機械学習を用いたヘイトスピーチの分析：
Wordfish モデルによる非市民的な表現の抽出

ロンドン政治経済学院方法論学部　研究員　渡辺耕平

ウェブサイトが公的な議論の場になる可能性はインターネットが登場した早い段階から期待されてきたが(Rheingold, 2000)、２ちゃんねるなどへの投稿が「便所の落書き」と形容されるように、匿名性の高いインターネット掲示板では、感情的な投稿が多く、建設的な議論が行われにくいとも指摘されている(松村, 三浦, 柴内, 大澤, & 石塚, 2004)。近年では、オンラインニュースサイトが公的な問題について議論をする場となることが期待されているが(Ruiz et al., 2011)、特定の個人や集団に対する攻撃的な言論、いわゆる、ヘイトスピーチをどのように防ぐかが課題となっている(Erjavec & Kovačič, 2012)。今年の６月には、日本で最も人気のあるオンランニュースサイトの一つ、Yahoo ニュースが、排斥主義的な投稿を減らすために、同一の利用者が大量のコメントを投稿することを禁止した。同サイトでは、24 時間の人間と機械による投稿の監視が行われているが、中国や韓国に対する否定的なコメントが依然として多くみられる(日浦 & 村田, 2017)。

ニュースサイトに近隣諸国に対する否定的なコメントがあふれることは、東アジアにおけるぎこちない国際関係、日本におけるナショナリズムの高まりなどの政治・社会的な問題と関連していると考えられ、ヘイトスピーチを含む投稿を発見し、分析することは、社会科学者にとっても重要な課題である。しかしながら、Yahoo ニュースの運営者による投稿の選別が十分に機能していないように、大量の投稿の中から不適切なものだけを選択することは、コメントに含まれる表現の多様性から人間にとっても、機械にとっても困難な作業と言える。

本稿では、欧米の社会科学において急速に発展している計量テキスト分析の手法を用いて、上記の困難を克服する方法を提案する。本方法は欧米の政治科学者の間で以前から知られている Wordfish (Slapin & Proksch, 2008)と呼ばれる文書の尺度化 (document scaling) のための教師なし学習モデルを応用したものである。このモデルはもともと選挙マニフェストの政治的イデオロギーを推定するために開発され、政治家のスピーチの分析(Lauderdale & Herzog, 2014)にも応用されてきたが、本稿では、これをコメントの市民性(civility)を推定するために用いる。なお、本手法は、R 言語において、筆者が開発に携わるテキスト分析パッケージ quanteda (Benoit et al., n.d.)を用いて実装されているため、読者が容易に再現および転用できる。[1]

[1] 同パッケージを用いた日本語の計量テキスト分析の具体的な方法については「R による日本語テキスト分析入門」（https://github.com/koheiw/IJTA/）を参照。

以下の節では、はじめにコメントに含まれる特異語の正確なトークン化の重要性を指摘した後、Wordfish がコメントに含まれる語の極性を推定する能力を持つことを示し、推定された極性値を用いた量的分析の応用例を提示する。本稿で用いられているデータは、筆者が Yahoo ニュースを 2017 年 7 月 2 日から 15 日の 2 週間、「韓国」というキーワードでニュースを検索し、記事の本文および 24 時間以内に投稿されたコメントをダウンロードすることによって作成した。このデータは、計 90 のニュースメディアから供給された 1,458 個の記事と 90,607 件のコメントを含んでいる。

共起分析による特異語のトークン化

文書をコンピュータープログラムが効率的に処理できる形式に変換する処理は、一般的にトークン化(tokenization)と呼ばれる。ヨーロッパ言語の計量テキスト分析では、文を空白によって分割することで語をトークンとして得るが、日本語は語が空白によって区別されていないため、より高度な前処理を求められる。これまで、日本語のトークン化には Mecab や Chasen などの形態素解析ソフトが用いられてきたが、これらのプログラムは辞書に登録されていない語彙を正確に分割できない。quanteda は形態素解析ソフトに依存しないが、日本語のトークン化に辞書を用いるため、コメントに含まれる特異な表現を正確に分割できない。コメントに含まれる特異な表現は、ヘイトスピーチの分析においてとりわけ重要であると考えられるため、本稿の分析では共出現分析(collocation analysis)を用いてトークンを修正した。ここでの共出現分析には、Blaheta & Johnson (2001)によって提案されたモデルを用い、連続する漢字もしくはカタカナを抽出し、統計的に強く結びついた組を発見する。強く結びついた組は連結し、以降の分析で一つの語として取り扱う。表１は、共起分析によって発見されたカタカナもしくは漢字の組の中から最も強く結びついたものを示している。「シ ネ」（死ね）、「タカ リ」（集り）、「トン チン カン」、「売春 婦 像」、「国交 断絶」、「嫌 韓」は個々としては意味がはっきりしないが、組としてはヘイトスピーチを構成する表現であると考えらえる。これらの表現は、トークンの修正を行わない従来の分析手法では見落とされることが多い。

表１：カタカナと漢字の共起語

rank	collocation	count	lambda	z	collocation	count	lambda	z
1	シ ネ	288	9.54	86.18	韓国 人	3093	4.38	185.81
2	レッド ライン	442	11.57	78.49	慰安 婦	4634	10.20	151.57
3	ム ン	243	8.91	75.09	日 韓	1122	6.53	150.35
4	コリア ン	155	8.40	64.95	米 軍	800	6.78	133.94
5	ノ ム	89	9.80	55.74	被害 者	926	7.40	132.21
6	アン シ	79	7.95	55.00	売春 婦	1588	7.78	129.62
7	ゴール ポスト	114	11.30	54.65	婦 像	875	6.20	123.63
8	コ レ	66	8.76	50.78	経済 制裁	527	7.03	114.92

9	タカ リ	66	8.85	50.63	在 日	643	7.17	113.97
10	ダイ ハン	198	12.68	49.67	日本 政府	1042	4.58	112.18
11	スマ ホ	127	12.31	49.30	日 米	558	5.79	109.12
12	ア レ	72	9.50	48.31	婦 問題	773	4.48	108.02
13	パ ク	212	11.24	48.28	国交 断絶	893	10.63	105.80
14	レッド チーム	66	6.71	48.12	支持 率	652	10.05	102.50
15	ク ネ	62	6.62	46.43	億 円	402	8.70	101.72
16	ギャラ ク	80	9.46	45.34	自分 達	368	6.72	95.59
17	ライ ダイ	200	13.07	45.13	監督 経験	340	6.98	93.86
18	ク シー	86	9.66	44.31	東京 五輪	299	8.09	93.54
19	トン チン	39	9.59	42.78	朝鮮 戦争	338	6.23	92.84
20	チン カン	39	9.75	42.08	国家 間	312	6.43	90.63
21	ナ メ	79	11.48	40.32	国際 社会	282	7.83	88.63
22	ボ ミ	35	10.76	39.61	自 民	322	9.53	88.38
23	タイ ハン	33	8.59	38.79	間 違	383	8.95	88.22
24	ライ タイ	33	8.82	38.39	安倍 政権	298	6.13	86.65
25	ブ ン	53	8.44	37.52	米 韓	353	5.25	85.65
26	イ ク	34	7.39	36.81	再 交渉	233	7.99	85.49
27	ピン ポイント	30	11.26	36.24	嫌 韓	263	6.08	79.49
28	バレ バレ	26	8.29	35.98	旭日 旗	887	11.31	78.30
29	ヘタ レ	47	9.86	35.91	核 実験	229	7.74	77.84
30	マス ゴミ	104	11.91	35.80	再 協議	191	8.59	76.69

Wordfish による語の極性の推定

Wordfish は与えられた文書に含まれる語に対して統計的に最適な重みづけを与えることによって、文書の極性を推定する機械学習モデルである。このモデルは教師なし学習に基づくため、社会科学的に意味のある結果を常に生み出すわけではないが、本稿のデータからは Coe, Kenski, & Rains (2014)が論じたコメントの市民性―非市民性(civility-uncivility)の極性を抽出できているように思われる。彼らの定義によれば特定の個人や集団に対する誹謗および中傷、不誠実および下品な発言を含むコメントは非市民的とされる。図 1 は韓国のフランチャイズ企業におけるパワハラ問題についての記事に対するコメントを Wordfish を用いて分析した結果を示しているが、図の左側に集中している「息の根」「止める」「売春」「小汚い」「ゲス」などの語は非市民的な語であると言えるだろう。対照的に図の右側には非市民的な語は見られず、この記事のコメントにおいて Wordfish が良好に機能していることがわかる。表 2 は、10 個以上のコメントを集めた 884 個の記事に対して、個別に Wordfish を適用して、極性の平均値がもっとも強い語を示してある。表の左側の語がニュースの本文に使われるような市民的な語であるのに対し、右側の語の多くは否定的なコメントに使われることの多い非市民的な語であり、Wordfish が他の記事のコメントにおいても良好に機能していることがわかる。

しかしながら、Wordfish を利用する際は、語の極性の方向が恣意的であることに注意しなくてはならない。上述の例では、非市民的な語に負の極性が与えられていたが、Wordfish 自体は、どちらが非市民的であるかの判断をすることができず、モデルによっては非市民的な語に正の極性が与えられる。筆者は、本分析において恣意的な極性の方向を統一するために、簡易な期待値最大化(expectation–maximization)アルゴリズムを作成し、自動的にすべてのモデルの極性の方向を同じにした後に、手動で非市民的な語が負の極性を得るように調整した。このアルゴリズムは、語の極性値が全モデルを通じて最も強く相関するように、極性パラメーターの方向を反転する処理を繰り返すもので、これを 884 個のモデルに対して適用すると 10 回程度の反復を通じて、およそ半分のモデルの極性の方向が反転した。

図 1 ：Wordfish によって推定された語の極性

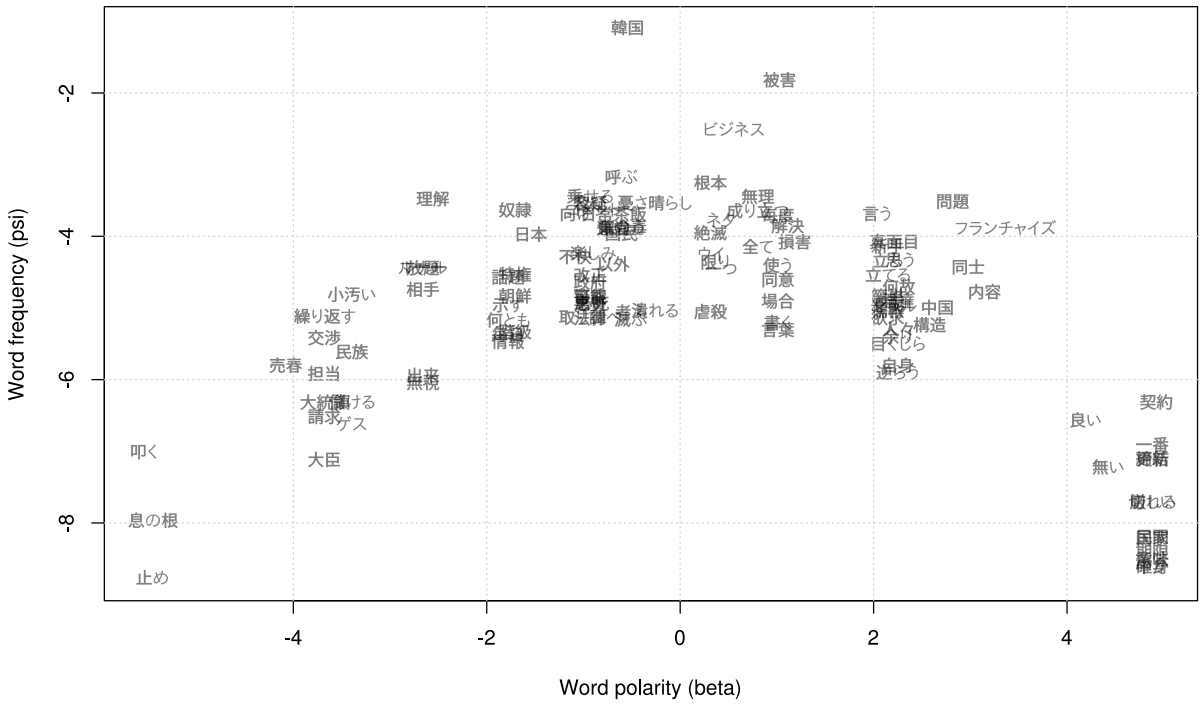


表 2 ：すべての記事を通じて最も極性が強い語

word	polarity	word	polarity
日本	0.91	大丈夫	-0.14
韓国	0.66	意味	-0.13
中国	0.41	www	-0.12
国民	0.38	早く	-0.11
世界	0.32	ダメ	-0.10
経済	0.31	妄想	-0.08
問題	0.30	無理	-0.07

米国	0.29	悪い	-0.07
政府	0.28	レッド	-0.07
に対して	0.25	笑える	-0.07
アメリカ	0.24	相手	-0.07
受け	0.24	南朝鮮	-0.07
北朝鮮	0.22	ww	-0.07
企業	0.20	バカ	-0.07
状況	0.20	相変わらず	-0.07
政権	0.19	コウモリ	-0.06
政策	0.19	期待	-0.06
政治	0.18	何で	-0.06
必要	0.18	良い	-0.06
に対する	0.18	刈り上げ	-0.06
多い	0.17	キモ	-0.06
歴史	0.17	オリンピック	-0.06
慰安	0.17	カン	-0.06
被害	0.17	無駄	-0.06
開発	0.17	お願い	-0.05
思う	0.17	電話	-0.05
ロシア	0.16	マジ	-0.05
自国	0.16	爆笑	-0.05
制裁	0.16	恥ずかしい	-0.05
結果	0.16	野郎	-0.05

Wordfish モデルから抽出された非市民的な語から、キーワード辞書を作成することで、ニュースサイトにおけるヘイトスピーチの計量分析を容易に行える。表３には、トピックモデル(LDA)を用いて判別したニュース記事の話題を示してあり、Topic 4 はスポーツ、Topic 5 と 10 は芸能、Topic 8 は従軍慰安婦に関する話題であると解釈できる。コメントに含まれる非市民的な語の頻度を話題別に集計すると図２のようになり、スポーツと芸能に関するニュース記事に多くの非市民的コメントが集まることが見て取れる。だだし、従軍慰安婦に関するニュースに非市民的なコメントが集まらないわけではなく、図３に示してあるように平均で一記事あたり 266 件と、多数のコメントが寄せられるため、非市民的な語の相対的頻度が低下していると理解すべきである。

表３：トピックモデルによって判別されたニュース記事の話題

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	大統領	米国	北朝鮮	ツアー	ソン	中国	韓国	韓国	代表	韓国
2	首脳	韓国	ミサイル	アンダー	リーグ	問題	企業	日本	監督	日本
3	北朝鮮	日本	発射	女子	韓国	thaad	市場	慰安	韓国	ファン
4	会談	貿易	弾道	アン	試合	米国	ウォン	問題	チーム	アルバム
5	トランプ	交渉	米国	日本	所属	国際	日本	女性	東芝	メンバー
6	韓国	自動車	韓国	通算	ジュンギ	ロシア	サービス	ソウル	選手	写真

7	韓米	関税	icbm	最終	ヘギョ	制裁	販売	政府	契約	グループ
8	問題	fta	実験	キム	結婚	解決	ドル	明らか	五輪	公開
9	対話	欧州	大陸	打差	事務所	主席	経済	被害	日本	公演
10	首相	eu	キロ	プロ	発表	政府	事業	調査	サッカー	デビュー
11	両国	協定	金正	ゴルフ	fc	に対する	投資	合意	sk	思い
12	ドイツ	自由	開発	出場	写真	会議	会社	受け	ハイ	今回
13	平和	経済	火星	タイ	シーズン	向け	産業	関連	アジア	ドラマ
14	共同	連合	日本	賞金	航空	地域	生産	委員	出場	人気
15	カ国	市場	攻撃	選手	キム	台湾	世界	国民	連合	ステージ
16	会議	協議	発表	試合	ファン	カ国	サムスン	歴史	交渉	映画
17	米国	政府	訓練	プレー	伝え	世界	海外	家族	最終	ツアー
18	強調	産業	委員長	韓国	ドラマ	企業	開発	事実	委員	感じ
19	に対する	要求	脅威	ハム	チーム	措置	昨年	機関	経験	時間
20	合意	合意	国防	ミニ	ミン	立場	基地	伝え	決定	出演

図２：非市民的な語の話題別の頻度

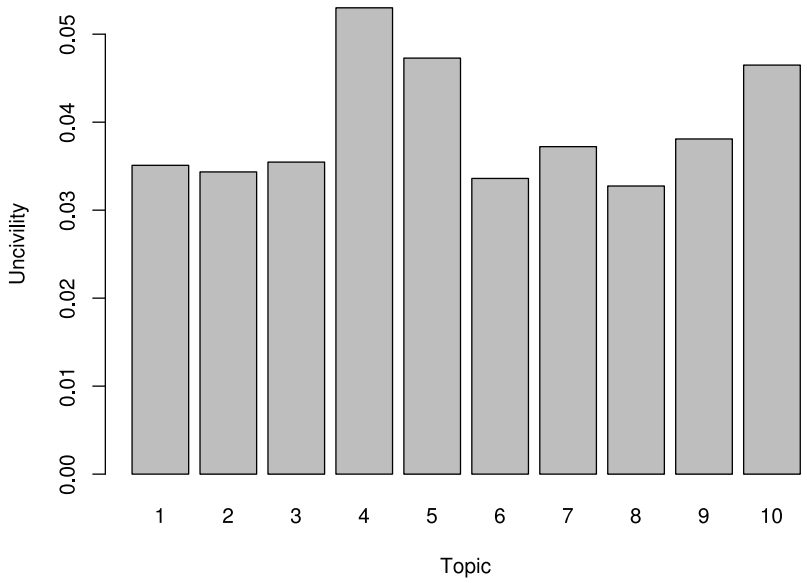
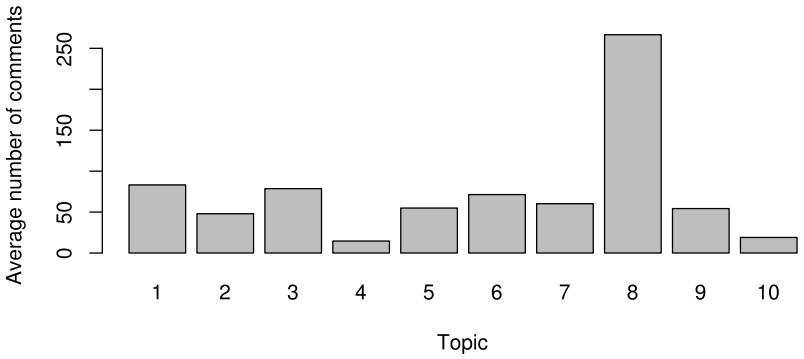


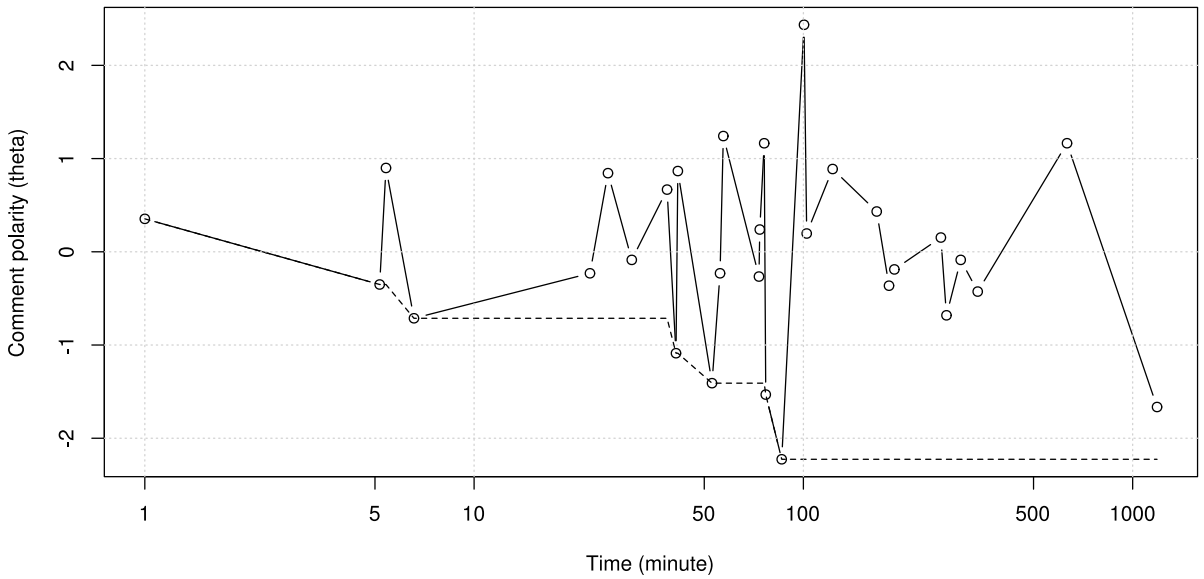
図３：記事に対するコメントの平均数



Wordfish によるコメントの極性の推定

Wordfish は、語の極性だけではなく、コメントの極性も同時に推定するため、極性パラメーターの方向を統一することで、個々のコメントについてのさまざまな分析を行うことができる。筆者は、非市民的なコメントが議論全体に及ぼす影響に関心があり、図４に示してあるようにコメントの極性の変化を分析する予定である。この図では、コメントの極性が全体としては不規則に推移しているように見えるが、下部の破線が示しているように非市民的なコメントの極性が段階的に強くなっていることがわかる。

図４：コメントの極性と投稿された時間



結論

本稿では、まず、ヘイトスピーチ分析の前処理として共起分析を行うことで、従来の計量テキスト分析の手法では見落とされがちである重要な特異語を発見し、適切にトークン化できることを示した。そして、これまでもっぱら政治イデオロギーの分析に用いられてきた Wordfish に対して、パラメーターの極性の方向を一致させるアルゴリズムを適用することで、ニュースサイトにおけるコメントの分析にも活用できることを示した。さらに、Wordfish で推定した語とコメントの極性値を用いて、ニュースの話題ごとの非市民的なコメントの分布および非市民的なコメントが議論全体に与える影響についての予備的な分析を行った。

Wordfish を用いて抽出したキーワード辞書を用いて分析した結果、スポーツと芸能関係のニュースで非市民的なコメントの頻度が相対的に高いことが示された。スポーツに関するニュースが非市民的なコメントを集めやすいのは、国際的なスポーツ大会が読者の排外的な意識を高めること、また、政治や経済についてよりも、スポーツについてコメントを書くことの方が容易であり、低い社会階層の利用者でも投稿しやすいからだと考えられる。また、芸能ニュースにおいて非市民的なコメントが多く見つかるのは、芸能人の容姿に関するあからさまな中傷が行われているからであると思われる。従軍慰安婦問題に関する記事は圧倒的に多くのコメントを集めることから、ほかのニュースに対するコメントと直接比較することが困難であり、区別して分析する必要があるように思われる。時系列の分析では、非市民的なコメントが議論全体の質に与える影響は見られなかったものの、それらの極性が段階的に強くなっていることから、非市民的な利用者が、同様の傾向を持つ者を刺激し、より過激なコメントを投稿させている可能性を見出せる。このような発言の過激化のメカニズムの解明は重要な課題であるため、より系統的な方法で大規模に分析する必要があると考えている。

参考文献

Benoit, K., Watanabe, K., Nulty, P., Obeng, A., Wang, H., Lauderdale, B., & Lowe, W. (n.d.). *quanteda: Quantitative Analysis of Textual Data*. Retrieved from http://quanteda.io

Blaheta, D., & Johnson, M. (2001). Unsupervised Learning of Multi-Word Verbs. In *Proceeding of the Acl/Eacl 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations* (pp. 54–60).

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Erjavec, K., & Kovačič, M. P. (2012). “You Don’t Understand, This is a New War!” Analysis of Hate Speech in News Web Sites’ Comments. *Mass Communication and Society*, *15*(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Lauderdale, B. E., & Herzog, A. (2014). Measuring Political Positions from Legislative Debate Texts on Heterogenous Topics. Retrieved from http://www.alexherzog.net/files/Lauderdale_Herzog_2015.pdf

Rheingold, H. (2000). *The Virtual Community: Homesteading on the Electronic Frontier*. MIT Press.

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public Sphere 2.0? The Democratic Qualities of Citizen Debates in Online Newspapers. *The International Journal of Press/Politics*, *16*(4), 463–487. https://doi.org/10.1177/1940161211415849

Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x

日浦統, & 村田悟. (2017, July 31). ネットのコメント欄に変化 規制を強化、マナーも求める：朝日新聞デジタル. 朝日新聞デジタル. Retrieved from http://www.asahi.com/articles/ASK7Z5RZXK7ZUTIL01V.html

松村真宏, 三浦麻子, 柴内康文, 大澤幸生, & 石塚満. (2004). 2 ちゃんねるが盛り上がるダイナミズム. 情報処理学会論文誌, *45*(3), 1053–1061.