM) のみを用いた場合を除き、どの素性組み合わせにお

いても F 値 0.65 - 0.70、正答率も 60% 後半 - 70% 程度であった。感情度に続きこの分類でも時間的特徴 (TM) と環境的特徴 (EV) は、素性に加えることにより評価値が下がることもあるなど、影響はほとんどないという結果となった。境界値 100 の場合は言語的特徴 (LI) のみの場合と、表層的特徴 (SF) と言語的特徴 (LI) を組み合わせた場合との結果に大きな差は見られないが、境界値 300 の場合は表層的特徴 (SF) と言語的特徴 (LI) を組み合わせた場合、学習結果が改良されている（表 5.9)。

**表 5.9.** 関連コメントツイート数の分類：実験 1

| 特徴 | コメント数 $\geq 100$ | | | コメント数 $\geq 300$ | | |
|---|---|---|---|---|---|---|
| | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| SF | 0.652 | 0.632 | 64.3% | 0.656 | 0.625 | 64.1% |
| LI | 0.674 | 0.667 | 67.1% | 0.666 | 0.659 | 66.2% |
| TM | 0.433 | 0.632 | 55.3% | 0.491 | 0.538 | 51.6% |
| SF + LI | 0.675 | 0.690 | 68.3% | 0.707 | 0.700 | 70.3% |
| SF + TM | 0.652 | 0.632 | 64.3% | 0.629 | 0.624 | 62.6% |
| LI + TM | 0.675 | 0.663 | 66.9% | 0.675 | 0.657 | 66.6% |
| SF + LI + TM | 0.675 | 0.691 | 68.3% | 0.702 | 0.694 | 69.8% |
| SF + LI + TM + EV | 0.674 | 0.691 | 68.3% | 0.702 | 0.697 | 69.9% |

実験２からは記事 (LI-AR) と見出し (LI-HE)、関連リンク (LI-RL) の影響度を見ると（表 5.10）と、各々の特徴を単独で用いた場合、記事 (LI-AR) が F 値・正答率ともに最も高くなること、またヤフー作成の見出し (LI-HE-y) とニュース記事の見出し (LI-HE-n) ではニュース記事の見出し (LI-HE-n) を特徴として用いた場合の方が、特に境界値 300 の場合に評価値が高く、コメント数の伸びへの影響が高いこと、またどちらの境界値の場合もヤフーの見出し（LI-HE-y) を特徴として加えても、ニュース記事の見出し（LI-HE-n）のみの場合と結果に差がほとんどないことがわかる。境界値の違いによる結果の差としては、分類境界値を 100 とした場合は記事 (LI-AR) の他に見出し (LI-HE)・関連リンク (LI-RL) を特徴に加えることで F 値、正答率が改善されているのに対して、分類境界値を 300 とした場合は記事 (LI-AR) の他の特徴 (LI-HE, LI-RL) を加えても改善は見られないことが挙げられる。

**表 5.10.** 関連コメントツイート数の分類：実験２

| 特徴 | コメント数 $\geq 100$ | | | コメント数 $\geq 300$ | | |
|---|---|---|---|---|---|---|
| | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| LI-AR | 0.647 | 0.640 | 64.4% | 0.663 | 0.662 | 66.2% |
| LI-HE-y | 0.632 | 0.504 | 57.8% | 0.594 | 0.492 | 54.9% |
| LI-HE-n | 0.632 | 0.593 | 61.3% | 0.616 | 0.583 | 60.0% |
| LI-HE | 0.625 | 0.591 | 60.9% | 0.622 | 0.583 | 60.3% |
| LI-AR + LI-HE | 0.669 | 0.665 | 66.7% | 0.663 | 0.656 | 66.0% |
| LI-AR + LI-HE + LI-RL* | 0.674 | 0.667 | 67.1% | 0.666 | 0.659 | 66.2% |

最後に文字特徴 (LI-Character)・固有表現特徴 (LI-NamedEntity)・感情特徴 (LI-Sentiment) の

内では文字特徴（LI-Character）がコメントの伸びに最も寄与していることがわかり、境界値 100 の場合では全言語特徴 (LI) を用いた学習結果とほぼ同等の結果が得られた。しかし境界値 300 の場合は文字（LI-Character）特徴が他の２つよりも評価値は最も高いものの、文字 (LI-Character)・固有表現 (LI-NamedEntity)・感情特徴 (LI-Sentiment) のいずれの場合も結果はほぼ変わらなかった (表 5.11)。

**表** 5.11. 関連コメントツイート数の分類：実験 3

| 特徴 | コメント数 $\geq 100$ | | | コメント数 $\geq 300$ | | |
|---|---|---|---|---|---|---|
| | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| LI-Character | 0.670 | 0.668 | 66.9% | 0.660 | 0.650 | 65.5% |
| LI-NamedEntity | 0.663 | 0.612 | 63.9% | 0.638 | 0.646 | 64.2% |
| LI-Sentiment | 0.636 | 0.610 | 62.3% | 0.665 | 0.614 | 64.1% |

### 5.3.2　Yahoo!ニュース内コメント数の分類結果

ここからは Yahoo!ニュース内コメント数の分類を境界値 200、700 で行った結果を示す。実験 1 からコメントツイート数の分類と同様、表面的特徴 (SF)・言語的特徴 (LI)・時間的特徴 (TM)・環境的特徴 (EV) の内では言語的特徴 (LI) がコメント数の伸びに最も影響が大きく、時間的特徴 (TM) と環境的特徴 (EV) の影響が非常に小さいことが読み取れるが、両者を比べるどちらの境界値の場合も学習結果の F 値・正答率が Yahoo!ニュースコメント数の分類の場合の方がおよそ高い結果が得られていることがわかる。しかしながら、表層的特徴 (SF) のみを用いた場合を比べて見ると、コメントツイート数の分類の場合の方が Yahoo!ニュース内コメント数の分類の場合よりも評価値が高い (表 5.9)。

**表** 5.12. Yahoo!ニュース内コメント数の分類：実験 1

| 特徴 | コメント数 $\geq 200$ | | | コメント数 $\geq 700$ | | |
|---|---|---|---|---|---|---|
| | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| SF | 0.597 | 0.658 | 63.0% | 0.645 | 0.604 | 62.5% |
| LI | 0.697 | 0.712 | 70.8% | 0.734 | 0.745 | 74.0% |
| TM | 0.565 | 0.505 | 53.7% | 0.475 | 0.567 | 52.5% |
| SF + LI | 0.707 | 0.724 | 71.6% | 0.726 | 0.741 | 73.4% |
| SF + TM | 0.649 | 0.673 | 66.1% | 0.657 | 0.627 | 64.3% |
| LI + TM | 0.698 | 0.720 | 71.0% | 0.722 | 0.747 | 73.5% |
| SF + LI + TM | 0.708 | 0.723 | 71.6% | 0.727 | 0.749 | 73.8% |
| SF + LI + TM + EV | 0.709 | 0.723 | 71.6% | 0.730 | 0.749 | 74.0% |

言語的特徴 (LI) で記事 (LI-AR) と見出し (LI-HE)、関連リンク (LI-RL) の影響の大きさを比べると、記事 (LI-AR) が特徴として最重要であることはこれまで通りであるが、境界値 700 の場合には見出し (LI-HE) のみを特徴とした場合と、記事 (LI-AR) のみを特徴とした場合の結果に大きな差が出ており、記事と見出しを組み合わせて特徴とした場合（LI-AR+LI-HE）も記事のみ（LI-AR）の

場合からの結果の改善はほぼない。ニュース記事の見出しのみ (LI-HE-n) の場合とヤフー作成の見出しを加えた場合（LI-HE）にほとんど結果が変わらないことは、コメントツイート数の分類の場合と同様である（表 5.13)。

**表 5.13**. Yahoo!ニュース内コメント数の分類：実験 2

| 特徴 | コメント数 ≥ 200 | | | コメント数 ≥ 700 | | |
|---|---|---|---|---|---|---|
| | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| LI-AR | 0.689 | 0.697 | 69.3% | 0.741 | 0.748 | 74.5% |
| LI-HE-y | 0.665 | 0.604 | 63.7% | 0.600 | 0.613 | 60.7% |
| LI-HE-n | 0.672 | 0.645 | 65.9% | 0.607 | 0.645 | 62.7% |
| LI-HE | 0.680 | 0.649 | 66.5% | 0.590 | 0.647 | 62.1% |
| LI-AR + LI-HE | 0.698 | 0.719 | 70.9% | 0.741 | 0.751 | 74.6% |
| LI-AR + LI-HE + LI-RL* | 0.697 | 0.718 | 70.8% | 0.734 | 0.745 | 74.0% |

最後に文字特徴（LI-Character）・固有表現 (LI-NamedEntity)・感情特徴 (LI-Sentiment) の影響の大きさを比べると、境界値 200 の場合は文字 (LI-Character) を特徴とした場合と固有表現 (LI-NamedEntity) を特徴とした場合に結果に大きな差は見られないが、境界値 700 の場合は文字 (LI-Character) を特徴とした場合が他の２つの場合よりも格段に良い結果が得られている（表 5.14)。

**表 5.14**. Yahoo!ニュース内コメント数の分類：実験 3

| 特徴 | コメント数 ≥ 200 | | | コメント数 ≥ 700 | | |
|---|---|---|---|---|---|---|
| | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| LI-Character | 0.678 | 0.716 | 69.8% | 0.738 | 0.750 | 74.5% |
| LI-NamedEntity | 0.690 | 0.688 | 68.9% | 0.659 | 0.664 | 66.1% |
| LI-Sentiment | 0.658 | 0.665 | 66.2% | 0.659 | 0.685 | 67.2% |

以上より関連コメントツイート数の分類と Yahoo!ニュース内コメント数の分類においては、表層的特徴 (SF)・言語的特徴 (LI)・時間的特徴 (TM)・環境的特徴 (EV) の中では、時間的特徴 (TM) と環境的特徴 (EV) の影響力は非常に小さく、言語的特徴 (LI) の影響が最も大きいこと、特に文字特徴 (LI-Character) の影響が感情度の場合とは異なり大きいことがわかった。

### 5.3.3 カテゴリー別コメント数の分類結果

ここからはカテゴリー別にコメント数の伸びの分類を行った結果を示す。ただしデータ数が正例・負例ともに十分である場合のみ実験を行うこととし、境界値 200 での Yahoo!ニュース内のコメント数の分類は、コンピューター・サイエンスカテゴリーでは実施しない。また特徴の組み合わせは実験 1 の場合（表層的特徴 (SF)、言語的特徴 (LI)、時間的特徴 (TM)、環境的特徴 (EV) の組み合わせ）とした。

境界値における正例・負例のデータ数については表 5.15、表 5.16 に示す。

**表 5.15**. カテゴリー別コメントツイート数

| カテゴリー | コメント数 ≥ 100 | | コメント数 ≥ 300 | |
|---|---|---|---|---|
| | 正例 | 負例 | 正例 | 負例 |
| 国内 | 390 | 241 | 113 | 518 |
| 国際 | 189 | 376 | 32 | 533 |
| 地域 | 311 | 388 | 76 | 623 |
| 経済 | 262 | 197 | 76 | 383 |
| エンターテイメント | 467 | 135 | 139 | 463 |
| スポーツ | 289 | 405 | 33 | 661 |
| コンピューター | 121 | 72 | 29 | 164 |
| サイエンス | 96 | 108 | 16 | 188 |

**表 5.16**. カテゴリー別 Yahoo!ニュース内コメント数

| カテゴリー | コメント数 ≥ 200 | | コメント数 ≥ 700 | |
|---|---|---|---|---|
| | 正例 | 負例 | 正例 | 負例 |
| 国内 | 229 | 126 | 76 | 279 |
| 国際 | 144 | 268 | 34 | 378 |
| 地域 | 108 | 156 | 32 | 232 |
| 経済 | 136 | 127 | 37 | 226 |
| エンターテイメント | 353 | 193 | 94 | 452 |
| スポーツ | 277 | 338 | 34 | 581 |
| コンピューター | 32 | 102 | 3 | 131 |
| サイエンス | 26 | 81 | 9 | 98 |

**国内**

境界値 100 のコメントツイート数の分類では、これまでの全カテゴリーの場合と比べ全体的に評価値は低いものの、特徴による影響の大きさの違いという面ではあまり差は見られない（表 5.17)。境界値 300 のコメントツイート数の分類でも、この傾向は変わらない（表 5.19)。境界値 200 の Yahoo!ニュース内コメント数の分類も、これまでの全カテゴリーの場合と評価値とも非常に似ているが、国内カテゴリーのみで分類を行った場合の方が、表層的特徴（SF）を単独で特徴として用いた場合の結果が比較的良いと言える（表 5.17)。

**国際**

国際カテゴリーにおいては、境界値 100 のコメントツイート数の分類の場合・境界値 200 の Yahoo!ニュース内コメント数の分類の場合共に、全カテゴリーでの分類の場合に比べ表層的特徴 (SF) の影響よりも時間的特徴 (TM) の影響が大きく、時間的特徴 (TM) がコメント数の伸びに影響している。また境界値 100 のコメントツイート数の分類では、言語的特徴 (LI) のコメント数の伸びへの寄与が見られるが、境界値 200 の Yahoo!ニュース内コメント数の分類では、言語的特徴 (LI) のみを特徴と

した場合の評価値が比較的低く留まっているのはもちろん、表層的特徴 (SF) と時間的特徴 (TM) を組み合わせた場合の結果が最良で、さらに言語的特徴を加える (SF + LI + TM) ことで、結果が多少ではあるものの悪くなっている（表 5.17)。

**地域**

地域カテゴリーについては全カテゴリーの場合と比べて際立った点は見られず、似通っていると言える（表 5.17)。

**経済**

経済カテゴリーでは、境界値 100 のコメントツイート数分類において非常に珍しく、表層的特徴・言語的特徴・時間的特徴（SF + LI + TM）を組み合わせて学習を行った場合と環境的特徴も組み合わせに加えた場合（SF + LI + TM + EV）の結果を比べた場合に後者の方がわずかではあるものの F 値・正解値共に高い。しかしながら、境界値 200 の Yahoo!ニュース内コメントではこのような傾向は見られない。またこの境界値 200 の Yahoo!ニュース内コメントの分類の特徴としては表層的特徴 (SF) の影響が微小であることが挙げられる（表 5.17)。

**エンターテイメント**

エンターテイメントカテゴリーは既に述べたようにコメント数が伸び易いカテゴリーであるが、境界値 100 のコメントツイート数分類においては言語的特徴 (LI) の影響が他の特徴よりも大きく見られるものの、境界値 300 のコメントツイート数分類・境界値 200 の Yahoo!ニュース内コメント数の分類においては言語的特徴 (LI) の影響は全カテゴリーの場合に比べ小さく、境界値 300 のコメントツイート数分類では時間的特徴 (TM) にコメント数の伸びが大きく左右されている（表 5.18、表 5.19)。

**スポーツ**

スポーツカテゴリーに関しては他のカテゴリーと比べ、境界値 100 のコメントツイート数の分類における F 値・正答率が比較的低い（表 5.18)。

**コンピューター**

コンピューターカテゴリーでは Twitter 上では比較的コメントが伸び易いカテゴリー（5.2）であるが、境界値 100 のコメントツイート数分類では全カテゴリーの場合に比べ、言語的特徴 (LI) の寄与が小さい（表 5.18)。

**サイエンス**

サイエンスカテゴリーの境界値 100 のコメントツイート数分類は、言語的特徴 (LI) が他の特徴に比べ結果に対する影響が大きい（表 5.18)。

**表 5.17.** カテゴリー別コメント数の分類 1 ：実験 1

| | 特徴 | コメントツイート数 $\geq 100$ | | | *Yahoo*!ニュース内コメント数 $\geq 200$ | | |
|---|---|---|---|---|---|---|---|
| | | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| 国内 | SF | 0.655 | 0.572 | 61.8% | 0.749 | 0.552 | 67.9% |
| | LI | 0.627 | 0.651 | 63.9% | 0.750 | 0.644 | 70.6% |
| | TM | 0.490 | 0.626 | 56.8% | 0.519 | 0.552 | 53.6% |
| | SF + LI | 0.644 | 0.650 | 64.7% | 0.718 | 0.667 | 69.4% |
| | SF + TM | 0.652 | 0.578 | 61.8% | 0.747 | 0.574 | 68.3% |
| | LI + TM | 0.639 | 0.648 | 64.3% | 0.763 | 0.654 | 71.8% |
| | SF + LI + TM | 0.645 | 0.642 | 64.3% | 0.760 | 0.647 | 71.4% |
| | SF + LI + TM + EV | 0.646 | 0.649 | 64.7% | 0.732 | 0.703 | 71.8% |
| 国際 | SF | 0.654 | 0.570 | 61.6% | 0.599 | 0.519 | 56.3% |
| | LI | 0.726 | 0.718 | 72.2% | 0.629 | 0.649 | 63.9% |
| | TM | 0.602 | 0.658 | 63.2% | 0.589 | 0.667 | 63.2% |
| | SF + LI | 0.732 | 0.717 | 72.5% | 0.626 | 0.644 | 63.5% |
| | SF + TM | 0.679 | 0.681 | 68.0% | 0.639 | 0.696 | 67.0% |
| | LI + TM | 0.738 | 0.721 | 73.0% | 0.648 | 0.664 | 65.6% |
| | SF + LI + TM | 0.740 | 0.720 | 73.0% | 0.648 | 0.664 | 65.6% |
| | SF + LI + TM + EV | 0.746 | 0.686 | 72.0% | 0.609 | 0.665 | 63.9% |
| 地域 | SF | 0.623 | 0.569 | 59.8% | 0.633 | 0.687 | 66.2% |
| | LI | 0.644 | 0.658 | 65.1% | 0.712 | 0.732 | 72.2% |
| | TM | 0.573 | 0.483 | 53.2% | 0.592 | 0.565 | 57.9% |
| | SF + LI | 0.656 | 0.656 | 65.6% | 0.724 | 0.739 | 73.1% |
| | SF + TM | 0.623 | 0.569 | 59.8% | 0.618 | 0.649 | 63.4% |
| | LI + TM | 0.625 | 0.671 | 65.0% | 0.744 | 0.747 | 74.5% |
| | SF + LI + TM | 0.649 | 0.653 | 65.1% | 0.737 | 0.735 | 73.6% |
| | SF + LI + TM + EV | 0.641 | 0.673 | 65.8% | 0.726 | 0.728 | 72.7% |
| 経済 | SF | 0.667 | 0.663 | 66.5% | 0.521 | 0.564 | 54.3% |
| | LI | 0.696 | 0.668 | 68.3% | 0.681 | 0.738 | 71.3% |
| | TM | 0.554 | 0.601 | 57.9% | 0.543 | 0.581 | 56.3% |
| | SF + LI | 0.702 | 0.677 | 69.0% | 0.678 | 0.734 | 70.9% |
| | SF + TM | 0.654 | 0.689 | 67.3% | 0.644 | 0.606 | 62.6% |
| | LI + TM | 0.689 | 0.665 | 67.8% | 0.678 | 0.740 | 71.3% |
| | SF + LI + TM | 0.696 | 0.668 | 68.3% | 0.684 | 0.736 | 71.3% |
| | SF + LI + TM + EV | 0.724 | 0.690 | 70.8% | 0.661 | 0.726 | 69.7% |

**表** 5.18. カテゴリー別コメント数の分類２：実験１

| | 特徴 | コメントツイート数 $\geq 100$ | | | *Yahoo*!ニュース内コメント数 $\geq 200$ | | |
|---|---|---|---|---|---|---|---|
| | | F 値 (+) | F 値 (−) | 正答率 | F 値 (+) | F 値 (−) | 正答率 |
| エンターテイメント | SF | 0.625 | 0.517 | 57.8% | 0.595 | 0.471 | 54.1% |
| | LI | 0.669 | 0.657 | 66.3% | 0.617 | 0.631 | 62.4% |
| | TM | 0.525 | 0.562 | 54.4% | 0.577 | 0.679 | 63.5% |
| | SF + LI | 0.669 | 0.664 | 66.7% | 0.619 | 0.635 | 62.7% |
| | SF + TM | 0.620 | 0.535 | 58.1% | 0.599 | 0.618 | 60.9% |
| | LI + TM | 0.657 | 0.662 | 65.9% | 0.637 | 0.648 | 64.2% |
| | SF + LI + TM | 0.664 | 0.654 | 65.9% | 0.644 | 0.656 | 65.0% |
| | SF + LI + TM + EV | 0.647 | 0.634 | 64.1% | 0.629 | 0.655 | 64.2% |
| スポーツ | SF | 0.625 | 0.465 | 55.9% | 0.537 | 0.549 | 54.3% |
| | LI | 0.583 | 0.610 | 59.7% | 0.630 | 0.641 | 63.5% |
| | TM | 0.538 | 0.611 | 57.8% | 0.579 | 0.631 | 60.6% |
| | SF + LI | 0.595 | 0.618 | 60.7% | 0.625 | 0.638 | 63.2% |
| | SF + TM | 0.602 | 0.554 | 58.0% | 0.621 | 0.594 | 60.8% |
| | LI + TM | 0.595 | 0.622 | 60.9% | 0.650 | 0.661 | 65.5% |
| | SF + LI + TM | 0.605 | 0.623 | 61.4% | 0.618 | 0.664 | 64.3% |
| | SF + LI + TM + EV | 0.614 | 0.597 | 60.6% | 0.644 | 0.665 | 65.6% |
| コンピューター | SF | 0.696 | 0.614 | 66.0% | — | — | — |
| | LI | 0.702 | 0.564 | 64.6% | — | — | — |
| | TM | 0.632 | 0.684 | 66.0% | — | — | — |
| | SF + LI | 0.709 | 0.569 | 65.3% | — | — | — |
| | SF + TM | 0.657 | 0.689 | 67.4% | — | — | — |
| | LI + TM | 0.692 | 0.620 | 66.0% | — | — | — |
| | SF + LI + TM | 0.662 | 0.612 | 63.9% | — | — | — |
| | SF + LI + TM + EV | 0.671 | 0.545 | 61.8% | — | — | — |
| サイエンス | SF | 0.612 | 0.551 | 58.3% | — | — | — |
| | LI | 0.680 | 0.629 | 65.6% | — | — | — |
| | TM | 0.491 | 0.608 | 55.7% | — | — | — |
| | SF + LI | 0.670 | 0.606 | 64.1% | — | — | — |
| | SF + TM | 0.613 | 0.636 | 62.5% | — | — | — |
| | LI + TM | 0.748 | 0.472 | 65.9% | — | — | — |
| | SF + LI + TM | 0.670 | 0.606 | 64.1% | — | — | — |
| | SF + LI + TM + EV | 0.680 | 0.652 | 66.7% | — | — | — |

**表** 5.19. カテゴリー別コメント数の分類（その他）：実験 1

| | 特徴 | コメントツイート数 $\geq 300$ | | |
|---|---|---|---|---|
| | | F 値 (+) | F 値 (−) | 正答率 |
| 国内 | SF | 0.538 | 0.719 | 65.0% |
| | LI | 0.676 | 0.678 | 67.7% |
| | TM | 0.540 | 0.582 | 56.2% |
| | SF + LI | 0.664 | 0.672 | 66.8% |
| | SF + TM | 0.608 | 0.705 | 66.4% |
| | LI + TM | 0.687 | 0.684 | 68.6% |
| | SF + LI + TM | 0.682 | 0.690 | 68.6% |
| | SF + LI + TM + EV | 0.667 | 0.687 | 67.7% |
| エンターテイメント | SF | 0.607 | 0.455 | 54.3% |
| | LI | 0.577 | 0.536 | 55.8% |
| | TM | 0.679 | 0.701 | 69.1% |
| | SF + LI | 0.580 | 0.556 | 56.8% |
| | SF + TM | 0.677 | 0.703 | 69.1% |
| | LI + TM | 0.614 | 0.609 | 61.2% |
| | SF + LI + TM | 0.607 | 0.601 | 60.4% |
| | SF + LI + TM + EV | 0.620 | 0.603 | 61.2% |

### 5.3.4 関連コメント数の分類に関する考察

関連コメント数の伸びの予測に関しても、Twitter 上のコメントよりも Yahoo!ニュース内のコメントの方が予測し易いという結果になったが、表層的特徴と言語的特徴に注目して結果を見ると、Twitter のコメント数の方が比較的表層的特徴の影響を受け易く、Yahoo!ニュース内のコメント数の方が言語的特徴の影響を受け易いことがわかる。つまり、Twitter を利用してコメントを発信する人の方が、記事そのものの内容よりもカテゴリーや配信元などの情報に左右され易いと解釈できる。理由としては、Twitter のユーザの方が記事内容を読まずにコメントを残している場合が多いという可能性や、Yahoo!ニュースのコメント欄では同記事に対して他人が書き込んだコメントを読んだ上で自らコメントを残すことができるために、記事の要点が明確にされやすく記事に対する理解が自然と深まっているという可能性も考えられる。

感情度の回帰・コメント数の分類ともに時間的特徴の寄与はあまり見られてこなかったが、例外的にエンターテイメントカテゴリーでコメントツイート数が全記事の上位 10% に入るかという分類を行った場合に非常に高い寄与が見られた。分類の境界となっているコメント数 300 という値は、エンターテイメントカテゴリーの中では上位 25%（図 5.1）であり、境界値による時間的特徴の影響度の違いも可能性としては考えられたが、国内カテゴリーにおいてもコメント数 300 という境界値はカテゴリー内のおよそ上位 25% にあたり、国内カテゴリーの結果からは時間的特徴の影響が見られない（表 5.19）為、エンターテイメントカテゴリー特有の傾向である可能性が高い。しかしながら本実験ではデータ数が限られているので、更なる実験の必要がある。

# 第 6 章

# 結論

ソーシャルメディアを通じて新たな発展を遂げるジャーナリズムへの注目が研究者の中で高まっているという昨今の研究動向と、社会に存在する望ましい報道発表手法の解明、ニュース提供画面の最適化というニーズを勘案し、効果的な言語特性の使用手法または言語特性から予想されるユーザの反応を考慮したニュース提供手法の確立、という将来的な研究において基礎となるであろう知見の導出を目的として本研究は行われた。そして研究の中でニュース記事に含まれるあらゆる言語特性を用いたユーザコメントの分析手法を提案し、提案手法を用いてニュースコンテンツがユーザコメントへ与える影響度をコメントの感情度とコメント数の伸びという２つの基準から、ソーシャルサイトとポータルサイトという２種類のメディアにおいて分析した。

本研究の独自性は、

1. 言語的特徴を網羅的に使用する
2. 複数メディアのコメントを研究対象とする
3. コメントの感情度分析を機械学習で行う

という３点にあり、得られた実験結果と新知見は次の通りである。第一に、言語的特徴を網羅的に使用したニュースコンテンツのコメントへの影響分析を行うことにより、結果として執筆者にとって可変的な特徴である、言語的特徴がユーザのコメントに少なからず影響を与えていることを示唆する実験結果を得た。よって、将来的に各言語特性の影響研究を行うことの有効性が明らかになった。

第二に、ポータルサイト（Yahoo!ニュース）とソーシャルサイト（Twitter）という複数メディアを用いて実験を行うことにより、コメントの感情度の回帰・コメント数の伸びの分類共に Yahoo!ニュース内のコメントよりもコメントツイートで精度が下がるという結果を得た。理由としては、実際に Twitter のコメントは記事からの予測が難しいという可能性の他に、Twitter 上のコメントは独自に考案したアルゴリズムによる抽出の過程を踏んでいるため、ノイズが完全に除去されず、結果に影響を与えているという可能性が考えられる。よって”social news” など新しいジャーナリズムに関する研究が進んでいくであろう今後、より良いコメント抽出のアルゴリズムが確立されていくことで、両メディアによるコメント特徴の違いが一層明白になることが期待される。

第三に機械学習によるユーザコメントの感情度予測により、提案手法による実験から Yahoo!ニュース内のコメントでは 0.693 という決定係数を得、なおかつニュース記事の感情とコメントの感情との相関から、記事と同程度のポジティブまたはネガティブな感情をコメントが示しやすい傾向があることを明らかにすることができた。しかしながらユーザの多くが記事に共感しやすいのか、記事に共感した場合にコメントをし易いのか、という点には議論が残っており、今後の研究が待たれる状態である。

以上のように本提案手法による実験の結果、コメントの感情度・数の伸びに対する言語的特徴の影響を確認し新しい知見を得ることができたものの、コメント数の伸び分類、殊更 Twitter 上のコメント数の伸び分類では、予測結果は芳しくなかった。

いずれのメディア・コメントによる反応についても、素性として使用した表層的素性・言語的素性・時間的素性・環境的素性のみでは結果のすべてを説明することができなかったということは、従来のジャーナリズムから新しいジャーナリズムへの移行に伴い、ソーシャルサイトで個人の属するコミュニティやポータルサイトで既に書き込まれている周囲の発言といった人々の相互作用が見逃せないようになっており、それらを考慮せずにユーザのコメントを予測することには今の時代では限界がある、ということなのかもしれない。しかしながら言語的特徴のコメントへの影響度を複数のメディアで様々なデータ、特徴の組み合わせから調べたことは今後の研究の発展と、より実用的知識の創出へと繋がっていくと期待したい。

# 謝辞

本研究を進めるにあたり、坂田・森研究室の多くの皆様にご指導いただきましたこと、この場を御借りして深く感謝し申し上げます。

特に指導教員の森純一郎特任講師、並びに共同研究室の坂田一郎教授には論文の題目決定から、設計・執筆に至るまで、非常に多くのアドバイスを頂きました。

株式会社ホットリンク兼本学研究員の榊剛史氏、同じく本学研究員の原忠義氏、博士課程３年の丸井淳己氏には研究アイデアのブラッシュアップや基礎知識の取得、論文校正等で大変お世話になりました。

修士１年の株田氏と山下氏、そして同期の石塚くん、河津くんからは常に良い刺激を受け、何も知識がない状態からここまで研究を進めることができました。

わずか半年の期間でしたが坂田・森研究室に所属し、皆様の多大なご協力の下、研究テーマの設定から論文執筆に至る全ての過程を踏み、一つの研究をやり遂げることができたこと、またそのような機会を頂けたことに、大変感謝するとともに、この経験を今後に生かせるよう精一杯努力して参りたいと思っております。皆様への感謝の念を綴ると共に、今後の研究室の更なる発展を願って、謝辞とさせて頂きます。

どうもありがとうございました。

平成 27 年 2 月 5 日
東京大学工学部システム創成学科
知能社会システムコース ４年
上子 優香

# 参考文献

[1] Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The Pulse of News in Social Media: Forecasting Popularity." ICWSM. 2012.

[2] Artzi, Yoav, Patrick Pantel, and Michael Gamon. "Predicting responses to microblog posts." Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012.

[3] Tatar, Alexandru, et al. "Predicting the popularity of online articles based on user comments." Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

[4] Java, Akshay, et al. "Why we twitter: understanding microblogging usage and communities." Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007.

[5] Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[6] Subašić, Ilija, and Bettina Berendt. "Peddling or creating? Investigating the role of Twitter in news reporting." Advances in Information Retrieval. Springer Berlin Heidelberg, 2011. 207-213.

[7] Petrović, Saša, et al. "Can Twitter replace Newswire for breaking news?." ICWSM. 2013.

[8] Osborne, Miles, and Mark Dredze. "Facebook, Twitter and Google Plus for Breaking News: Is There a Winner?." Proceedings of the International Conference on Weblogs and Social Media. 2014.

[9] Castillo, Carlos. "Traffic prediction and discovery of news via news crowds." Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

[10] Lehmann, Janette, et al. "Transient News Crowds in Social Media." ICWSM. 2013.

[11] Lehmann, Janette, et al. "Finding news curators in twitter." Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

[12] Saez-Trumper, Diego, Carlos Castillo, and Mounia Lalmas. "Social media news communities: gatekeeping, coverage, and statement bias." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[13] Lerman, Kristina, and Rumi Ghosh. "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks." ICWSM 10 (2010): 90-97.

[14] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. 2010.

[15] 熊本忠彦, 田中克己: Web ニュース記事を対象とする喜怒哀楽抽出システム, インタラクション 2005, pp.25-26(2005).

[16] 高野憲悟，萩原将文: 感情関連語を用いた感情推定法の提案とニュースサイトのアクセス解析への応用，感性工学研究論文集，Vol.11, No.3, pp.495-502(2012).

[17] 内藤和宏, 榎堀優, 梶田将司, 間瀬健二: Twitter コメントに含まれる感情語がイベント印象に与える影響の評価, インタラクション 2012, pp.871-876(2012).

[18] 中村明: 感情表現辞典, 東京堂出版 (1993).

[19] O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." ICWSM 11 (2010): 122-129.

[20] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.

[21] Popescu, Ana-Maria, and Marco Pennacchiotti. "Detecting controversial events from twitter." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[22] Popescu, Ana-Maria, and Marco Pennacchiotti. "" Dancing with the Stars," NBA Games, Politics: An Exploration of Twitter Users' Response to Events." ICWSM. 2011.

[23] Tsagkias, Manos, Maarten de Rijke, and Wouter Weerkamp. "Linking online news and social media." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[24] Shi, Bichen, Georgiana Ifrim, and Neil Hurley. "Be In The Know: Connecting News Articles to Relevant Twitter Conversations." arXiv preprint arXiv:1405.3117 (2014).

[25] 邱起仁, 櫨山淳雄: ニュース記事に関連するツイート収集手法の提案とその評価, 情報処理学会研究報告. EC, エンタテインメントコンピューティング 2014-EC-31(57), pp.1-6(2014)

[26] Kothari, Alok, et al. "Detecting Comments on News Articles in Microblogs." ICWSM. 2013.

[27] Štajner, Tadej, et al. "Automatic selection of social media responses to news." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

[28] Lakkaraju, Himabindu, Julian J. McAuley, and Jure Leskovec. "What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media." ICWSM. 2013.

[29] Mahmud, Jalal. "Why Do You Write This? Prediction of Influencers from Word Use." Eighth International AAAI Conference on Weblogs and Social Media. 2014.

[30] Tsagkias, Manos, Wouter Weerkamp, and Maarten De Rijke. "Predicting the volume of comments on online news stories." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[31] Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval." Journal of documentation 28.1 (1972): 11-21.

[32] 栗田多喜夫: サポートベクターマシン入門, 産業総合技術研究所, http://home.hiroshima-u.ac.jp/tkurita/lecture/svm.pdf, (参照 2015-02-01).

[33] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一: 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).

[34] 東山昌彦, 乾健太郎, 松本裕治: 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第 14 回年次大会論文集, pp.584-587(2008).

[35] 奥村倫弘: ヤフー・トピックスの作り方, 光文社新書 (2010)