4I1-3

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

Predicting the influence of online news on social media using comprehensive textual features

上子 優香 *1 Yuka Kamiko　榊 剛史 *2*3 Takeshi Sakaki　原 忠義 *3 Tadayoshi Hara　森 純一郎 *3 Junichiro Mori　坂田 一郎 *3 Ichiro Sakata

*1東京大学工学部システム創成学科
Department of Systems Innovation, Faculty of Engineering, The University of Tokyo

*2株式会社ホットリンク
Hottolink, Inc.

*3東京大学大学院工学系研究科
School of Engineering, The University of Tokyo

Recent widespread use of social media has enabled us to observe feedback from the public on online news through comments. Hence some research has attempted to reveal the influence of provided news items on readers by proposing the method to predict the number of comments on the news from several features of its contents. Even though online news articles are mostly composed of texts, however, a limited number of studies have focused on textual features. In this research, we take comprehensive textual features into account and predict not only the number but also the sentiment of comments on news items. In addition, we provide the comparison between two major social media services, which allow us to explore the wider range of possibilities for revealing the influence of provided online news contents on readers.

1. 序論

近年のソーシャルメディアの普及に伴い、ジャーナリズム、特にニュースコンテンツへの人々の反応が、投稿・コメントといった形で飛躍的に顕在化しつつある。また、これらの反応を解析することで、発信されたコンテンツが人々へ与える影響を分析し、様々な応用事例へと活用する可能性について、多くの研究が試みられている。既存の研究においては人々への影響の指標としてニュースコンテンツに関する投稿数・引用数・コメント数などを用いて、これらの指標をコンテンツが持つ特徴から予測する試みがなされている [Bandari 12][Tsagkias 09] が、いずれの研究も主に表層的な特徴に着目しており、本質的なところであるはずのニューステキスト自体にはあまり言及していない。また、ソーシャルメディア毎による予測可能性の差異についてもまだ十分に研究は進んでいない。

本研究では、ニュースコンテンツがコメント数やコメントの感情へ与える影響について、従来着目されていたコンテンツの表層的特徴のみならず、 ニューステキストの言語的特徴に関しても、それらを利用し各指標を予測する機械学習モデルを学習することで、各特徴量への回帰を試みる。また、具体的なコメントデータとして、Twitter 上でのコメント、Yahoo!ニュースのコメント欄、それぞれについて本アプローチでの分析を行うことで、複数のメディアでの影響の差異を考察する。

2. ソーシャルメディア上のニュースの影響度予測手法

2.1 影響度予測手法概要

本研究のフレームワークを図 1 に示す。図 1 のようにニュースコンテンツとそのコンテンツに対するコメントをデータとして与え、学習させることによりコメント予測を行うことができる。本研究ではテキストデータを主とするニュースコンテンツの素性を基にサポートベクターマシンを用いて予測モデルを構築し、コメントの影響度をコメントの感情度・コメント数という指標で判断することによって、ニュースコンテンツの影響を分析する。以下、各指標および素性の生成方法について説明する。

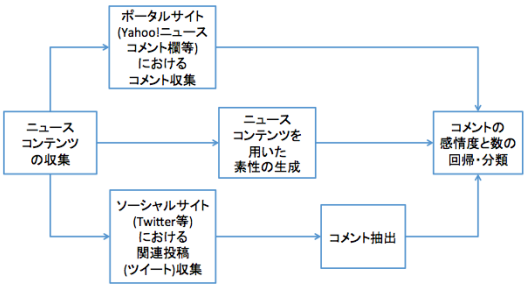


図 1: 本研究のフレームワーク

2.1.1 コメントの感情度回帰

感情をネガティブであるほど 0、ポジティブであるほど 1 に近づくよう数値化したものを「感情度」と呼ぶ（定義は 2.2 節を参照）。本研究においてはコメントの感情度の回帰を行うが、記事に対するコメントの感情度は、その記事に対して付いた各コメントの感情度の平均とし、コメントが 5 件以上ある記事を対象として学習を行う。

2.1.2 コメント数の分類

収集したコメントデータから、各サイトでコメント数が全記事のおよそ上位 50%、上位 10%になる境界値を設定する。そしてある記事に付与されるコメント数が、全体の上位 50%以上または未満か、全体の上位 10%以上または未満か、という 2 種類の分類を行う。

連絡先: 東京大学大学院工学系研究科技術経営戦略学専攻
〒 113-8656 東京都文京区本郷 7-3-1 工学部 3 号館 203
TEL: 03-5841-1161
E-mail: kamiko@ipr-ctr.t.u-tokyo.ac.jp

2.2 ニュース記事を用いた素性の生成

予測モデルを構築する為、収集したニュースデータから素性を生成する。使用する素性は表層的素性・時間的素性・環境的素性・言語的素性という 4 つのセットに大きく分けられ、詳細は表 1・2 の通りである。

表 1: 素性リスト 1

表層的素性 (SF)	時間的素性 (TM)	環境的素性 (EV)
カテゴリ	曜日	天気
配信元	時刻	気温
リンク先	配信時間	日経平均株価

表 2: 素性リスト 2

言語的素性 (LI)					
		素性名	素性セット名		
			記事 (AR)	見出し (HE)	関連リンク (RL)
素性セット名	文字 (CH)	N-gram(N=2,3)	○		
		TF-IDF	○		
		記号・数字		○	
		平仮名・片仮名		○	
		文字数	○	○	
		品詞	○		
	固有表現 (NE)		○	○	
	感情 (SE)	客観性	○		
		感情語	○	○	○
		各感情語の数	○	○	○
		感情度	○		
		類似度		○	

客観性・感情語・各感情語の数・感情度の 4 つの素性では日本語評価極性辞書 *1 に収録された感情語のマッチングを行い、ポジティブ・ニュートラル・ネガティブのいずれかの感情を示す言葉を抜き出す。辞書では感情語に対して主観・客観というラベル付けがされている為、客観性を次のように定義する。

$$\text{客観性} = \frac{1}{o+s}\begin{pmatrix} o & s \end{pmatrix}\begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}$$

(o : 客観語の数, s : 主観語の数)

感情語という素性は、テキスト中の辞書とマッチした語とその出現回数である。感情度については次のように定義する。

$$\text{感情度} = \frac{1}{p+e+n}\begin{pmatrix} p & e & n \end{pmatrix}\begin{pmatrix} 1.0 \\ 0.5 \\ 0.0 \end{pmatrix}$$

(p : ポジティブ語の数, e : ニュートラル語の数, n : ネガティブ語の数)

ただしコメント数の分類の際には 3.2 節で述べる対象データの事前調査に基づき、感情度の代わりに感情の非中立性

$$\text{感情の非中立性} = |\,\text{感情度} - 0.5|$$

を用いる。

*1 http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources

3. データの収集及び感情度に関する事前調査

3.1 データセット

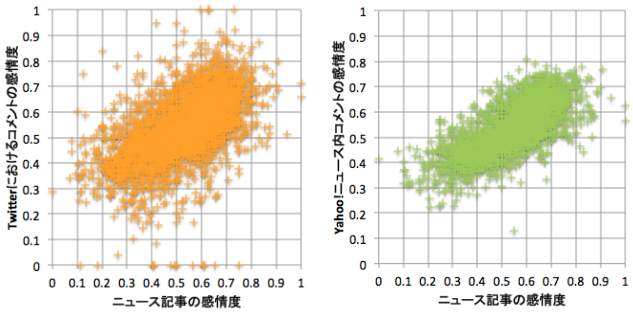
ニュースデータは Yahoo!ニュース *2 から、コメントデータは同じくポータルサイト Yahoo!ニュース、そしてソーシャルサイト Twitter*3 から入手する。まず Yahoo!ニュースにおいてトピックスとして取り上げられたニュースコンテンツの収集を行い、その後 Yahoo!ニュースのコメント欄に投稿されたコメントと、Twitter に投稿された関連ツイートの収集を行う。関連ツイートの収集には Twitter Search API を利用し、ニュース記事の見出しまたは URL が含まれるツイートをニュースコンテンツの関連ツイートとして取得する。関連ツイートからは、ユーザが独自に付与した「コメント」部分を抽出する。ニュース記事本文の反復や見出しは「コメント」に含まない。

本研究においては 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 4,078 件をデータとして収集した。Yahoo!ニュース内のコメントは、トピックスとしてニュース記事が取り上げられてから 3 日後に取得を行った。Twitter の関連ツイートはニュースがトピックスとして掲載された 12、24、48、72 時間後にデータを取得し、およそ関連ツイート数が頭打ちになった 24 時間後のデータを使って、関連ツイートからコメントの抽出を行った。

以上のように収集したデータからコメント数が全記事のおよそ上位 50%、上位 10%になる境界値を調査し、Twitter におけるコメントの境界値を 100 コメントと 300 コメント、Yahoo!ニュース内におけるコメントの境界値を 200 コメントと 700 コメントと決定した。

3.2 感情度の事前調査

ニュース記事の感情度とコメントの感情度との相関を調べた結果、図 2 のようにニュース記事の感情度と Twitter 上のコメントの感情度との相関係数は 0.59、Yahoo!ニュース内のコメントの感情度との相関係数は 0.71 となり、ニュース記事の感情にコメントの書き手は共感し易い傾向があるようであった。

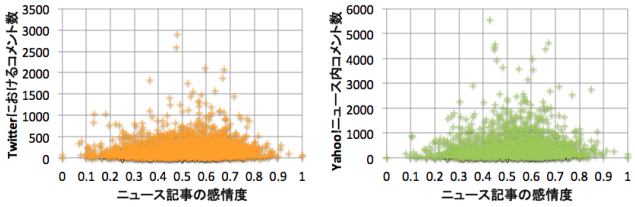


[1]Twitter [2]Yahoo!ニュース

図 2: ニュース記事の感情度とコメントの感情度の関係

しかしながら、ニュース記事の感情度とコメント数との関係を見ると、図 3 のように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られた。よってコメント数の分類においては、2.2 節で既に述べたように感情度の代わりに感情の非中立性を用いることとした。

*2 http://news.yahoo.co.jp
*3 https://twitter.com



[1]Twitter [2]Yahoo!ニュース

図 3: ニュース記事の感情度とコメントの数の関係

4. 実験

4.1 素性の組み合わせ

以下の実験 1 から実験 3 の素性の組み合わせから特徴ベクトルを作成し、機械学習により予測モデルを構築してコメントの感情度回帰・コメント数の分類を行った（省略文字の表す具体的な素性については表 1・2 を参照）。実験 1 はおよそ従来研究で用いられてきた素性と言語的素性との比較を行うこと、実験 2・3 は素性数が多い言語的素性を素性取得要素別、素性の性質別に細分化することにより、言語的素性の中でも特に影響力の高い素性を明らかにすることを目的とした。

実験 1-1: SF
実験 1-2: LI (LI-AR + LI-HE + LI-RL)
実験 1-3: SF + LI
実験 1-4: SF + LI + TM
実験 1-5: SF + LI + TM + EV
実験 2-1: LI-AR
実験 2-2: LI-HE
実験 2-3: LI-AR + LI-HE
実験 3-1: LI-CH
実験 3-2: LI-NE
実験 3-3: LI-SE

4.2 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 プログラムを作成・実行することにより行い、形態素解析・品詞の分類には MeCab[*4] を用いた。TF-IDF の計算には Python の機械学習ライブラリである scikit-learn[*5] 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は、事前実験の結果に基づきサポートベクターマシンを用いることとし、実装に LIBSVM[*6] を使い線形カーネルを用いた学習を行った。

4.3 評価方法

本実験ではコメントの感情度回帰・コメント数の分類共に 10 分割の交差検定を行い、回帰の評価には平均二乗誤差 (MSE) と決定係数 (R^2) を、分類の評価には F 値 (F_1) と正答率 (Acc.) を用いた。

*4 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
*5 http://scikit-learn.org
*6 http://www.csie.ntu.edu.tw/ cjlin/libsvm/

5. 実験結果と考察

5.1 感情度の回帰によるニュース影響度予測

実験 1 から実験 3 までの感情度の回帰結果は表 3 - 5 の通りである。表 3 よりニュースコンテンツの表層的素性 (SF) や時間的素性 (TM)、環境的素性 (EV) ではなく、言語的素性 (LI) がコメントの感情度に最も影響を与えているという知見が得られた。表 4 から記事本文 (LI-AR) が見出し (LI-HE) や関連リンク (LI-RL) に比べ結果に寄与が大きいこと、表 5 から文字 (LI-CH)、固有表現 (LI-NE)、感情 (LI-SE) という素性セットの中で、感情が最も評価値が高いものの、文字による結果も感情を用いた場合との差は十分小さいことがわかる。

プラットフォームの比較からは、Twitter よりも Yahoo!ニュースに投稿されたコメントの方がニュースコンテンツからコメントの感情度が予測し易いことが読み取れ、その理由としてはツイートからコメント抽出をした際に残ったノイズが Twitter での結果へ悪影響を与えた可能性の他に、Yahoo!ニュースの同一ページ内にコメントを書き込むため、投稿までの時間の短さや他サイトへの移動の少なさからコンテンツの影響を強く受け易いという可能性が挙げられる。

表 3: コメントの感情度回帰結果：実験 1

素性セット	Twitter		Yahoo!ニュース	
	MSE	R^2	MSE	R^2
SF	0.0110	0.273	0.0058	0.398
LI	0.0084	0.448	0.0030	0.690
SF+LI	0.0083	0.449	0.0030	0.693
SF+LI+TM	0.0083	0.448	0.0030	0.692
SF+LI+TM+EV	0.0084	0.446	0.0030	0.691

表 4: コメントの感情度回帰結果：実験 2

素性セット	Twitter		Yahoo!ニュース	
	MSE	R^2	MSE	R^2
LI-AR	0.0087	0.424	0.0034	0.646
LI-HE	0.0105	0.320	0.0044	0.544
LI-AR+LI-HE	0.0085	0.434	0.0031	0.684
LI-AR+LI-HE+LI-RL	0.084	0.448	0.0030	0.690

表 5: コメントの感情度回帰結果：実験 3

素性セット	Twitter		Yahoo!ニュース	
	MSE	R^2	MSE	R^2
LI-CH	0.0090	0.406	0.0037	0.616
LI-NE	0.0108	0.285	0.0051	0.477
LI-SE	0.0089	0.421	0.0036	0.640

5.2 コメント数の分類によるニュース影響度予測

実験 1 から実験 3 までのコメント数の分類結果は表 6 - 11 の通りである。上位 50%以上・未満および上位 10%以上・未満という 2 種類の分類を行った為、各実験ごとに 2 種類の結果を得ている。

表 6・7 より、コメント数の分類においても Twitter よりも Yahoo!ニュースで予測が容易であるという結果が得られた。表層的素性 (SF) と言語的素性 (LI) をそれぞれ単独で用いた場合の結果を両プラットフォームで比較すると、Twitter のコメント数の方が表層的素性の影響を比較的受け易く、Yahoo!ニュース内のコメント数の方が言語的素性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、記事の中身よりもカテゴリや配信元といった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する他人のコメントを読んだ上で自身のコメントを投稿できる為、記事に対する理解が深まっているという可能性がある。

表 8・9 からはコメントの感情度同様、コメント数においても記事本文 (LI-AR) の特徴が結果に与える影響が大きいこと、更にはコメント数が上位 10%以上になるか否かには見出しよりも記事本文の影響が大きいことがわかる。

表 10・11 からはコメント数には感情 (LI-SE) よりも文字 (LI-CH) の特徴が影響を与えており、Yahoo!ニュース内のコメント数には固有表現 (LI-NE) の影響も大きいことが読み取れる。ニュースコンテンツの感情はコメント投稿者の感情に訴える部分があるが、コメントを投稿するか、には感情よりも文字の特徴の方が重要であるようだ。

表 6: コメント数の分類結果：実験 1（境界値：上位約 50%）

素性セット	Twitter		Yahoo!ニュース	
	F_1	Acc.	F_1	Acc.
SF	0.652	64.3%	0.597	63.0%
LI	0.674	67.1%	0.697	70.8%
SF+LI	0.675	68.3%	0.707	71.6%
SF+LI+TM	0.675	68.3%	0.708	71.6%
SF+LI+TM+EV	0.674	68.3%	0.709	71.6%

表 7: コメント数の分類結果：実験 1（境界値：上位約 10%）

素性セット	Twitter		Yahoo!ニュース	
	F_1	Acc.	F_1	Acc.
SF	0.656	64.1%	0.645	62.5%
LI	0.666	66.2%	0.734	74.0%
SF+LI	0.707	70.3%	0.726	73.4%
SF+LI+TM	0.702	69.8%	0.727	73.8%
SF+LI+TM+EV	0.702	69.9%	0.730	74.0%

表 8: コメント数の分類結果：実験 2（境界値：上位約 50%）

素性セット	Twitter		Yahoo!ニュース	
	F_1	Acc.	F_1	Acc.
LI-AR	0.647	64.4%	0.689	69.3%
LI-HE	0.625	60.9%	0.680	66.5%
LI-AR+LI-HE	0.669	66.7%	0.698	70.9%
LI-AR+LI-HE+LI-RL	0.674	67.1%	0.697	70.8%

表 9: コメント数の分類結果：実験 2（境界値：上位約 10%）

素性セット	Twitter		Yahoo!ニュース	
	F_1	Acc.	F_1	Acc.
LI-AR	0.663	66.2%	0.741	74.5%
LI-HE	0.622	60.3%	0.590	62.1%
LI-AR+LI-HE	0.663	66.0%	0.741	74.6%
LI-AR+LI-HE+LI-RL	0.666	66.2%	0.734	74.0%

表 10: コメント数の分類結果：実験 3（境界値：上位約 50%）

素性セット	Twitter		Yahoo!ニュース	
	F_1	Acc.	F_1	Acc.
LI-CH	0.670	66.9%	0.678	69.8%
LI-NE	0.663	63.9%	0.690	68.9%
LI-SE	0.636	62.3%	0.658	66.2%

表 11: コメント数の分類結果：実験 3（境界値：上位約 10%）

素性セット	Twitter		Yahoo!ニュース	
	F_1	Acc.	F_1	Acc.
LI-CH	0.660	65.5%	0.738	74.5%
LI-NE	0.638	64.2%	0.659	66.1%
LI-SE	0.665	64.1%	0.659	67.2%

6. 結論

本研究ではニュースコンテンツの言語特性がユーザコメントへ及ぼす影響度を分析する手法を提案し、実験によって Twitter と Yahoo!ニュースに投稿されるコメントの感情度・コメントの数は、ニュースコンテンツの言語的素性の影響を少なからず受けていることを明らかにした。言語的素性の中でも特に影響度が大きい単語や記号などをより細かい分析を進めることによって突き止め、実用的な知見を得ることが今後望まれる。本研究が言語特性に関する研究の発展へと繋がっていくことを期待したい。

参考文献

[Bandari 12] Bandari, R., Asur, S., & Huberman, B.: The Pulse of News in Social Media: Forecasting Popularity, ICWSM, pp. 26-33(2012).

[Tsagkias 09] Tsagkias, M., Weerkamp, W., & De Rijke, M.: Predicting the volume of comments on online news stories, Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1765-1768(2009).

[小林 05] 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).

[東山 08] 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).