令和５年

修士論文草稿

タイトル

指導教員：相馬 隆郎

東京都立大学大学院

電子情報システム工学域

学修番号：22861651

氏名：西原涼介

論文要旨

ここから論文要旨

目 次

第 I 部
はじめに

1　研究背景

近年, Amazon や楽天市場などの大手 EC サイトをはじめ、数多くの EC サイトが普及し, その利用者も急増している. そして, 商品を購入する際に EC サイトのレビューを参考にしている利用者の割合は約 70%と言われていて, その中でもレビューの信頼性を重要視している人が多いことが明らかになっている. また, 多くの企業にとって, EC サイトのレビューからユーザーの嗜好や意見を分析し，マーケティングに活用することが重要な課題となっている．そのため，EC サイトのレビューの信頼性や参考になるかどうかを評価する評判分析や口コミ分析, レビューを様々なトピックに分類する文書分類に関する研究が多く行われている. 例えば, 関連研究の項で詳しく紹介する「機械学習を用いた自然言語処理による商品レビューの評価」[1] では, Amazon の商品レビューを機械学習を用いて参考になる順に並びかえるシステムの構築, 及びその評価に関する研究を行っている。また近年では, 従来の EC サイトや商品の Web ページ以外にも, YouTube のような動画投稿サイトや X(旧 Twitter) や Instagram などの SNS で自社製品・サービスの宣伝を行う企業が増えてきている. それにつれて, 商品を購入する際に SNS や YouTube 上でその商品を宣伝している投稿を参考にしている人も増加している. そのため, SNS や YouTube 上の広告に対するユーザーのコメントも, 他のユーザーが商品の購入を検討する際の重要な判断材料になり得ると考えられる. つまり, SNS や YouTube 上での商品の宣伝に対するコメントは, EC サイトのレビューと同等の機能を持ち, その信頼性や参考になるかどうかが重要になるため, 評判分析や文書分類の研究の対象になると考えられる. ここで, SNS や YouTube は商品レビューのページとは異なり, 誰でも気軽にコメントを投稿できたり, その投稿内容も自由という特性上, 商品やサービスに関係ないコメントが多数存在する.

そこで, 本研究では分析対象を YouTube 上で自社製品やサービスを宣伝している動画に対するユーザーのコメントとし, トピックモデルの一種である Biterm Topic Model による商品に関するトピック抽出を用いて, その動画に対するユーザーのコメントから, 宣伝している商品やサービスに対して関連性が高いコメントを抽出するシステムの作成, 及び作成したシステムの人手に対する精度の検証を行った.

本論文の第 I 部では, EC サイトのレビューにおける評判分析やトピックモデルを用いた文書分類に関する関連研究の紹介, また本研究の研究目的を明確に説明する. 第 II 部では, 本研究で用いる二つのトピックモデルの説明, 及び提案手法のシステムや実装方法について説明する. 第III部では, 実際の YouTube 上の動画に対するコメントを用いた実験結果を述べる. 第IV部では, 実験結果をもとに考察した提案手法の有効性や将来性について述べる.

2　関連研究

本研究を進めるにあたり, 研究テーマの方向性決めや研究課題の発見, 及び本研究で用いている技術に関して参考にした論文を 4 つ紹介する.

2.1　機械学習を用いた自然言語処理による商品レビューの評価 [1]

この論文では, ユーザーが商品レビューを読んで参考になったかどうかを評価する機能が備わっていない EC サイトの場合に, 数多くあるレビューから参考になる情報を探す必要がある問題に着目し, 機械学習を用いた自然言語処理の手法で分析, 評価を行い, レビューを参考になる順番に並び替えるシステムの構築を目的としている. そして並び替えた順番が正しいかどうかを評価するために, クイックソートを利用した新しい評価法である QE 法を提案している.

図 1 はこの論文で提案されている, レビューを参考になる順番に並び替えるシステムの概要図である. はじめに, インターネット経由で Amazon の商品レビューのデータ取得し, 学習用データと評価データに分ける. 学習段階では, レビュー文章の正規化や各前処理を施し, 教師データとして準備する. この研究では, 全角数字やアルファベットを半角に変換したり, 数字は全て 0 に置換, アルファベットは全て小文字に変換などの正規化を行っている. また, 日本語形態素解析システムである MeCab を用いて形態素解析を行い, 品詞ごとに “_” で分割する. その後, 活用語の原型への変換, 及びストップワード除去を行っている. 例えば,「ロボットは２４時間働けるのでＡＩに仕事をとられる.」という文章の場合, 正規化と前処理を施すことで,「ロボット_0_働ける_ai_仕事_とる.」となる. この一連の処理を学習用データに施した後, 機械学習の際に用いる素性の抽出を行う. この研究はレビューを参考になる順序に並べ替えることが目的のため, 素性には単語の出現

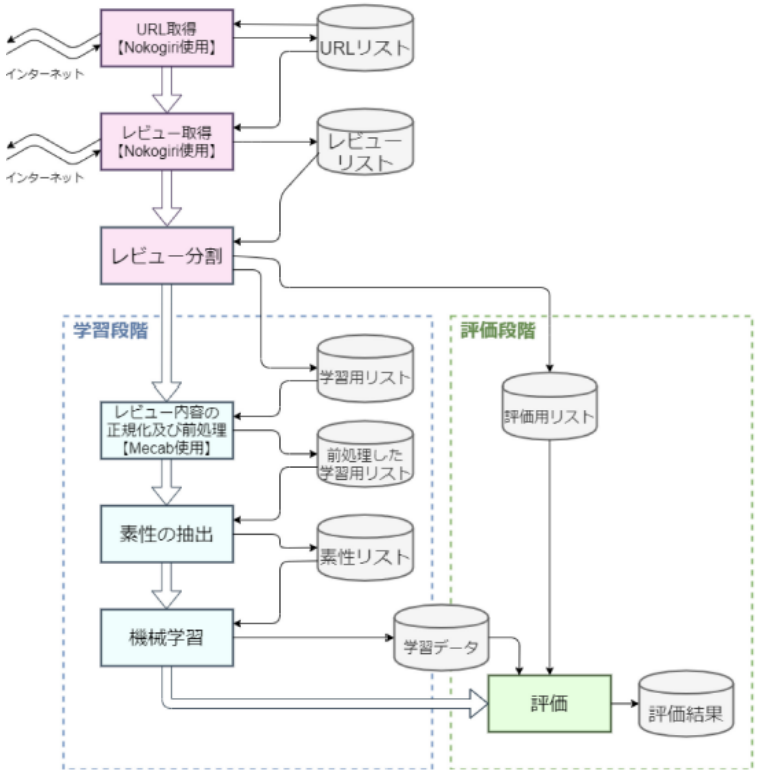

図 1: システム概要図

頻度を用いている．目的変数をレビューが参考になる確率 P とし, 抽出した素性を用いてロジスティック回帰により学習する．ロジスティック回帰のモデル式は式（1）で示される．θ_i は素性の重み, N は素性の数を表している．

$$P = \frac{1}{1 + exp(\theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_N x_N)} \tag{1}$$

次に学習したモデルを用いて, 評価用データに対して実験を行ない, 提案システムの精度を検証している．この研究の提案システムの精度の評価は, 実際の商品ページのレビューの並び順との一致率で評価している．正解の並び方を L_R, 提案システムによる並べ方を L_P としたとき, それぞれの要素の一致率を P_{match} としている．例えば, 以下の並び方のとき, $P_{match} = 100\%$ となり最も良い結果となる．

$$L_R : \{1, 2, 3, 4, 5\}$$
$$L_P : \{1, 2, 3, 4, 5\}$$

しかし, 以下のように並び方の評価としては良い結果と言える場合でも, 5 件のレビュー中 1 件のみ一致していることになり, $P_{match} = 20\%$ と低い結果になる．

$$L_R : \{1, 2, 3, 4, 5\}$$
$$L_P : \{4, 1, 2, 3, 5\}$$

このように正しい評価が行えない場合を解決するため, この研究ではクイックソートを利用した新しい評価法の QE 法 (Quicksort Evaluation method) を提案している．QE 法ではピボットを中央値とし, 昇順にするために要素を入れ替えた回数 S_{count} と, 要素数における最大の入れ替え回数 S_{max} を用いた式（2）により, 評価値 P_{QE} を求めている．なお, S_{max} は全ての要素が逆順の場合にクイックソートで昇順に入れ替えた回数である．

$$P_{QE} = 1 - \frac{S_{count}}{S_{max}} \tag{2}$$

実際の商品レビュー 52,403 件を取得し, そのうち 51,403 件を学習用データ, 1,000 件を評価用データに分けて実験を行い, 提案システムの精度を評価した結果を表 1 に示している．ここで, 登場回数 F とは学習の素性とするか決定するための単語の出現回数である．表 1 から, $F = 5000$, 学習率 $\eta = 1.7$ のときに評価値 $P_{QE} = 0.814$ と最大になる．従って, この論文で提案しているシステムはレビューを参考になる順序に並び替える手法として有効であると言える．

しかし, この論文では Amazon の商品レビューを分析の対象としていて, 素性には単語の出現頻度を用いているため, 提案システムが成り立つにはしっかり商品をレビューしている文章を学習させる必要がある．そのため, この論文で提案されているシステムでは YouTube で商品を宣伝している動画や, SNS の投稿に対するコメントを学習させた場合に上手く学習できなかったり, 精度が悪くなってしまうことが考えられる．なぜならば, YouTube の動画や SNS の投稿に対するコメントというのは誰でも気軽にでき, 内容も自由であるため, 商品のレビューのようなコメントの数が Amazon の商品レビューに比べると少ないからである．また, 一文の長さも短いことが多く, 素性となり得る単語の抽出も難しいと考えられる．そこで, 本論文ではそのような問題を解決するための手法を第 II 部で提案する．

表 1: 登場回数と学習率の組み合わせごとの評価値 P_{QE}

登場回数 F	素性数 N	学習率 η										
		1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
100	2083	0.788	0.783	0.800	0.806	0.796	0.766	0.750	0.762	0.769	0.747	0.768
200	2083	0.788	0.783	0.800	0.806	0.796	0.766	0.750	0.762	0.769	0.747	0.768
500	1472	0.774	0.748	0.781	0.782	0.769	0.769	0.763	0.751	0.787	0.774	0.726
1000	1058	0.728	0.794	0.746	0.781	0.758	0.813	0.792	0.762	0.795	0.776	0.784
2000	701	0.782	0.756	0.781	0.712	0.737	0.734	0.722	0.800	0.795	0.769	0.718
5000	363	0.759	0.773	0.774	0.801	0.764	0.763	0.772	0.814	0.757	0.765	0.755
10000	207	0.795	0.804	0.795	0.809	0.789	0.782	0.794	0.781	0.741	0.787	0.759

2.2 単語の出現頻度と類似性に基づいたトピックモデル洗練化手法 [2]

この論文では, 第Ⅱ部で後述するトピックモデルの一種の Latent Dirichlet Allocation(以下 LDA) を自然言語文書に適用する際の改善案を提案している. 通常, トピックモデルを自然言語文書に適用する際には, 前処理として分類に不必要なストップワードの除去を行うことが多いが, 一般的にストップワードリストに含まれている単語を除去するだけでは, 特定の文書にのみ頻出する特徴的な単語を除去することが出来ず, トピックモデルの精度に影響を及ぼすという問題が存在する. また, トピックモデルによって分類したトピックには, 類似したトピックが複数出現し, 分類の精度が下がるという問題も存在する.

そこでこの論文では, 前処理として分析対象としている文書から適切なストップワードリストを作成する方法を提案している. また, トピックモデルを適用後の後処理として, トピックを構成している単語の類似度からトピック間の距離を算出し, 類似しているトピックを統合することでより正確なトピック分類を可能にする手法を提案している. 図 2 は提案手法の全体像である.

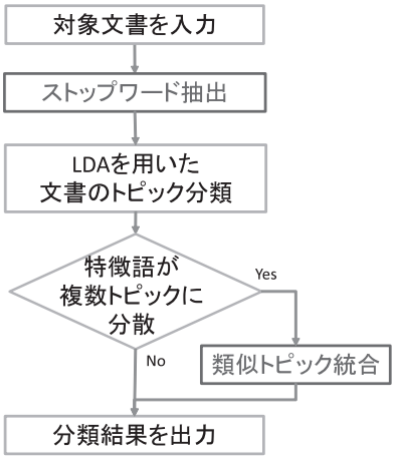

図 2: 提案手法の流れ

ストップワードリストを作成する手順を図 3 で示している. この手法では, まず対象としている文書全体に対して出現率が高い単語をストップワードとして抽出する. 出現率の算出にはは DF(Document Frequency) を用いている. DF とは, 文書全体に対してある単語 T が含まれる文書数のことであり, 事前に設定した閾値よりも高い DF 値を持つ単語をストップワードリストに加える. 次に, 抽出した単語と意味的に類似している単語をさらにストップワードとしてリストに加え

る．word2vec を用いて文章中の各単語を周辺の単語から学習し, 単語の分散表現を得て単語間の類似度を算出する．それによりある単語 T の類似単語を抽出することができ, ある閾値以上の類似度を示した単語を全てストップワードリストに加える．これにより, DF 値が高くない場合でも文書の特徴を表しにくい単語をストップワードリストに加えることが可能になる．

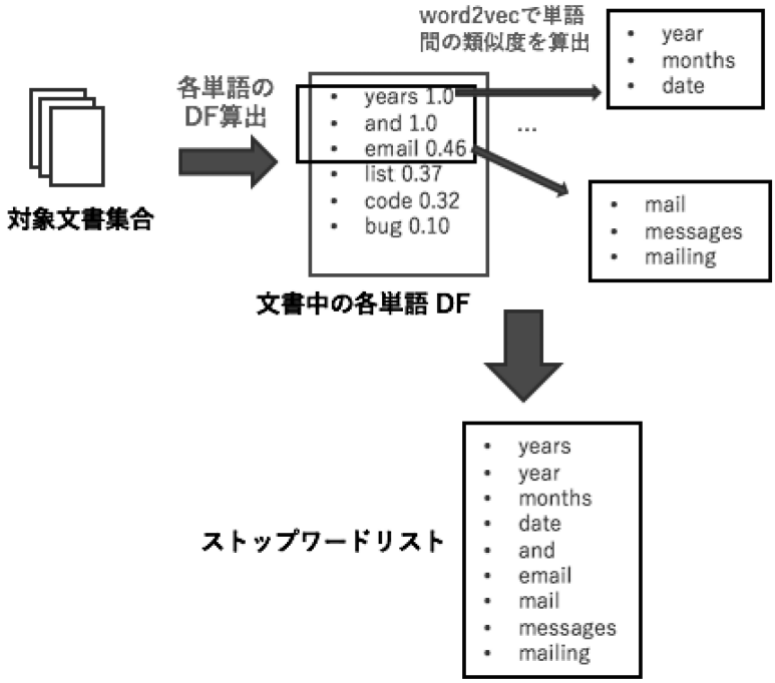

図 3: ストップワード抽出手法の流れ

トピックモデル適用後の後処理では, 分類結果に似たよったトピックが存在する場合にそれらのトピックを統合する処理を行うことを提案している．類似トピックの判断基準には, TF-IDF cos 類似度推定法が用いられている．これは, cos 類似度の計算に使用するベクトルの成分を TF-IDF で算出したものにした手法である．この研究では, 分類した各トピックの単語集合に対して TF-IDF cos 類似度を利用したクラスタリングを行い, その結果に従ってトピックを統合する手法を提案している．

以上の提案手法を LDA のよるメーリングリストのトピック分類に適用し, 評価項目に基づいて比較することで提案手法の有効性の評価を行っている．評価項目として, 一般的なストップワードリストを用いた手法と提案手法を比較している．また, 後処理として類似トピックの統合を行った場合についても比較を行っている．トピック分類の正確さを評価する指標として, 適合率, 再現率, それらの調和平均の F 値を採用している．実験結果を表 2 に示している．表 2 より, LDA を自然言語に適用したトピック分類では, 前処理としてストップワード除去を行うことで分類精度が大幅に向上していることが分かる．また, 一般的なストップワードリストを用いた場合と提案手法によるストップワードリストを用いた方法を比較すると F 値が向上している．また, 後処理を行った場合の F 値はさらに向上していることから, トピック分類の正確性を向上させる提案手法の有効性を確認できている．

本研究ではトピックモデルの分類精度よりも抽出する単語の質を重視しているため, この研究で提案されているストップワードリスト作成方法を用いると抽出したい単語をストップワードリストに加えてしまう恐れがある．また, 本研究で分析の対象としている YouTube のコメントは, 書

表 2: ストップワード数（$\#N$）, 適合率, 再現率, F 値の結果

	$\#N$	適合率	再現率	F 値
ストップワードなし	なし	0.047	0.497	0.086
Fox ストップワードリスト	425	0.237	0.397	0.297
Poisson ストップワードリスト	1238	0.355	0.422	0.385
RAKE	500	0.091	0.423	0.149
提案手法 (前処理)	761	0.411	0.418	0.414
提案手法 (前処理+後処理)		0.489	0.453	0.470

き言葉よりも話し言葉で書かれていたり, 若者言葉が使われていることが多く, 一般的なストップワードリストだけでは不要な単語を除去することが困難である. そのため, 前処理で形態素解析を行い, 特定の品詞のみ抽出してストップワードとして除去する, または特定の品詞のみでトピック分類を行う手法で実験を行った.

2.3 関連研究 3

関連研究 3 保留

3　研究目的

前項までの関連研究を踏まえた上で, 本研究の研究目的を明らかにする. 本研究では YouTube 上で商品やサービスを宣伝している動画に対する視聴者のコメントを分析対象とする. 対象としているコメントのうち, その動画で紹介している商品・サービスと関連性が高いコメントを抽出するシステムの作成を目的としている. [1] でレビューを参考になる順序に並び替えているように, 本研究では後述する提案手法により商品との関連性が高い順にコメントを並び替え, 他のユーザーがその商品の購入を検討する際に参考にするコメントを取得しやすいシステムの作成を目指す. また, 作成したシステムにより抽出した商品との関連性が高いコメントが, 人手の評価に対してどれほどの精度で抽出できているかを評価し, 最終的な提案手法の精度としている.

第 II 部
提案手法

第 II 部では本研究の提案手法, 及び提案手法で用いている主要な技術について説明する.

4　トピックモデル

提案手法の説明に先立ち, 本提案手法において主要な技術であるトピックモデルについて説明する. トピックモデルとは,

4.1　Latent Dirichlet Allocation

LDA の説明

4.2　Biterm Topic Model

BTM の説明

5　提案手法

ここでは, 前項で説明した Bitem Topic Model を用いて, YouTube 上で自社製品やサービスを宣伝している動画に対するユーザーのコメントから, 宣伝している商品やサービスに対して関連性が高いコメントを抽出するシステムを提案する. また, 提案したシステムの精度を検証する方法についても説明する.

5.1　データ収集

実験に用いる YouTube のコメントは, YouTube Data API v3 を用いて取得した. YouTube Data API v3 は Google Cloud Console で API キーを作成し, API を有効化することで様々な YouTube データにアクセスすることが可能になる. そのうち, YouTube のコメントに関連するものとして表 3 のようなデータが挙げられる.

表 3: YouTube Data API v3 で取得できるコメント情報

項目	内容
videoID	コメントした動画の ID
textDisplay	現在表示されているコメント
textOriginal	最初に投稿されているコメント
authorDisplayName	コメント投稿者の名前
authorProfileImageUrl	コメント投稿者のアイコン
authorChannelUrl	コメント投稿者のチャンネル
authorChannelId	コメント投稿者のチャンネル ID
likeCount	コメントに付いたいいねの数
publishedAt	コメントの投稿日
updatedAt	コメントの最終更新日

本研究では, 表 3 のうち textDisplay(現在表示されているコメント) のみを抽出し, 実験を行なった. また, YouTube のコメントには元のコメントの他に別のユーザーが返信しているケースも多いが, 本研究では返信しているコメントは扱わず, 元のコメントのみを抽出し実験の対象としている. 抽出したコメントの一例を図 4 に示す.

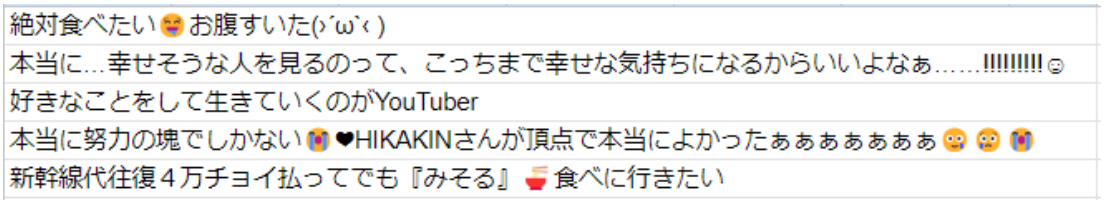

図 4: 抽出したコメントの一例

5.2 前処理手法

トピックモデルを含む自然言語処理の様々な手法において, テキストデータに対する前処理は非常に重要である. Web テキストを扱う場合には HTML タグや JavaScript のコードが含まれることもあり, 前処理としてそのようなノイズを除去する必要がある. また, 本研究では YouTube の動画に対するコメントをテキストデータとして扱うが, YouTube のコメントや SNS の投稿には絵文字や顔文字, URL, 話し言葉などを含んでいることが多い. そのため, 本研究でもトピックモデルによる分類を行う前の前処理は非常に重要である. 本研究で行った主な前処理とその簡単な説明を以下に示す.

5.2.1 クリーニング処理

空白、改行文字を削除

半角空白や全角空白, 及び “\n” などの改行文字を空文字に変換する.
例:「おはよう　今日はいい天気ですね」→「おはよう今日はいい天気ですね」

記号除去

“!” や “#”, 及び全角記号を除去する. また, YouTube のコメントの特性上顔文字が使われることも多いため, それらを除去する目的でもある.
例:「今晩、友達と映画を見に行く予定です！ 楽しみですね (ˇ ω ˇ)」→「今晩友達と映画を見に行く予定です楽しみですね」

絵文字除去

顔文字と同様に YouTube のコメントで使われることが多い. 感情分析などでは絵文字から情報を取得することもあるが, 本研究では感情に関する情報の抽出, 分析は行わないため絵文字は除去する.

数字を 0 に置換

自然言語処理の様々な手法において, 数字は意味を為さないことが多いため, 1 つ以上連続している数字を全て 0 に置換することが多い. 本研究でも数字に関する情報を必要としないため, 数字はすべて 0 に置換する.
例:「目標は年間で 10 回のイベントを開催することです」→「目標は年間で 0 回のイベントを開催することです」

単語の正規化

単語の正規化とは, 単語の文字種の統一, つづりや表記ゆれなどを無くすことである. この処理を行うことで同じ意味で異なる表記や形態の単語が同じ形になり, テキストの処理や解析が容易になる. 単語の正規化にはいくつか種類があり, 例えば

- テキスト内のアルファベットを全て小文字に変換する
- 半角カナを全角に統一する
- 辞書を用いた単語の統一

などがある.
例:「Google でﾈｺの写真を検索してください」→「google でネコの写真を検索してください」

連続長音記号除去, 繰り返し文字のまとめ

話し言葉や若者言葉でよくある「きたーーーーーー」や「うおおおおおお」など, 連続して長音記号が含まれている場合や同じ単語が繰り返されているものを削除, または一つにまとめる処理を行った.

例：「食べたーーーーい！！」→「食べたーい」

その他の前処理

スパムの可能性があるため, URL を含むコメントを削除した. また, YouTube の特性上外国人のコメントも多く存在したため, 日本語と英語以外の言語を含んでいるコメントを削除した.

以上の前処理を図 4 のコメントに適用した結果を図 5 に示す.

絶対食べたいお腹すいた
本当に幸せそうな人を見るのってこっちまで幸せな気持ちになるからいよなぁ
好きなことをして生きていくのがyoutuber
本当に努力の塊でしかないhikakinさんが頂点で本当によかったぁ
新幹線代往復0万チョイ払ってでもみそる食べに行きたい

図 5: 前処理後のコメント

5.2.2 MeCab による形態素解析及び分かち書き

次に, トピックモデルで学習する際に必要な文章の分かち書きを行う. 分かち書きとは, 自然言語処理の様々な手法において文章を単語や形態素などの最小単位に分割する処理のことである. この処理を行うことで, 言語解析や機械学習の際にテキストをより扱いやすい形態で実験を行うことが出来る. また, 英文の場合は単語間にスペースが明示的に存在するため分かち書きは必要ない場合が多いが, 日本語に関しては単語間のスペースがなく, 文章を単語単位に分割する処理を行わないと機械が単語を認識し解析することが難しくなるため, 分かち書きが必要である.

本研究では MeCab[参考文献] を利用して形態素解析, 及び分かち書きを行った. MeCab は京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所の共同研究で開発されたオープンソースの日本語形態素解析エンジンであり, 日本語の文法や単語の品詞情報をもとに文章を形態素に分解したり, 品詞の付与などが可能である. 本研究では MeCab を利用して対象のコメントに対して形態素解析を行い, 形態素に分割後スペース区切りで繋ぐことで分かち書きを行う. MeCab には最初から分かち書きを行う機能も含まれているが, 形態素解析によって抽出した品詞をストップワード除去に用いるためこの手法で分かち書きを行う.

また, 形態素解析の精度はエンジンのアルゴリズムの精度に加え, 形態素解析辞書の精度にも左右される. そのため, 形態素解析の目的にあった辞書を指定し, 解析することが重要となる. 表 4 は MeCab で形態素解析を行うときに主に用いられている辞書を比較したものである. 通常, MeCab での形態素解析には標準搭載されている mecab-ipadic を用いるが, mecab-ipadic は基本的な文法や専門用語に強い反面, 辞書の更新がないため新しい単語や固有表現に弱いという特徴がある. 本研究の分析対象である YouTube の動画に対するコメントは比較的新しい言葉や固有名詞などを含むことが多いため, mecab-ipadic に多数の web 上の言語資源から得た新語を追加し, カスタマイズした mecab-ipadic-NEologd を本研究では形態素解析の辞書に用いている.

表 4: 形態素解析辞書の比較

形態素解析辞書	特徴
mecab-ipadic	IPA コーパスをもとにした MeCab に標準搭載されている IPA 辞書. 基本的な日本語の文法や専門用語などの固有名詞に強いが, 辞書の更新がないため新しい言葉や固有名詞に弱い.
UniDic-mecab	言語学・国語学や音声情報処理など, より多様な目的に適した辞書. 「短単位」という揺れが少ない斉一な単位を見出し語に採用している.
mecab-ipadic-NEologd	多数の web 上の言語資源から得た新語を追加しカスタマイズした MeCab 用のシステム辞書. 辞書の更新が行われるので, 新しい固有表現に強い.

MeCab による形態素解析, 及び分かち書きの処理を行った後, 品詞によるストップワード除去と一般的なストップワードリストを用いたストップワード除去を組み合わせることで, 最終的に実験に用いるテキストの形式として整形する. 図 5 に分かち書きからストップワード除去までの処理の流れを示す.

5.3 BTM によるトピック抽出

5.4 文章生成

5.5 文章間の類似度計算

5.6 提案手法の精度検証

第 III 部
実験結果

6　実験目的、仮説

第 IV 部
参考文献

[1] 機械学習を用いた自然言語処理による商品レビューの評価
[2] 単語の出現頻度と類似性に基づいたトピックモデル洗練化手法