クラシファイドサイトにおけるユーザー投稿のクオリティーチェック自動化の検討

佐藤兼示[1]　桝井文人[1]　ミハウプタシンスキ[1]　荒田真輝[1]　鈴木智之[2]

[1]北見工業大学　情報システム工学科テキスト情報処理研究室
[2]株式会社ジモティー

{f1610800462, m1852400016}@std.kitami-it.ac.jp
{ptaszynski, f-masui}@cs.kitami-it.ac.jp
tomoyuki@jmty.jp

1　はじめに

Web サイトの一つに数行程度の短い広告文章を分類・集約して表示するクラシファイドサイトがある．このクラシファイドサイトで扱われる内容は，商品売買，物々交換，住居の空室，求人，イベント開催の告知など，様々な内容が掲載されている．クラシファイドサイトはユーザーが自由に広告文章を掲載することができる性質上，同じ内容の投稿を複数回掲載する重複投稿や，違法行為に関する投稿など，運営会社が定めた規約に違反した内容の投稿がある．

このような規約に違反した内容の投稿はサイト運営会社によるクオリティチェック (QC) でサイト上から削除される．QC とは，QC 担当者が手作業で投稿を確認，違反した内容の投稿であれば削除を行うことである．

しかしながら，サイトの投稿の数は膨大であり手作業での QC にかかるコストが課題となっている．

そこで本研究では，機械学習を用いてクラシファイドサイトのジモティー[1]でユーザーの投稿がクラシファイドサイトの投稿規約に違反した内容を含んでいないかどうかを自動判定する QC 自動化システムの実現を目指し，実現しようとするシステムに用いる投稿文に対する前処理と分類器の最適な組み合わせを実験的に検討したので報告する．

[1]株式会社ジモティーが運営している物々交換やメンバー募集などの投稿が掲載されているクラシファイドサイト

2　関連研究

文章に対してデータの前処理と機械学習を用いて自動分類する手法はこれまでも数多く提案されている．

山岡ら [1] は形態素解析と NaiveBayes の組み合わせで Twitter の投稿からツイートを投稿したユーザーの趣味を推定する手法を提案している．

野寄ら [2] は SVM やニューラルネットワークなどの機械学習のみを用いてソフトウェアのバグレポート中の文章に 5 種類の意味的ラベルを付与する手法を提案している．

竹岡 [3] は形態素解析と fasttext を用いて商品・サービスのカテゴリーを超えた競合関係をテキストマイニングによって分析・可視化する手法を提案している．

これらの機械学習による自動分類研究において，多くの学習アルゴリズムや特徴量抽出法が提案されているが，現実の課題やタスクに対する適切な組み合わせについては十分に整理されていない．また，単語や文章の意味で分類する研究はあるが，本研究で扱う投稿規約を基にした分類で，言葉の意味を用いた分類だけで十分であるかは明らかにされていない．

本研究では，数種類の前処理と分類器の組み合わせの中からクラシファイドサイトにおける QC に適した組み合わせについて議論する．

3　投稿 QC 自動化について

クラシファイドサイト「ジモティー」上には膨大な数の投稿が存在する．その中には，同じ内容の重複投稿や，違法行為に関する投稿など，運営会社が定めた

規約に違反した投稿が含まれる．これらの投稿に対して，ジモティーではQC担当者が手作業で全ての投稿を確認し，違反した内容の投稿であれば削除を行うことで対応している．

しかし，手作業によるQCコストの大きさが問題になっている．

本研究では，この問題に対処するために，機械学習を用いてユーザー投稿が投稿規約に違反しているか否かを自動判定するシステムの構築を目指す．

本稿では麻薬や医療器具などのジモティーで取り扱えない物の名称を要注意ワードとし，要注意ワードが含まれる投稿を「赤投稿」，要注意ワードを含まない投稿を「白投稿」と呼ぶことにする．

白投稿に含まれる非承認投稿の数は赤投稿に含まれる非承認投稿の数に比べて極めて少ない．そのため，白投稿は赤投稿に比べて非承認投稿である可能性が低く，QCにかかる時間も赤投稿より短い．

よって，もし投稿文と要注意ワードの部分一致で分類された赤投稿を機械学習による自動分類でさらに白投稿と赤投稿に判別できれば全体的なQCコストを低減できる可能性が高い．

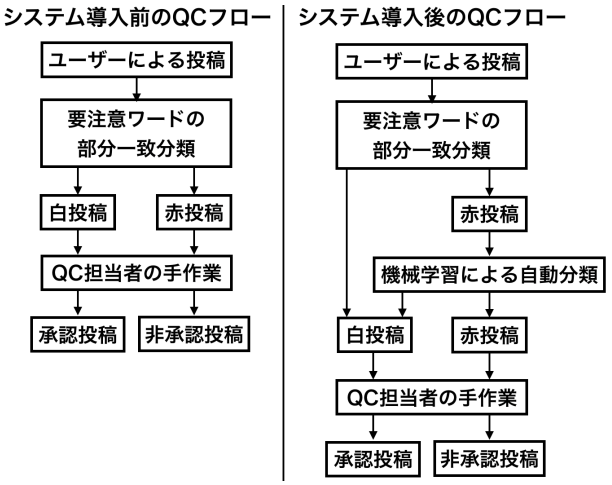

図 1: システム導入前後のQCフロー

4 投稿分類実験

冒頭でも述べたように，機械学習を用いたユーザー投稿のQC自動化システム構築に適した前処理と分類器の組み合わせを選出するため，様々な前処理と分類器を組み合わせ，分類実験を行った．

佐藤ら [5] は白投稿と赤投稿を区別せずに機械学習を用いて承認投稿と非承認投稿に分類する実験を行ったが，QC自動化システム適した前処理と分類器の組み合わせを見出せなかったと報告している．これは，重複投稿など非承認投稿になる原因が投稿文にない投稿がデータセットに含まれるからではないかと推測している．

よって，本稿では要注意ワードを含む赤投稿で分類実験を行う．

以下，実験に用いたデータセットと実験手順について説明する．

4.1 データセット

実験では，実際の赤投稿データ38万件を用いた．さらに，赤投稿は承認投稿と非承認投稿それぞれ19万件から成る．

4.2 前処理

MeCab[2]とCaboCha[3]を用いて前処理を行う．

機械学習の研究において範ら [4] のように単語分割を使った実験が多くあるため本研究でも前処理として単語分割 (tokenization) を用いた．

QCでは重複投稿などで非承認投稿になることもあるため，分類において精密な言語学的ルールが存在するかどうか不明である．そこで，文法情報が付与された単語分割+原型化 (lemmatization)，単語分割+品詞化 (Part-of-Speech tagging)，文節区切り (chunking)，係り受け解析 (dependency) の前処理も適用した．

単語分割
文章に対して単語毎に空白区切りの処理を行う．

単語分割+原型化
単語分割で区切った単語を原型化する処理を行う．

単語分割+品詞化
単語分割で区切った単語を単語を表す品詞に置き換える．

文節区切り
文章に対して文節毎に空白区切りの処理を行う．

[2] https://taku910.github.io/mecab/
[3] https://taku910.github.io/cabocha/

係り受け解析
　文節区切りで区切った文節に係り受け構造の階層情報を付与する．

本稿では前処理を行って空白で区切られたものを素性と定義する．

4.3　分類器

前処理を行ったデータセットを素性ごとに TF-IDF 値で表現する．TF-IDF 値の計算には weka[4]の StringToWordVector を用いた．

次に，TF-IDF 値で表現されたデータセットを分類器を用いて承認クラス，非承認クラスに分類する．クラスの分類には DecisionTrees，kNN，SVM，NaiveBayes，畳み込みニューラルネットワーク (CNN) を用いる．

4.4　赤投稿のタイトルを用いた分類実験

赤投稿の 38 万件のデータセットを用いて，赤投稿のタイトルに対してを，文節区切りの前処理を用いて特徴抽出する．それらを CNN，Naive Bayes，SVM，kNN を用いて承認クラスと非承認クラスに分類する．

5　実験結果

上で述べた前処理と分類器を組みわせ分類実験を実施した．実験は 10 分割検定を用いて実施し，その結果に対して分類性能を評価した．

評価尺度は Precision と Recall を用いた．ただし，本タスクにおいては Recall がより重視される．

非承認投稿がサイトに掲載させる承認クラスに分類されてしまうことを優先的に抑制するため，非承認投稿の Recall が最も高いものを QC 自動化システムに適した前処理と分類器の組み合わせとして優先的に採用する．非承認投稿の Recall が高い場合，Precision と F-値も考慮して実験結果を考察する．

5.1　赤投稿のタイトルを用いた分類実験

分類実験の結果を図 2 に示す．

図 2 より，すべての分類器で非承認投稿の Recall が低くなった．

[4]https://www.cs.waikato.ac.nz/~ml/weka/

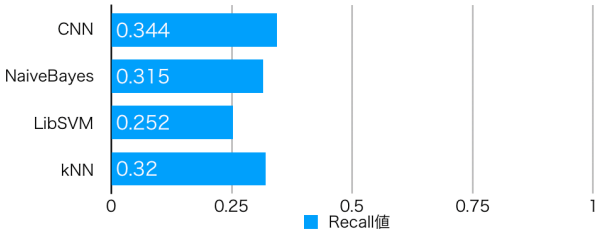

図 2: 赤投稿タイトルの非承認投稿の Recall

5.2　赤投稿の本文を用いた分類実験

赤投稿のタイトルを用いた分類実験で最も非承認投稿の Recall が高かった文節区切りと CNN を用いて赤投稿の本文に対して分類を行った．その他に比較対象として単語分割，単語分割+原型化，係り受け解析の前処理と CNN を組み合わせて分類を行った．分類結果を図 3 に示す．

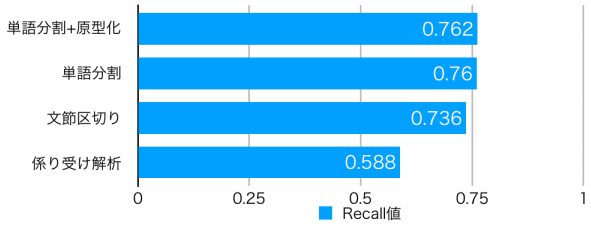

図 3: 赤投稿本文の非承認投稿の Recall

表 1: 承認投稿の分類結果

前処理	Precision	Recall	F-値
単語分割	0.748	0.712	0.729
単語分割+原型化	0.747	0.702	0.724
文節区切り	0.698	0.609	0.651
係り受け解析	0.582	0.571	0.576

表 2: 非承認投稿の分類結果

前処理	Precision	Recall	F-値
単語分割	0.725	0.760	0.742
単語分割+原型化	0.719	0.762	0.740
文節区切り	0.653	0.736	0.692
係り受け解析	0.582	0.588	0.583

図 3 より，単語分割の前処理を行った場合，非承認

投稿の Recall が 0.760 でタイトルを用いた分類と比べて QC 自動化システムに適した結果になった．さらに表 1，2 より，同じく単語分割の前処理を行った場合，承認投稿と非承認投稿のそれぞれの Precision，Recall，F-値全ては 0.7 を上回って最も QC 自動化システムに適した結果になった．

6 考察

赤投稿のタイトルを用いた分類実験では QC 自動化システムに適した結果が見出せなかった．これは投稿のタイトルに分類で用いる素性が少ないことが原因であると考察した．また，非承認になる可能性の高い要注意ワードが赤投稿のタイトルにはあまり含まれていなく，投稿の本文に含まれているのではないかと考察した．そのため，投稿について詳細に書かれている本文を用いて赤投稿の分類実験を行った．

赤投稿の本文を用いた分類実験において用いた前処理を比較すると，単語分割+原型化が最もユニークな素性の数が少なく，係り受け解析が最もユニークな素性の数が多いと言える．

単語分割:ぼく　わたし　あなた　は　の

文節区切り:ぼくは　わたしは　あなたは　ぼくの　わたしの　あなたの

図 4: ユニークな素性の例

例えば，図 4 のように単語分割と文節区切りを比較すると，単語分割のの 5 つのユニークな素性を用いて，文節区切りのの 6 つのユニークな素性を表すことができる．係り受け解析は文節区切りのユニークな素性にさらに階層の情報も加わるため，ユニークな素性の数は文節区切りより係り受け解析の方が多いと言える．ここで図 3 より，ユニークな素性の数が多いほど非承認投稿の Recall が低くなっている．したがって，ユニークな素性が少なければ，分類性能が向上するのではないかと推測した．しかし，表 1，2 より，ユニークな素性の数は単語分割+原型化の方が単語分割より少ないが F-値は単語分割の方が高い．よって，上記の推測は間違いで単語分割のユニークな素性の数がこのシステムに最も適していると考察した．

今後さらに分類性能の向上を図るために，今後は言語モデルの複雑さを上げて実験を行う．今回の実験では前処理を行ったデータを Bag of Words で表現したが，word2vec や glove などの単語分散表現を用いて言語モデルの複雑さを上げることによって，より QC 自動化システムに適した分類結果につながるのではないかと考察したため今後実験を行う．

7 おわりに

本研究の目標としてクラシファイドサイトにおけるユーザーの投稿を承認か非承認のどちらかに自動的に分類するシステムの構築を目指す．そのために本稿ではシステムに最適なデータ前処理と分類器の組み合わせを選出する実験を行った．

はじめに，白投稿と赤投稿のタイトルに対して分類実験を行ったが，どちらの結果も QC 自動化システムに適さない結果となった．これは要注意ワードがタイトルにあまり含まれていなく，本文に含まれているからではないかと考え，赤投稿の本文に対して分類実験を行った結果，QC 自動化システムに適した結果になった．

分類性能の向上を図るため，今後は word2vec や glove を用いて言語モデルの複雑さを上げて実験を行う．

参考文献

[1] 山岡 拓生， 佐野 睦夫 : ”ナイーブベイズ法に基づく SNS を利用したペルソナ推定”， The 32nd Annual Conference of the Japanese Society for Artificial Intelligence， 2018.

[2] 野寄 祐樹， 鷲崎 弘宣， 深澤 良彰， 鹿糠 秀行，大島 敬志， 土屋 良介:”バグレポートの検索向上のための機械学習による文章単位の自動ラベリング”，情報処理学会第 80 回全国大会.

[3] 竹岡 志朗:”機会学習を活用したテキストマイニングー口コミを用いた商品・サービスカテゴリーの横断分析”，桃山学院大学経済経営論集 第 59 巻第 4 号.

[4] 範 暁蓉，二宮 崇:”対訳語抽出における Decipherment 法と文脈に基づく手法の比較”，平成 27 年度 AAMT/Japio 特許翻訳研究会 報告書

[5] 佐藤兼示，桝井文人，プタシンスキミハウ，荒田真輝，鈴木智之:”クラシファイドサイトにおける掲載不可投稿の検出のための機械学習による分類”，言語獲得と理解研究会報告 LAU Technical Reports, 2020.