2021年度（令和3年度）　卒業論文

機械学習を利用した
賛成を得やすい文章の分析

電気情報系学科 卒業論文検印	
学科長	

指導教員

村田真樹
村上仁一

鳥取大学工学部　電気情報系学科

自然言語処理研究室

B18T2111B　三木 謙志

概要

本研究はYahoo!ニュースのコメント欄を利用して賛成を得やすい文章の特徴の発見を目的とする．賛成を得やすい文章とはYahoo!ニュースの同じ記事に対してコメントされた文章を比較し，賛成した人数がより多く，コメント時刻がより最近のコメントを賛成を得やすい文章だと定義している．比較した時の賛成を得やすい文章と賛成を得にくい文章を文章対として機械学習に入力し，どちらが賛成を得やすい文章かを推定させた．推定させた結果，BERT，最大エントロピー法，SVMの順で正解率が高くなり，一番正解率が高いBERTで0.7506となった．素性分析の結果，賛成を得やすい文章の素性だと納得できる素性の発見はできていないが,「バカ」や「幼稚」などの賛成を得にくい文章の素性だと納得できる素性はいくつか発見した．

目 次

表 目 次

第1章　はじめに

本研究はYahoo!ニュースのコメント欄を利用して賛成を得やすい文章の特徴の発見を目的とする．賛成した人数とコメント時刻の情報を使用し，同じ記事に対する2つのコメントのどちらが賛成を得やすい文章かを機械学習を用いて推定させる．自動分類のみならず，なぜ自動分類できたかの理由を素性分析の技術を利用し，賛成を得やすい文章にどのような特徴があるかを分析する．この分析は賛成を得やすい文章の作成につながると考えている．

本研究の主な主張点を以下に整理する．

- 本研究ではME，SVM，BERTの機械学習を利用して実験を行った結果，BERT＞ME＞SVMの順で性能が高くなった．一番性能が良かったBERTの正解率は0.7506となり，一番低いSVMは0.6734となった．
- 機械学習での素性分析を行った結果，賛成を得やすい素性として「うーん」,「予防」,「現場」などが得られた．しかし，これらの単語が要因で賛成を得やすい文章になっているとは考えにくい．逆に，賛成を得にくい素性として得られた「バカ」,「幼稚」などは賛成を得にくい素性だと考えることができる．
- 文字数の素性分析の結果，賛成を得やすい文章は100文字や200文字より大きい文字数の文章が良いとされた．これは賛成を得やすい文章にはコメントの説明にある程度の文字数が必要だと考えることができる．

本論文の構成は以下の通りである．

第2章ではこれまでの関連する研究を説明する．

第3章では本研究における，賛成を得やすい文章の判定方法と判定に利用する技術を説明する．

第4章では賛成を得やすい文章の自動判定とその評価を行う．

第5章では考察を行い，効果的な利用方法を考察する．

第6章ではまとめを行う．

第2章　先行研究

本章では，先行研究について記述する．

2.1 節では，石黒 [1] が行った日本語学習者の作文における文章構成と説得力の関係の研究について記述する．

2.2 節では，端ら [2] が行った感動を与える文の自動取得と分析の研究について記述する．

2.3 節では，村田ら [3] が行ったユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析の研究について記述する．

2.4 節では，2.1 節，2.2 節，2.3 節以外の先行研究を簡単に記述する．

2.1　日本語学習者の作文における文章構成と説得力の関係

石黒 [1] は，日本語学習者が執筆した作文の文章構成と，その作文が有する説得力の高さとの関係を調査している．文章構成については，段落分けと文章型の観点から検討を行った．段落分けについては，一定の話題のレベルで区切るようにし，段落構成に一貫性を持たせること，段落分けを細かくしすぎないようにし，話題のまとまりを明確にすることが重要であることを示した．また，文章型については，超級レベルに達していない学習者には，両括型か頭括型でしっかり書けるように指導するのが望ましいこと，尾括型を使う場合には，冒頭部で焦点となる疑問や観点をあらかじめ示すと読みやすくなり，分括型は主題文を繰り返すのではなく，主題文を間接的な表現に言い換えるようにすると，文脈効果が高まり説得力のある文章になること，中括型や潜括型は一般に避けたほうがよいことを示した．

2.2　感動を与える文の自動取得と分析

端ら [2] は，人は日々感動を求め，感動によって動かされる生き物であるということから「感動」と「文」に重きを置き，感動を与える文か否かの自動判定に関する研究

を行うことによって，感動を与える文の作成支援を行った．自動判定は教師有り機械学習であるSVMやパターンマッチングによって行っている．また，感動を与える文で多く使われる単語を収集することで，感動を与える文の言語的特徴を明らかにしている．

その結果，感動を与える文に多く出現する単語として,「人生」,「人々」,「幸福」,「友情」,「青春」,「恋愛」などが得られた．これらを用いた文は感動的な文を作成する際に役立つものと考えられる．

2.3 ユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析

村田ら [3] は，情報の重要度を決める要因を明らかにし，その知見に基づき情報の重要度を自動推定するシステムを構築することを目標に一般の人が重要と考える情報の分析につながるようにアンケート調査に基づく情報の重要度の研究を行った．どのような情報を重要と考えるかは個々人によって異なるため，本研究ではユーザ個人ごとに異なる情報の重要度について特に焦点をあてて分析した．ユーザごとの興味をアンケートにより抽出しその結果を利用してユーザごとに異なる情報の重要度について調査を行った．実験の結果，60%以上の被験者が重要と考える一般的な情報の重要度は，80%以上の精度でもとめることができることがわかった．また，ユーザ個人の考える情報の重要度は約65%で推定できた．興味情報が機械学習で重要とされた上位500個の単語の方と有意に重なりが多かった被験者は53人で，下位500個の単語の方が重なりが多かった被験者は2人であった．53人と2人は検定で有意差があるため，ユーザ個人の興味情報が，そのユーザの重要な記事の判断と相関があることがわかった．

2.4 種々の先行研究

文献 [4]，文献 [5]，文献 [6] はSNSで拡散されやすい文章を研究している．研究結果として，文章が述べている分野に関係なく，ポジティブ感情を示す絵文字を使用している文章よりもネガティブ感情を示す絵文字を使用している文章のほうが拡散されやすく，拡散されるスピードも早いという結果が得られている．絵文字のみならず，全体的にネガティブな文章のほうが拡散されやすいという結果も得られている．また，多くの人の共通認識や固定観念などが書かれている文章も拡散されやすくなっている．文

献 [7] では，政治のスピーチで説得力のある発話にタグを付与したコーパスを作成している．文献 [8] では，説得を行う文章の構造に関する調査をしている．説得力のある文章は，賛成を得やすい文章と類似するため，これらの研究は本研究に関連する．しかし，これらの研究は，文章構造や文章のパターンに基づき説得力について分析を行っている．他にも似たような研究として村田らの研究 [9] [10] も存在している．

第3章　提案手法

本章では，本研究の提案手法の説明を記述する．

3.1 説では，提案手法の大まかな流れについての説明を記述している．

3.2 節では，本研究で使用する機械学習法である最大エントロピー法 (ME) についての説明を記述している．

3.3 節では，本研究で使用する機械学習である ME で求まるα値を正規化した正規化α値についての説明を記述している．

3.4 節では，本研究で使用する機械学習法である SVM についての説明を記述している．

3.5 節では，本研究で使用する機械学習法である BERT についての説明を記述している．

3.6 節では，機械学習に入力するデータの作成方法についての説明をしている..

3.7 節では，機械学習での推定方法についての説明をしている．

3.8 節では，機械学習で使用する素性について記述している．

3.1　提案手法

本研究では賛成を得やすい文章と賛成を得にくい文章の二つを利用して文章対を作成し，ME，SVM，BERT の三種類の機械学習に入力し，どちらが賛成を得やすいかを推定させる．自動分類のみならず，なぜ自動分類できたかの理由を素性分析の技術を利用し，賛成を得やすい文章にどのような特徴があるかを分析する．この分析は賛成を得やすい文章の作成につながると考えている．

3.2　ME

ME とは, あらかじめ設定しておいた素性 $f_j(1 \leq j \leq k)$ 集合を F とするとき, 式 (3.1) を満足しながらエントロピーを意味する式 (3.2) を最大にするときの確率分布

$p(a,b)$ を求め, その確率分布にしたがって求まる各分類の確率のうち, もっとも大きい確率値を持つ分類を求める分類とする方法である.

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad for \quad \forall f_j(1 \leq j \leq k) \tag{3.1}$$

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)log(p(a,b)) \tag{3.2}$$

ただし, A, B は分類と文脈の集合を意味し, $g_j(a,b)$ は 文脈 b に素性 f_j があってなおかつ分類が a の場合 1 となり, それ以外で 0 となる関数を意味する. また, $\tilde{p}(a,b)$ は, 既知データでの (a,b) の出現の割合を意味する.

式 (3.1) は, 確率 p と出力と素性の組の出現を意味する関数 g をかけることで 出力と素性の組の頻度の期待値を求めることになっており, 右辺の既知データにおける期待値と, 左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として, エントロピー最大化 (確率分布の平滑化) を行って, 出力と文脈の確率分布を求めるものとなっている.

3.3 正規化 α 値

正規化 α 値とは，ME で求まる α 値を全分類先での合計が 1 となるように正規化した値である．また，素性 a と分類先 b の対によって定まる値であり，素性 a のみが適用される場合に分類先 b となる確率に相当する．各素性の．分類先ごとに与えられた正規化 α 値が高いほど，その分類先であることを推定するのに重要な素性であることを意味する．

3.4 SVM

SVM は，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔 (マージン) が大きいものほどテストデータで誤った分類を

する可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行う．

3.5 BERT

BERT は，Bidirectional Encoder Representations from Transformers の略で,「Transformer による双方向のエンコード表現」と訳され，2018 年 10 月に Google の Jacob Devlin らの論文 [11] で発表された自然言語処理モデルである．従来の機械学習では，大量のラベルのついたデータを用意させ，処理を行うことで課題に取り組む．しかし従来の手法に対し，BERT は事前学習でラベルのないデータをはじめに大量に処理を行う．その後，ファインチューニングで少量のラベルのついたデータを使用することで課題に対応させる．

3.6 データ作成

データの作成には Yahoo!ニュースのコメント欄にある，時刻と賛成した人数の情報を利用する．2021 年 9 月〜12 月の期間に Yahoo!ニュースの 1,000 コメント以上投稿された記事を対象に，コメント，賛成数，時刻の情報を 300 記事分収集している．同じ記事に投稿された 2 つのコメントにおいて，賛成した人数がより多く，コメント時刻がより最近という 2 点を満たすコメントを賛成を得やすい文章，もう一方を賛成を得にくい文章だと定義する．そのような賛成を得やすい文章と賛成を得にくい文章を文章対として大量に作成する．ただし，一度でも文章対を作成する際に使用されたコメントは他の文章対では使用しないこととする．例を以下に掲載する．

- コメント A，賛成：4,234，コメント時刻:4 時間前
- コメント B，賛成：3,823，コメント時刻:2 時間前
- コメント C，賛成：6,923，コメント時刻:8 時間前
- コメント D，賛成：2,182，コメント時刻:5 時間前

コメント A を対の一方として文章対を作成する場合，定義に当てはめると文章対として使用できるもう一方はコメント D のみである．コメント B を対の一方として利用する場合も同様に文章対として使用できるもう一方はコメント D のみであり，コメン

トCは賛成した人数は一番多いがコメント時刻が一番古いためどのコメントとも文章対を作成することはできない.

よって，この4つのコメントで定義に当てはまり，文章対として使用できるコメント対はコメントAとコメントD，コメントBとコメントDのみである．この時，賛成を得やすい文章はコメントA，コメントBとなり，コメントDは賛成を得にくい文章となっている．しかし，本研究では文章対作成の際に一度使用したコメントは利用しないので，文章対として使用できるのはコメントAとコメントDの文章対かコメントBとコメントDの文章対のどちらか一つの文章対となっている．このように定義にあてはまる文章対をコメントデータから大量に作成する.

3.7 推定方法

作成した文章対の一方の文章を左側，もう一方を右側として，賛成を得やすい文章が左側の文章なら文章対の先頭に“L”を，賛成を得やすい文章が右側の文章なら文章対の先頭に“R”を付与したものを学習データとする. このような学習データである文章対を入力とし,“L”,“R”を出力とする. 入力が与えられるとそれに対する出力を推定できるように3種類の機械学習 (ME，SVM，BERT) で学習する.

3.6節の文章対を学習データとした場合の例を以下に示す. 例のように作成した文章対の”L”と”R”と文章対の左右を反転させたものも同時に学習データとして機械学習に入力する.

- L, コメントA, コメントD
- R, コメントD, コメントA

例のような学習データを機械学習で上述の文章対の場合に“L” or “R”となるように学習する．そして別の文章対を入力しその場合の“L” or “R”を推定する．左側の文章が賛成を得やすいと判断した場合は“L”を，右側の場合は“R”を出力する.

3.8 素性

ME，SVMでは文章対において左側の文章にある単語は「L:単語」，右側にある単語は「R:単語」と「コメントの文字数」を素性として利用する．BERTに入力する際は素性の入力は不要なので文章対のみを入力している.「コメントの文字数」の素性は

「L:〜以下」,「L:〜より大きい」,「R:〜以下」,「R:〜より大きい」としており，〜には「10」,「20」,「50」,「100」,「200」,「500」,「1,000」のいずれかの数字がコメントの文字数に応じて入る．また,「コメントの文字数」を素性とせずに単語だけを素性とする実験も行っている．

ME は正規化 α 値，SVM は分離平面を用いて素性分析を行う．ME では，正規化 α 値の高いものが重要な素性となる．SVM では「L:単語」などの 1 単語を入力し，分離平面からの距離が大きいものが重要な素性となる．文章対を用いた BERT の素性分析は困難なため現時点では行えていない．

第4章　実験

本章では，本研究で行った実験の説明を記述する．

4.1 節では，実験を行った際のデータの内容と実験結果についての説明を記述している．

4.2 節では，ME の素性分析の結果についての説明を記述している．

4.3 節では，SVM の素性分析の結果についての説明を記述している．

4.4 節では，ME と SVM の素性分析の結果から得た情報をまとめて記述している．

4.5 節では，機械学習が誤った推定をした例をいくつか掲載している．

4.6 節では，ME，SVM，BERT の条件による有意差についての説明を記述している．

4.1　推定実験

1,000 コメント以上を持つ記事を対象に 300 記事分のコメントを収集し，同記事内の賛成した人数とコメント時刻の 2 つの条件を満たすコメントから文章対を作成している．対とするコメントの組み合わせは賛成を得やすい文章の定義を満たすコメントからランダムに決定している．

作成した 16,342 組の文章対のうち，8,172 組を学習データ，8,170 組をテストデータとする．BERT では 8,172 組の学習データのうち，2,045 組を検証データ，6,129 組を訓練データとして実験を行っている．表 4.1 は機械学習により“L” or “R”のどちらの文章が賛成を得やすいかを推定したときの正解率を示している．

表 4.1: 機械学習の性能評価

文字数素性	有	無
ME	0.7215	0.7209
SVM	0.6788	0.6734
BERT		0.7506

表 4.1 に示されているように BERT，ME，SVM の順で正解率が高くなっていることがわかる．一番性能が高い BERT で 0.7506 という正解率を得た．また，今回の実験では ME，SVM ともに文字数の素性の有無で性能が大きく変化することはなかったが文字数の素性を利用するほうが正解率がわずかに高くなった．

4.2 ME の素性

ME の素性分析の結果で得られた賛成を得やすい素性，賛成を得にくい素性を上位 30 個並べる．コメントの文字数の素性は上位 10 個並べる．

表 4.2: ME の素性

順位	賛成を得やすい素性	α 値	賛成を得にくい素性	α 値
1	うーん	0.8656	バカ	0.1001
2	予防	0.8598	ケース	0.1488
3	人材	0.8459	わ	0.1568
4	現場	0.8459	死刑	0.1677
5	杏	0.8322	バラマキ	0.1777
6	任意	0.8298	基準	0.1777
7	進ん	0.8264	こいつ	0.1793
8	類	0.8260	幼稚	0.1799
9	持た	0.8220	理不尽	0.1801
10	宮内庁	0.8146	育成	0.1816
11	サントリー	0.8088	ヤフコメ	0.1884
12	立場	0.8087	奴	0.1886
13	なかなか	0.7999	やん	0.1907
14	過ぎる	0.7982	くん	0.1943
15	国家	0.7949	ゴリ押し	0.1954
16	現状	0.7931	もん	0.1965
17	何で	0.7900	創価学会	0.1972
18	によって	0.7838	数	0.1985
19	まし	0.7827	貴方	0.1985
20	意見	0.7818	なさい	0.2030
21	支給	0.7798	迷惑	0.2069
22	とき	0.7758	お前	0.2093
23	ホント	0.7718	代わり	0.2097
24	準備	0.7717	無職	0.2116
25	市民	0.7715	反する	0.2122
26	すごい	0.7691	コイツ	0.2131
27	認める	0.7690	いちいち	0.2136
28	負担	0.7685	よかっ	0.2169
29	立候補	0.7685	結構	0.2169
30	調査	0.7666	たく	0.2179

表 4.3: ME の素性 (文字数素性無し)

順位	賛成を得やすい素性	α 値	賛成を得にくい素性	α 値
1	うーん	0.8944	バカ	0.1056
2	人材	0.8410	ケース	0.1517
3	現場	0.8393	わ	0.1598
4	予防	0.8382	死刑	0.1733
5	任意	0.8252	幼稚	0.1752
6	杏	0.8226	育成	0.1827
7	宮内庁	0.8186	理不尽	0.1841
8	進ん	0.8145	こいつ	0.1847
9	持た	0.8109	奴	0.1889
10	類	0.8106	やん	0.1899
11	過ぎる	0.8053	バラマキ	0.1922
12	サントリー	0.8046	基準	0.1934
13	立場	0.8039	貴方	0.1954
14	国家	0.7910	ヤフコメ	0.1986
15	なかなか	0.7882	ゴリ押し	0.2013
16	まし	0.7830	もん	0.2061
17	現状	0.7826	くん	0.2083
18	意見	0.7814	反する	0.2086
19	によって	0.7798	お前	0.2121
20	ホント	0.7796	創価学会	0.2122
21	とき	0.7793	数	0.2128
22	懲り	0.7740	なさい	0.2143
23	支給	0.7730	代わり	0.2157
24	家庭	0.7689	実験	0.2159
25	立候補	0.7677	たく	0.2173
26	負担	0.7673	無職	0.2179
27	やたら	0.7672	迷惑	0.2184
28	市民	0.7666	よかっ	0.2206
29	準備	0.7638	予約	0.2218
30	名前	0.7636	コイツ	0.2220

表 4.4: ME の素性 (文字数素性のみ)

順位	賛成を得やすい素性	α 値	賛成を得にくい素性	α 値
1	L100 より大きい	0.5869	L100 以下	0.4130
2	R100 より大きい	0.5863	R100 以下	0.4138
3	R20 以下	0.5583	R20 より大きい	0.4418
4	L20 以下	0.5563	L20 より大きい	0.4436
5	L200 より大きい	0.5457	L200 以下	0.4542
6	R200 より大きい	0.5456	R200 以下	0.4545
7	R50 以下	0.5196	R50 より大きい	0.4806
8	L50 以下	0.5183	L50 より大きい	0.4816
9	L10 より大きい	0.5100	L10 以下	0.4898
10	R10 より大きい	0.5095	R10 以下	0.4906

表 4.2 と表 4.3 から文字数の素性の有無では賛成を得やすい文章，賛成を得にくい文章の素性にはあまり変化が見られなかった．ME の文字数素性だけを抜き出したのが表 4.4 である．表 4.4 からは 100 字より多い文字数の文章が一番賛成を得やすい文量だと読み取れる．

4.3 SVM の素性

SVM の素性分析の結果で得られた賛成を得やすい素性，賛成を得にくい素性を上位 30 個並べる．コメントの文字数の素性は上位 10 個並べる．

表 4.5: SVM の素性

順位	賛成を得やすい素性	マージン	賛成を得にくい素性	マージン
1	予防	1.1266	幼稚	-1.0790
2	現場	1.0976	基準	-1.0733
3	認める	0.9750	創価学会	-1.0163
4	任意	0.9528	バカ	-0.9893
5	限界	0.9017	泥棒	-0.9803
6	うーん	0.8870	問う	-0.9716
7	すみ	0.8752	期日	-0.9410
8	子育て	0.8462	育成	-0.9268
9	経費	0.8377	ゴリ押し	-0.9227
10	製品	0.8154	失脚	-0.9087
11	実態	0.7956	ズレ	-0.9049
12	応援	0.7781	余っ	-0.8911
13	済む	0.7756	ケース	-0.8568
14	意地	0.7716	実験	-0.8329
15	宮内庁	0.7692	代わり	-0.8295
16	疾患	0.7621	予約	-0.8265
17	実際	0.7565	搬送	-0.8227
18	一連	0.7448	既に	-0.8141
19	分配	0.7399	いちいち	-0.8051
20	なし崩し	0.7377	貴方	-0.8014
21	守ろ	0.7272	ガソリン	-0.7860
22	すげ	0.7240	早め	-0.7859
23	持た	0.7183	歴代	-0.7810
24	自治体	0.7152	パート	-0.7792
25	自動車	0.7147	うっ	-0.7576
26	類	0.7023	格好	-0.7514
27	とき	0.6987	みなさん	-0.7383
28	見直し	0.6974	理不尽	-0.7344
29	先ず	0.6936	絡み	-0.7341
30	人材	0.6905	やん	-0.7333

表 4.6: SVM の素性 (文字数素性なし)

順位	賛成を得やすい素性	マージン	賛成を得にくい素性	マージン
1	現場	1.1244	幼稚	-1.0838
2	予防	1.0615	基準	-1.0281
3	任意	0.9417	バカ	-1.0081
4	認める	0.9054	泥棒	-0.9929
5	限界	0.8718	問う	-0.9911
6	すみ	0.8662	余っ	-0.9617
7	うーん	0.8660	創価学会	-0.9500
8	子育て	0.8354	期日	-0.9172
9	意地	0.8146	失脚	-0.9165
10	製品	0.8120	育成	-0.8986
11	実態	0.8103	ズレ	-0.8868
12	済む	0.7955	実験	-0.8842
13	応援	0.7919	ゴリ押し	-0.8752
14	宮内庁	0.7846	ケース	-0.8660
15	実際	0.7740	予約	-0.8516
16	経費	0.7543	代わり	-0.8254
17	守ろ	0.7391	歴代	-0.8075
18	見直し	0.7362	貴方	-0.8055
19	疾患	0.7283	既に	-0.7908
20	とき	0.7249	早め	-0.7906
21	分配	0.7194	格好	-0.7681
22	人材	0.7168	やん	-0.7640
23	上手い	0.7130	いちいち	-0.7572
24	一連	0.7081	みなさん	-0.7488
25	なし崩し	0.7042	搬送	-0.7481
26	自治体	0.6996	理不尽	-0.7384
27	払う	0.6962	奴	-0.7274
28	同調	0.6955	パート	-0.7240
29	すげ	0.6950	お前	-0.7149
30	自動車	0.6859	絡み	-0.7124

表 4.7: SVM の素性 (文字数素性のみ)

順位	賛成を得やすい素性	マージン	賛成を得にくい素性	マージン
1	L200 より大きい	0.1507	L200 以下	-0.1507
2	R200 より大きい	0.1507	R200 以下	-0.1507
5	L100 より大きい	0.1205	L100 以下	-0.1205
6	R100 より大きい	0.1205	R100 以下	-0.1205
3	L20 以下	0.0830	L20 より大きい	-0.0830
4	R20 以下	0.0830	R20 より大きい	-0.0830
7	L50 以下	0.0458	L50 より大きい	-0.0458
8	R50 以下	0.0458	R50 より大きい	-0.0458
9	L10 より大きい	0.0368	L10 以下	-0.0368
10	R10 より大きい	0.0368	R10 以下	-0.0368

表 4.5 と表 4.6 から ME と同様に文字数の素性の有無では賛成を得やすい文章，賛成を得にくい文章の素性にはあまり変化が見られなかった．SVM の文字数素性だけを抜き出したのが表 4.7 である．表 4.7 からは 200 字より多い文字数の文章が一番賛成を得やすい文量だと読み取れる．

4.4 素性分析

本研究では素性分析を用いて賛成を得やすい文章の重要素性の獲得をする．4.2 節と 4.3 節の表 4.2 から表 4.7 に ME，SVM の機械学習で獲得できた単語の上位 30 素性と文字数の素性の上位 10 素性を掲載している．素性分析の結果，良いとされた素性は「うーん」,「予防」,「現場」などが ME，SVM の上位 30 素性に共通して現れている．しかし，これらの単語が文章中に入っていることが要因で賛成を得やすい文章になっているとは考えにくい．逆に ME，SVM の悪い素性に共通して現れている「バカ」,「幼稚」などはマイナスな意味を持つ単語なので，これらの単語が含まれている文章は誰かを批判している文章の可能性が高く，賛成を得にくいのではないかと考えることができる．

また，ME では 100 文字より大きい文字数の文章が賛成を得やすいとされており，SVM では 200 文字より大きい文字数の文章が一番賛成を得やすいとされている．このことから，賛成を得やすい文章にはある程度の文章量が必要となっていることがわか

る．これは自分の意見の根拠を書くのにある程度の文字数が必要になってくるからではないかと考えることができる．

4.5　誤り例

本節では，以下に示すような 2 文を文章対とし ME でどちらが賛成を得やすいかを推定させた結果，間違えた例をいくつか紹介する．

以下の例では，賛成を得やすい文章は 1 なのだが，ME では 2 が賛成を得やすい文章だと判断された．

- 1. 日本以外の先進国ではそうかも知れないけど、ＣＯＣＯＡ一つ満足に動かせない日本では無理
- 2. またそれ用のアプリを作るの？

以下の例でも，賛成を得やすい文章は 1 なのだが ME では 2 が賛成を得やすい文章だと判断された．

- 1. 金額を見て、特養のあまりの安さにビックリしました。金額が上がったのはお気の毒ですが、他の方も言っておられるように今までが「安過ぎ」たのです。うちの母は特養に入れずサービス付き高齢者住宅と小規模多機能施設が併設されたところへ入所させていただきましたが、毎月の支払いは余裕で２０万超えてました。母の年金だけでは足りず不足分は私が補填し支払っていましたが、私が支払った補填分程度で特養に入れるんだと知ると、特養とそれ以外の施設の金額の開きの大きさに二度ビックリです
- 2. そりゃ、我々職員の給料も上がりませんよ。

以下の例では，賛成を得やすい文章は 2 なのだが ME は 1 が賛成を得やすい文章だと判断している．

- 1. これでまた信者が増えますね。
- 2. 所得上限を設けるそうですが、960 万円の所得でも子供が１人なのか、子供が４人なのかでも生活にかかるお金や状況は違いますよね？

SVM，BERT も誤り例は存在するが，ここでは ME が賛成を得やすい文章の推定を間違った例のみを紹介している．これらの誤り例は ME が賛成を得やすい文章の推定を間違った文章対の中からランダムに選んだ文章対を掲載している．

4.6 有意差検定

本節では ME，SVM，BERT の使用する素性などによって有意差が存在するのかについて説明する．表 4.8〜表 4.10 内にある数値は p 値を示し，p 値が 0.05 未満なら有意差有りと判断する．

表 4.8: 有意差検定 1

有意差	ME	SVM	BERT
ME		0.000	0.000
SVM	0.000		0.000
BERT	0.000	0.000	

表 4.9: 有意差検定 2

有意差	ME(文字数素性なし)	SVM(文字数素性なし)	BERT
ME(文字数素性なし)		0.000	0.000
SVM(文字数素性なし)	0.000		0.000
BERT	0.000	0.000	

表 4.10: 有意差検定 3

有意差	ME	SVM
ME(文字数素性なし)	0.391	0.000
SVM(文字数素性なし)	0.000	0.006

表 4.8 では，文字数素性有りの ME，SVM と BERT は全ての組み合わせで有意差が存在し，表 4.9 でも同様に文字数素性無しの ME，SVM と BERT は全ての組み合わせで有意差が存在する．

表 4.10 では文字数素性有りの ME と SVM，文字数素性なしの SVM と文字数素性有りの ME，SVM は有意差が存在するが，文字数素性無しの ME と文字数素性有りの ME は有意差が存在しない．これは文字数素性無しの ME と文字数素性有りの ME は文字数の素性の影響をあまり受けていないということになり，SVM は文字数素性の影響で有意差が存在するということになる．

第5章　考察

表 4.2〜表 4.7 にあるように賛成を得やすいとされた素性である「うーん」,「予防」,「現場」が使用された，賛成を得やすいコメントを掲載する．

- うーん．政治家とかの不法行為などはわかるんだけど，芸能人のプライベートを勝手に撮影して報道して良いとされているのはなぜ？スポーツ選手とかもそうだけど．
- 経口治療薬が出て来たらファイザーもモデルナ等予防薬はオワコン
- トイレの施工がきちんと管理出来たら一流の現場監督だと言われるくらいトイレは難しい．

上記は一例ではあるが，これらの単語が原因で賛成を得やすい文章になっているとは考えにくい．本実験では 1,000 コメント以上投稿された 300 記事を収集し，文章対を作成して機械学習に入力しているがデータ数が少ないためこのように賛成を得やすい文章の特徴の発見ができなかったのではないかと考える．

次に賛成を得にくいとされた素性である「バカ」,「幼稚」が使用されたコメントを掲載する．

- 地球はバカが多いから他の平和な星で暮らしたいです．
- 話し方も内容もあまりに幼稚で驚いた．

上記も一例ではあるが，これらの単語を用いた文章は皮肉や誰かを批判する文章になるため賛成が得にくい文章になっているのではないかと考える．

文章の文字数は ME，SVM ともに 100 文字や，200 文字のコメントが賛成を得やすい素性となっているので賛成を得やすい文章には自らの意見を伝えるためにある程度の文章量が必要になっているのではないかと考える．逆に，文章が長すぎても賛成を得やすいとならないのは Yahoo！ニュースのコメント欄では一目見たときに長すぎて，読む気を失っているのではないかと考える．

以上のことをまとめると賛成を得やすい文章を作成する際には以下の点を注意する必要があると考えることができる．

- バカや幼稚などの単語のみで人をバカにする意味がある単語の使用を避ける．
- 100 文字から 200 文字くらいに文章をまとめて作成する．

この 2 点を踏まえて，文章の作成を行うことで賛成を得やすい文章の作成が容易に可能になるのではないかと考える．

第6章　おわりに

本研究では賛成を得やすい文章の特徴の発見を目的としている．Yahoo!ニュース内にある 1,000 コメント以上持つ 300 記事からコメント時刻と賛成した人数の情報を使用し，同じ記事に対する 2 つのコメントを比較する．その際にコメント時刻がより最近であるコメント，賛成した人数がより多いコメントの 2 点を満たすコメントを賛成を得やすい文章だと定義する．

この 2 点を満たすコメントで文章対を大量に作成し，どちらの文章が賛成を得やすいかを教師あり機械学習を利用し推定させた．推定させた結果，BERT，ME，SVM の順で正解率が高くなった．一番性能が高い BERT の正解率が 0.7506 となった．素性分析の結果では，賛成を得やすいと納得できる素性は得られなかった．逆に，賛成を得にくいとされた「バカ」,「幼稚」などの素性は皮肉や誰かを批判する文章になるため賛成を得にくい文章になっていると考えられる．また．文字数の素性分析の結果，賛成を得やすい文字数は 100 文字や 200 文字より大きい文字数が良いとされた．これは賛成を得やすい文章にはコメントの説明にある程度の文字数が必要だと考えられる．

今後は分野によって良い素性は変化するかなどの分野依存性や高評価だけでなく定義の部分で低評価数も考慮した賛成を得やすい文章の分析を行い，今回の研究と比較し，違いの発見などを目指す．

謝辞

また，研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部知能情報工学科自然言語処理研究室の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授に心から御礼申し上げます．その他様々な場面で御助言を頂いた自然言語処理研究室の皆様に感謝の意を表します．

参考文献

[1] 石黒圭. 日本語学習者の作文における文章構成と説得力の関係. 一橋大学国際教育センター紀要, Vol. 8, pp. 3–14, 2017.

[2] 端大輝, 村田真樹, 徳久雅人. 感動を与える文の自動取得と分析. 言語処理学会第 18 回年次大会, pp. 303–306, 2012.

[3] 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 鳥澤健太郎. ユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析. 言語処理学会第 15 回年次大会, pp. 554–557, 2009.

[4] 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 the 3rd international web science conference*, pp. 1–7, 2011.

[5] Sho Tsugawa and Hiroyuki Ohsaki. Negative messages spread rapidly and widely on social media. In *Proceedings of the 2015 ACM on conference on online social networks*, pp. 151–160, 2015.

[6] 田中友理, 宮本聡介, 唐沢穣. ステレオタイプ情報はよりリツイートされるか? ツイートの言語表現に注目した検討. 人間環境学研究, Vol. 14, No. 2, pp. 165–170, 2016.

[7] Marco Guerini, Carlo Strapparava, and Oliviero Stock. Corps: A corpus of tagged political speeches for persuasive communication processing. *Journal of Information Technology & Politics*, Vol. 5, No. 1, pp. 19–32, 2008.

[8] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 46–56, 2014.

[9] 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 松岡雅裕, 井佐原均. 情報の重要度を決める要因の抽出・分析と重要度の自動推定. 言語処理学会第 14 回年次大会, pp. 907–910, 2008.

[10] Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Qing Ma, Ryo Nishimura, Yasuhiko Watanabe, Kouichi Doi, and Kentaro Torisawa. Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction. *Cognitive Computation, Volume 2, Issue 4*, pp. 272–279, 2010.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.