修 士 論 文

顔文字から見るSNS上の感情と社会トレンドについての研究

北陸先端科学技術大学院大学
知識科学研究科知識科学専攻

山口 和宏

2013年3月

修 士 論 文

顔文字から見るSNS上の感情と社会トレンドについての研究

指導教官 Dam Hieu Chi 准教授

北陸先端科学技術大学院大学
知識科学研究科知識科学専攻

1150036 山口 和宏

審査委員: Dam Hieu Chi 准教授 (主査)
國藤 進 教授
吉田 武稔 教授
由井薗 隆也 准教授

提出年月: 2013 年 2 月

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究背景

SNS が普及したことにより、社会構造に変化が起きている。90 年代までは人と人とのコミュニケーションは口頭、手紙、電話など一対一のコミュニケーションが主流であった。これに対し、インターネットが普及し SNS 上でのコミュニケーションが日常的となった現在では、ブログや Twitter、Facebook などでのコミュニケーションの割合が増加しており、複数の人と同時にコミュニケーションをおこなうことが可能となっている。他人の意見や情報を広く収集できるようになったことで、より正確に社会を理解できる機会が生まれている。同時に、個人が自由に情報発信することができ、社会への影響力を最大化する機会が生まれている。このため、SNS 発の社会現象や社会トレンドが続出している。具体例としては、アラブの春や反原発デモなどがあり、SNS 上での人々の活動は社会を理解する上で無視できないものとなっている。

この SNS と社会との関係について、いくつもの興味深い報告がある。政治分野では、2012 年のアメリカ大統領選挙においてオバマ氏の演説中に最大 52,000 件/分のツイートが記録されたのに対し、対立候補のロムニー氏の演説中では 14,000 件/分のツイートが記録された[1]。ツイート数と選挙の当落との関係については今後の研究が待たれるが、興味深い事例である。医療分野では、Twitter 上での”インフルエンザ”というキーワードについて検索をおこない、発言者がインフルエンザにかかっているかの分類をおこなったのちに感染症モデルを適用し、実際のインフルエンザ疾患者数の推定をおこなった試み [荒牧 12] が報告されている。経済分野では、Twitter の映画に関するツイート件数から映画の興行収入を予測した試み [AH10] や、Twitter 上の気分を推定し、ダウ工業平均株価 (DJIA) と併せて分析することで 3 日後の DJIA を 87.6%の精度で予測した試みが報告されている [BMZ10]。nature 誌においても、インターネット上と実社会上での人と人との繋がりについての知見をまとめた報告がおこなわれている [Gil12]。このように、SNS のビッグデータとデータマイニングの融合による社会の解明はあらゆる分野で応用が進められている。

一方、社会を解明する作業の一部として社会トレンドの理解がある。この社会トレンドへ影響を与える要素として、社会を構成する人々の感情が大きな要因となっていると考えられる。SNS はテキストを介したコミュニケーションプラットフォームであるため、SNS から感情を理解しようとする場合は自然言語処理技術を利用しておこなう。この自然言語処理には以下のタスクがある。

[1] http://popwatch.ew.com/2012/09/07/obama-twitter-record/

形態素解析 文を単語 (形態素) に分割し、各単語の品詞を特定する

構文解析 文節の修飾関係を特定する

語彙・意味解析 同義語の意味を特定する

照応解析 代名詞・指示語の推定や省略された名詞句の補完をおこなう

しかし、SNS 上では従来の自然言語処理では解析が困難な語が登場している。その一例を表 1.1 に示す。これらは主に Twitter 上でみられる。特に、文字数制限があり、リアルタイムで即時的なコミュニケーションがおこなわれる Twitter では、より速く・少ない文字数で意味や意思が伝達される傾向にある。そのために各々の利用者がより速く・少ない文字数でよりユーモラスに情報を伝達しようとする試行錯誤の結果、このような言葉が生まれてきた。この顔文字、崩れた表記、未知語を解析の対象とした分析は不自然言語処理と呼ばれ、特に顔文字を対象とした解析が近年注目を浴びている。

また、顔文字は読み手の感情に影響を与えることが加藤らの報告 [加藤 05] により示された。加藤らはさらに、電子メールを介したコミュニケーションにおいて、メールの読み手が書き手の感情をどのように解釈するかについて報告している [加藤 08]。これによると、読み手が解釈する書き手の感情は読み手自身の感情状態に依存するとされている。読み手が自身の感情状態に従って文章から感情を評価するならば、多量の文章を多くの人が読み、それぞれの書き手の感情を評価するとき、それらの文章の示す感情を総合すると読み手全員の感情を反映するのではないか、と考えられる。

一方 Twitter 上では多くの投稿があり、一つの投稿を多くの人が目にする。そのため、多くの投稿について読み手がどのように書き手の感情を解釈するのかを分析し統合すると、SNS 全体の感情を示すと考えられる。

そこで本研究では顔文字に着目し、その感情を解析することで SNS 上の感情を解析する。そして一般的にポジティブ/ネガティブなイベントと考えられるクリスマスと地震発生日を例に取り、それぞれのイベント時期について SNS 感情と社会トレンドとの関係について分析をおこなう。

1.2 関連研究

1.2.1 顔文字研究の始まり

顔文字に着目した研究は 90 年代後半から心理学や認知科学の分野でおこなわれてきた。井上らは顔文字などの記号列が書き手の感情を伝達することを示し [井上 97]、荒川らは謝罪文に付与された顔文字が読み手の怒りに与える影響について報告した [荒川 04]。また顔文字が言語表現で表すことのできない微妙な感情表現を補う情報である [登美 04] として、感情や配慮などの役割について研究がおこなわれてきた [川上 08, 登美 04]。

表 1.1: 自然言語処理に適さない語

種類	例
顔文字	(^o^) (T Д T)
崩れた表記	おはあり ぁ ν)ヵ ヾ ´ つう
未知語	キチョハナカンシャ

1.2.2 計算機科学と顔文字

一方心理学や認知科学などの社会科学以外の分野においても顔文字の重要性が認知されるようになってきた。人工知能やヒューマンコンピュータインタラクション、計算言語学などの分野では、コンピュータを介したテキストベースのコミュニケーションを発展させるため、オンラインでおこなわれるコミュニケーションにおいて、その文脈にふさわしい顔文字の生成と顔文字の表す意味を特定する研究が活発におこなわれた。Derks らはインターネット上のコミュニケーションにおける顔文字の使用について社会的な意味を調査し [DBvG07]、Maness は大学生によりおこなわれたチャットコミュニケーションを言語学的に分析し、常にコミュニケーションにおいて顔文字の使用は重要な意味を持っていることを明らかにした [Man08]。

1.2.3 顔文字の解析

顔文字を対象とした感情解析には、アンケート、カーネルを拡張した SVM、キネシクス理論によるアプローチなどがある。

アンケート

川上らは携帯電話に予め登録されている顔文字を対象に、アンケートにより顔文字が”喜び”、”哀しさ”、”怒り”、”楽しさ”、”焦り”、”驚き”、”強調”のそれぞれをどの程度表しているか、という視点で分析をおこない、データベースを作成した [川上 08]。

まず対象となる顔文字を”笑い”、”泣き”、”怒り”、”焦り”、”驚き”、”その他”の 6 クラスに分類に分類した。それぞれのクラスで近親性 (日常よく使用するか・よく見かけるか) についてアンケートにより評定を求め、各クラスで親近性の高い上位 4 個の顔文字 (”笑い”については顔文字の数が多かったため 8 個) と”その他”に分類した 7 個を加えた 31 個を最終的な分析対象の顔文字とした。

アンケート結果に対する評定値間の相関係数を算出し、川上は”喜び”と”楽しさ”はほぼ同一のものであるかのように扱われており、両者を別々に測定する必要はない。また、”哀しさ”と”焦り”を同時に表現しうる顔文字の例を上げ、ポジティブではない感情を適切に区別することは困難であると報告している。

カーネルメソッド

田中ら [TTO05] は顔文字の抽出を自然言語処理の一部であるチャンキング (構文解析) の一種と見なし、各文字について SVM ベースの形態素解析器である yamucha[2]を用いて、該当文字が顔文字の一部か否か判定した。次に Dynamic Time Alignment Kernel(DTAK), String Subsequence Kernel により類似度を測定し、k 近傍法、SVM により顔文字の分類をおこなった。カーネルの拡張には多項カーネル、RBF カーネルによりおこなった。

顔文字の感情については”Happy”, ”Sad”, ”Angry”, ”Surprised”, ”Action”, ”Wry Smile”を設定し、k 近傍法 (k=1, 10) と SVM(それぞれのカーネル拡張の有無) について感情を推定した。その結果、DTAK、多項カーネルを用いた SVM において、90.4%の精度で感情の推定が可能だと報告した。

キネシクス理論

Michal らはキネシクス理論 (動作学) に基づき、顔文字を目または口を表す意味領域 (kinemes) に分割し、顔文字の感情を推定するシステムを考案した [PDRA10, PMD+10]。

インターネット上の顔文字辞書から収集した顔文字を”目”+”口”+”目”のトリプレット (文法情報) に分割し、データベースを構築した。これらの感情辞書の感情ラベルについては辞書毎に異なるため、Michal らが作成した感情解析システム (ML-Ask)[PDS+09] を用いて統合した。統合先の感情は、中村の感情表現辞典 [中村 93] の”昂”、”驚”、”恥”、”喜”、”好”、”安”、”哀”、”厭”、”怒”、”怖”の感情種類を Russell の二次元感情モデル (ポジティブ-ネガティブ、活動的-非活動的の二軸)[A80] にマッピングしたものである。

入力文に顔文字が含まれるか判定し、含まれる場合は顔文字をデータベースと照合して感情を推定する。照合は顔文字辞書との完全一致、目・口・目のトリプレットとの一致、データベースに存在する目・口・目のトリプレットのすべての組み合わせの順でおこなわれる。感情の推定は、顔文字辞書との一致、目・口・目のトリプレットによる感情アノテーション、目・口・目の各トリプレットでの感情アノテーションの順でおこなわれる。その結果推定の精度は 93.5〜97.4%という高精度を達成した。

[2]http://chasen.org/ taku/software/yamcha/

1.2.4 問題意識

前節では Michal らが高精度で顔文字の感情を推定したことを示した。彼らの推定結果は顔文字と感情の一対一対応であるが、顔文字は文脈により多様な意味を示す。顔文字の多様な感情を抽出するためには単一次元ではなく、多次元で顔文字の感情を表現する必要があり、これに取り組む。

1.3 研究目的

本研究では顔文字からの感情解析により SNS 感情トレンドと社会トレンドとの相関性を明らかにする。そのために、次の二点に取り組む。

1. 顔文字による感情解析手法の確立
2. Twitter データからの感情抽出と社会トレンドへのアプローチ

1.4 知識科学的意義

人と人との円滑なコミュニケーションにおいては、相手の意思や感情を正しく理解することが重要である。対面でおこなわれるコミュニケーションにおいては、会話の内容だけでなく表情や身振りなど言語以外の情報が利用可能であるが、テキストベースのコミュニケーションでは言語以外の情報はほぼ欠落している。このため文章によるコミュニケーションでは齟齬が発生しやすいが、これを解消する一手段として顔文字がインフォーマルなコミュニケーションにおいて頻繁に利用されている。しかし顔文字は複数の意味で使われ、その解釈は読み手の主観に依っている。

本研究ではこの顔文字が表す感情をデータマイニング手法を用いて定量的に評価する。つまり、読み手が顔文字から解釈する感情という暗黙知を定量的に評価する点に意義がある。

また SNS の感情トレンドと社会のトレンドとの関係について分析をおこなう。SNS も社会も人間が構成するものであるため両者の間には何らかの関係や共通点が見られると推測されるが、SNS と社会との関係についての考察は未だ発展途上である。本研究ではこれらについて考察をおこなう。これにより、社会という複雑なシステムについて一つの知見が得られる可能性がある点に意義があると考える。

1.5 研究の流れ

研究の流れを図 1.1 に示す。

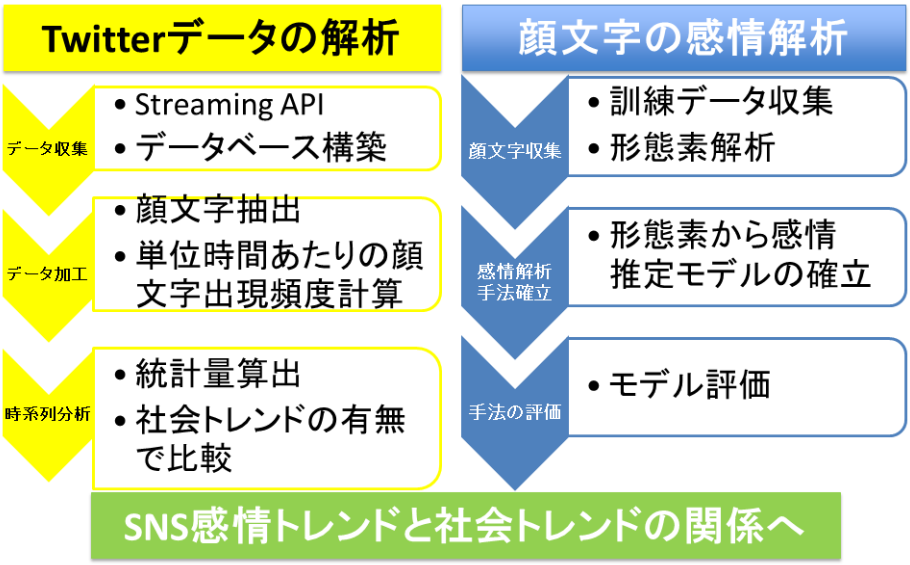


図 1.1: 研究の流れ

1.6 構成

本論文の構成を説明する。

第 2 章 データ収集について述べ、実際に使用するデータの概要も併せて説明する。

第 3 章 使用する分析手法について簡単に説明する。

第 4 章 事前分析として、主観により顔文字を感情毎に分類し、分析した結果を述べる。

第 5 章 顔文字の感情解析手法について述べる。

第 6 章 第 5 章で提案した手法を用いて再度分析した結果を述べる。

第 7 章 本論文の結論と今後の展望を述べる。

第2章　利用データ、顔文字の定義

2.1　データ収集

本研究ではTwitterの提供するデータを使用し、顔文字を抽出して考察をおこなう。データ収集にはTwitter APIのPythonライブラリであるtweepy[1] version1.9を使用した。以下で各APIデータの特徴と取得できるデータの概略を示す。なお本研究では日本語のツイートのみを分析の対象としている。

2.1.1　Search API

検索キーワードにマッチする直近3,000件のデータを取得する。キーワードは複数指定が可能で、論理和・積による検索も可能である。実際のデータ量は検索キーワードによる。データ収集期間は2011.12.03~2012.12.28である。

2.1.2　Streaming API

全ツイートから1%をランダムサンプリングした結果を取得する。データ量は約500,000,000ツイート/日程度であり、そのうち日本語のツイートは約490,000ツイート/日であった。データ収集期間は2012.05.16~2012.12.28である。

2.1.3　利用できる属性

本研究では上記APIを利用して取得したデータの内、ツイートID、ユーザーID、投稿日時、投稿元クライアント情報、投稿内容を利用した。その他に利用できる属性の一覧を付録Bに示す。

2.1.4　データベース作成

Streaming APIを用いて収集するデータは一日あたり13GBにのぼる。この膨大なデータから目的のデータを効率よく検索、利用するためにデータベースを構築した。構築したデータベース図を付録Aの図A.1に示す。

[1] https://github.com/tweepy/tweepy

2.2 顔文字

顔文字は古くはタイプライターの時代より使用され、現代では爆発的に普及してきている。顔文字教室[2] によると、現在までに 1 万種類以上の顔文字が考案されており、ウェブページなどで欠かすことのできないものとなっている。顔文字は数個の記号の組み合わせで構成され、表 2.1 に示すような種類に分けられる。本研究では顔文字に着目した分析をおこなうため、ここで本研究で扱う顔文字の定義と抽出方法について述べる。

表 2.1: 顔文字の種類

分類	顔文字
西洋式	:)
東洋式	(^o^)
日本式	(」・ω・)」うー！(／・ω・)／にゃー！
	∧＿∧ (･ｪ･)ﾌﾟｯ (　　) ｜　｜｜ (＿,,)_)

2.2.1 定義

Michal らは顔文字の定義を次のように述べている [PDRA10]。

顔文字とは 顔、姿勢などを表し、ユーザの感情を伝えるために頻繁に使われる文字列・記号列である。

日本において頻繁に見られる顔文字は、表 2.1 の東洋式、日本式の顔文字である。これらは顔の輪郭を示す記号として”(”と”)”で囲われたものが多い。そこで本研究ではこの定義に次の条件を付け加えた。

顔文字とは ”(”と”)”で囲われた文字列・記号列である。

顔文字とは 一つの顔により表現される文字列・記号列である。

顔文字の一例を表 2.2 に、本研究で扱わない顔文字の例を表 2.3 に示す。

[2]http://kaomoji.kyo-situ.com/

表 2.2: 顔文字の例

Σ（゜д゜lll）
(≧∇≦)
(T_T)

表 2.3: 顔文字として扱わないもの

理由	顔文字
”(”と”)”で囲われていない	（・д|柱|
同上	orz
複数の顔文字で構成されている	（ノД｀）ヾ（・ω・`*）

2.2.2 抽出方法

パターンマッチングにより顔文字を抽出した。以下に手順を示す。まず対象文字列について全角・半角記号の統一、空白文字の削除をおこない、”(”と”)”で囲われた文字列があるか判定した。該当部分が存在する場合は”∑”、”ノ”など、顔文字に頻繁に付随する文字を含めて抽出した。この頻繁に付随する文字は、筆者らの相談の上決定した。次に抽出した文字列について URL、”(笑)”など、顔文字の定義に合致しないものを削除した。

第3章　手法

3.1　本研究で使用するソフトウェア、分析手法

本研究で利用する分析手法及び分析に使用したソフトウェアについて説明する。

3.1.1　使用ソフトウェア

本研究では分析にフリーの統計処理ソフトウェアである R[1]、フリーの形態素解析器である MeCab[2]を利用した。

3.1.2　検定力分析

検定力分析は R において pwr パッケージ[3]で実装されており、これを利用した。

概要

統計学における仮説検定では、第一種の誤り、第二種の誤りに注意する必要がある。仮説検定における真実と判定の関係を表 3.1 に示す。表 3.1 に示すとおり、第一種の誤りとは帰無仮説が真であるが対立仮説を採択してしまう、第二種の誤りとは対立仮説が真であるが帰無仮説を採択してしまう問題であり、それぞれ確率 α, β で発生する。仮説検定では棄却したい仮説、たとえば比較群において差が無いこと、を帰無仮説として設定するが、第二種の誤りとは”差があるのにも関わらず、差がない”と判定することである。この第二種の誤りを犯さず正しく判定できる確率 $1-\beta$ を検定力という。有意水準 α、標本数 N、効果量 es(標準化された平均値の差)、検定力を用いて、より効果的な統計的検定をおこなうため、これらを吟味・考慮する方法を検定力分析という。

3.1.3　決定木

決定木は R において mvpart パッケージ[4]で実装されており、これを利用した。

[1]http://www.r-project.org/
[2]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[3]http://cran.r-project.org/web/packages/pwr/index.html
[4]http://cran.r-project.org/web/packages/mvpart/index.html

表 3.1: 統計的検定における真実と判定の関係

真実＼判定	帰無仮説を採択	対立仮説を採択
帰無仮説が真	正しい判断 $1-\alpha$	第一種の誤り α
対立仮説が真	第二種の誤り β	正しい判断 $1-\beta$(検定力)

表 3.2: データ概要

等級	大人子ども	性別	生死
1 等 :325	子ども: 109	女性: 470	死亡:1490
2 等 :285	大人 :2092	男性:1731	生還: 711
3 等 :706			
乗務員:885			

概要

決定木とは、データを分類し、if-then ルール形式で記述する分類手法である。東京図書の「データマイニング入門-R で学ぶ最新データ解析」で紹介している事例を引用し、簡単に具体例を示す。東京図書で公開している第 3 章、タイタニック号の生還者と死亡者について、役割、性別、大人子どもという属性を付与したデータを使用した[5] データの概要を表 3.2 に示す。

このデータから、”もし〜ならば ••• である”。というルールを作成する手法が決定木である。実際に作成したルールを図 3.1 に示す。赤い四角で囲まれている部分をノードと呼び、この図においては一番上の”死亡”のノードをルートと呼ぶ。分岐は上から下に向けて (または下から上に向かって) 一方向のみにおこなわれ、逆戻りすることはない。分岐の進行方向に従って、(この図では上から下に向けて分岐が進むので) 上のノードを親ノード、下のノードを子ノードと呼ぶ。決定木の中の部分木を枝またはブランチという。たとえば”男性”を含めそれより下の 5 個のノードの関係図の部分を、ノード”男性”のブランチという。ブランチの終点のノードをターミナルノードという。

決定木ではルートに近い分岐を生じさせている変数が基準変数に対して強い影響力を持っていると解釈する。今回の例では、決定木はまず”性別”が”生死”を分ける第一の要因

[5]http://www.tokyo-tosho.co.jp/books/ISBN978-4-489-02045-2.html

であると示している。この図で言うならば、もしルートノードにおいてあるデータの性別が”女性”であるならば、右側の枝に分岐する。そしてこのノード”女性”のブランチにおいて、470 名中 334 人が生還者であり、126 人が死亡者である。というルールが作成された。

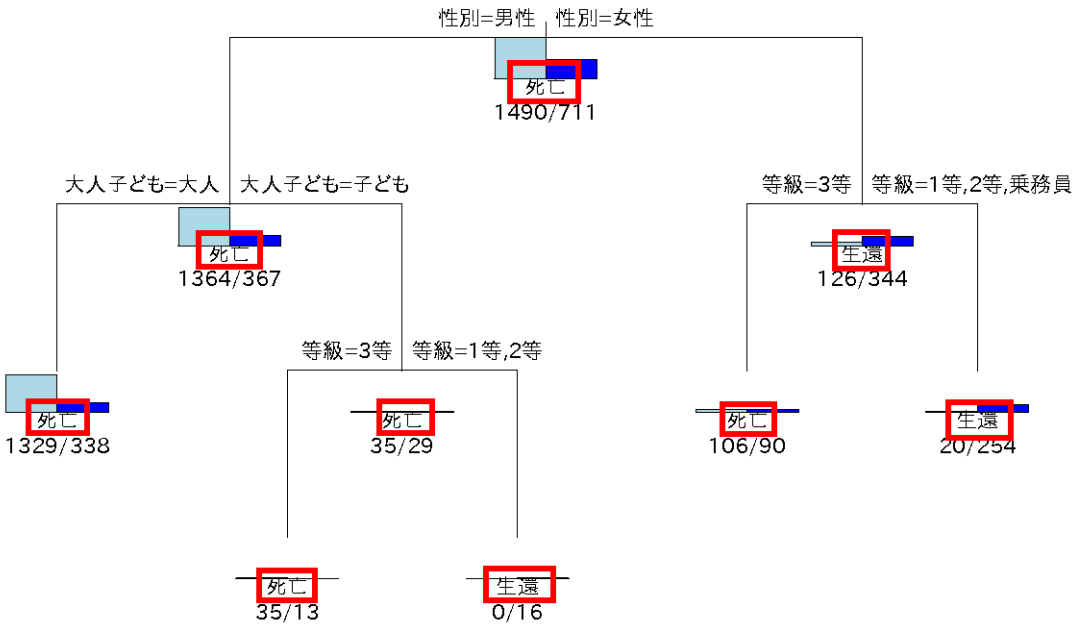


図 3.1: 決定木の例

アルゴリズム

今回の決定木は CART アルゴリズムにより作成した。CART では、説明変数を二進分岐させ、決定木を作成する。分岐の評価基準として、ジニ係数 (またはジニ分散指標とも呼ばれる) を用いている。分岐基準は決定木の分岐が生じる場所、つまり親ノードと子ノードの間において計算される。親ノード A は、すでに J 個の水準を持つカテゴリカルな基準変数 C によって $c_i(i = 1, 2, 3, ..., J)$ のように分割されている。

このとき親ノード A に属する観測値から任意に一つ選んで、それが c_j である確率を p_{Aj} とする。ここで、基準変数が水準 c_j である場合に 1、そうでない場合に 0 を取るダミー変数を考えると、水準 c_j の分散は $p_{Aj}(1 - p_{Aj})$ で表現できることが知られている。この分散は、$p_{Aj} = 0.5$ のときに 0.25 で最大となり、p_{Aj} が 1.0 と 0.0 に近づくに従って小さくなる。言い換えると、この分散が 0.5 であるとき、水準 c_j が当てはまる場合と当てはまらない場合が半々となり、最も判別のしにくい状態であることを表している。

ここから、不純度として親ノード A における基準変数 C の総分散を定義すると、

$$I(A) = i(P_A) = \sum_{j=1}^{J} p_{Aj}(1 - p_{Aj}) \tag{3.1}$$

となる。これがジニ係数と呼ばれるものである。

親ノード A が子ノード A_L と A_R に分岐する場合、以下の ΔI を最大化するような分岐基準を選択する。

$$\Delta I = P(A)I(A) - \{P(A_L)I(A_L) + P(A_R)I(A_R)\} \tag{3.2}$$

ここで、$P(\cdot)$ はそれぞれ分岐確率を表している。つまり、分岐確率を重みとする子ノードにおける不純度の平均 $P(A_L)I(A_L) + P(A_R)I(A_R)$ と、親ノードにおける不純度の差を計算することによって、分岐による誤分類の改善度を定義している。この分岐基準を候補にあがったすべての予測変数に関して計算し、値が最大になった予測変数で分岐をおこなう。

プルーニング

決定木は成長させればさせるほど見かけ上の成績が良くなる。基準変数が質的変数の場合には誤分類率が低くなり、量的変数の倍には予測の誤差分散が小さくなる。しかしこれは推定用のデータに対する成績であり、実際のデータにおける成績がより重要となる。決定木だけでなく他の学習器においても、複雑なモデルのほうが必ずしも成績が良いとは限らない。むしろ単純なモデルのほうが実務的な面、解釈容易性の面から見ても好ましい。そのため、予測への影響の少ない部分木を破棄するプルーニング (剪定、枝刈りとも呼ぶ) をおこなう。

決定木のプルーニング法には、推定用のデータのみを使う方法と、交差妥当化[6]用データや検証用データを併用する方法の 2 種類に大別される。以下で後者について流れを示す。

1. 検証用データを用い、推定用データに対する見かけ上の成績が頭打ちになるまで決定木を成長させる。

2. ターミナルノードを含む枝の中で、推定用データに関して成績のよくない部分に注目し、その枝があった場合とない場合の両方の成績を交差妥当化用データで計算する。

3. 交差妥当化用データに関して成績の落ちる枝はプルーニングする。

4. プルーニングした枝の部分はターミナルノードになるので、さらに 2、3 の過程を繰り返す。

[6]モデルの評価をおこなう場合に、そのモデルの母数の推定に用いたデータは利用せずに、それとは別に得られたデータへの当てはまりの良さを利用する方法。例えば、データを複数に分け、1 つの標本だけで母数を推定し、残りのデータで当てはまりを調べるという方法がある。

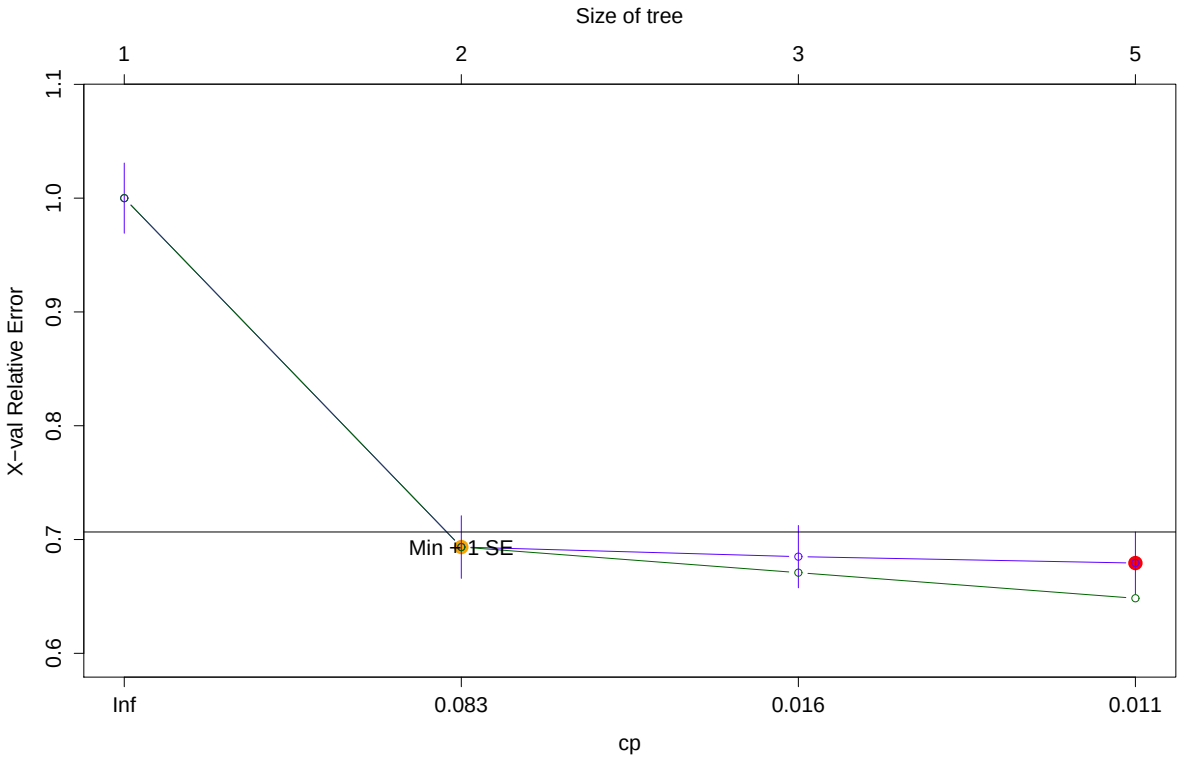


図 3.2: 決定木の大きさと複雑度
縦軸に予測変数の相対誤差を、横軸に木の大きさと複雑度を示す。

5. どのターミナルノードを含む枝をプルーニングしても、交差妥当化用データによる成績が下がるようであれば、プルーニングを終了する。

6. 度重なるプルーニングの過程で最終的な決定木と、交差妥当化用データは互いに統計的に独立ではなくなっていので、3 番目の独立したデータである検証用データを用意する。

7. 検証用データを用いて計算した最終的な決定木の誤分類率、誤差分散を、実践的運用における成績の目安とする。

プルーニングの程度を複雑度 (cp: complexity parameter) と呼び、この値が小さいほど決定木は複雑となる。複雑度と木の大きさと予測変数の相対誤差の関係を図 3.2 に示す。Min + 1SE とは、交差確認の結果から求められるリスクの最小値に標準偏差を足した値を示し、この図の例では cp=0.083 としてプルーニングをおこなう。この直線 Min + 1SE の下方で、最もその直線に近い点が示す複雑度を目安にプルーニングをおこなう手法を Min+1SE 法と呼び、頻繁に使用されている。本研究でもこの Min+1SE 法を用いてプルーニングをおこなう。

3.2 MeCabによる顔文字の形態素解析

3.2.1 形態素解析

形態素とは言語学の用語で、”意味の最小の単位”と説明される。例えば、”本を読んだ”というテキストは言語学の立場では”本”、”を”、”読ん”、”だ”と分割される。”読んだ”は五段活用動詞である”読む”と、過去を表す助動詞の”だ”で構成されていると解釈する。このようにテキストを形態素に分割することを形態素解析という。

3.2.2 顔文字の形態素

顔文字の解析をおこなうに当たり、MeCab を利用し、顔文字を目、口などの形態素に分類した。章 1.1 で述べたとおり自然言語処理は顔文字には適さないが、MeCab では分析者が独自に作成した辞書を作成して解析することができる。そこで本研究では、独自辞書として中島氏が公開している顔文字辞書[7]ver1.00 を利用した。

[7]http://www.haroperi.info/emoticon/annotated.html

第4章　事前分析—主観による感情解析

顔文字のみに焦点を当てた分析をおこなうに当たり、まず事前調査として顔文字のみでおこなう感情解析の効果を暫定的に確認した。

4.1　利用データ

2011.12.3 ~ 2011.12.28 において、毎日 10 分毎に Search API を用いてツイートデータを収集した。収集したデータの内、ツイートを一意に示す ID、ユーザー ID、投稿日時、投稿元クライアント情報、投稿内容の属性を利用した。一例を表 4.1 に示す。検索キーワードは”д OR Д OR ω OR ▽ OR ∇ OR △ OR > OR < OR ∀ OR ≦ OR ≧ OR ∪ OR ∧ OR ∨ OR _ OR __ OR ◇ OR □ OR Σ OR OR σ OR ε OR ▼ OR ゝOR ゞ” とした。表 4.2 にデータ量の要約を示す。

表 4.1: twitter から収集するデータの概要

ツイート ID	49139402818462
時間	23:49:45
ツイート	行かせていただきやす (‘・ω・´)
クライアント	Tween

表 4.2: 取得データ件数要約

	tweet count per day
Min.	12990
1st Qu.	13560
Median	17780
Mean	16060
3rd Qu.	18190
Max.	19140
sd.	2510.206

表 4.3: 顔文字の分類一例

喜	怒	哀
(*^^*)	(‘_ ′)	(・ω・‘=)
(^o^)	(‘д ′)	(′-ω-‘;)
(^-^)	(¯ ^ ¯)	(′• ω ˬ)
(^_^)	(° д °)	(′• ω • ˬ)

4.2 データ準備

ツイート ID によりツイートの重複を除外し、投稿日時を世界標準時から日本標準時に変換し、投稿内容における全角記号・半角記号を統一し、空白文字を削除した。各日において出現頻度上位 40 個の顔文字を集計し、合計 85 種類について主観により顔文字に”喜”、”怒”、”哀”の感情ラベルを割り当てた。顔文字と感情ラベルの一例を表 4.3 に示す。各日 10 分毎に顔文字の出現数を”喜”、”怒”、”哀”の各ラベルの感情量として集計した。

4.3 解析

以下の 3 点の視点から解析をおこなった。

1. 時間依存性
2. 投稿ツール依存性
3. 外的要因依存性

4.3.1 時間帯・曜日単位での感情比較

顔文字を含むツイートが人の手により投稿されていることを簡易的に確認する。ツイートは基本的に人の手により投稿されるため、投稿件数は時間依存性を持つ。この分析では顔文字の出現頻度を感情量としている。単位時間あたりの顔文字を含むツイート件数は単位時間当たりツイート件数に依存しており、顔文字の出現頻度も時間依存性を持つと考えられる。これを確認するため、全期間における 10 分単位の平均感情量の推移を図 4.1 に示す。図 4.1 では睡眠時間帯では感情量が少なく、通勤・通学時間帯や帰宅後である 19 時以降に感情量が増加傾向であることから、時間依存性があると考えられる。更に詳細に検討するため、全感情量の自己相関と偏自己相関を図 4.3,4.4、要約統計量を表 4.4、箱ひげ図を図 4.2 に示す。なお、図 4.3,4.4 において点線を超える場合は有意水準 $\alpha = 0.05$ にお

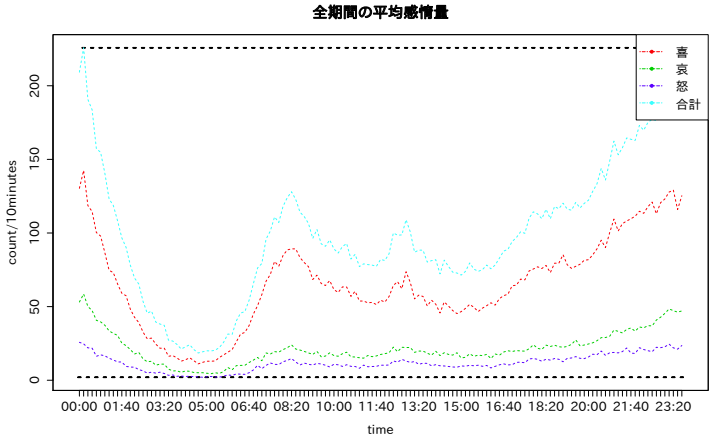


図 4.1: 全期間の感情量の推移

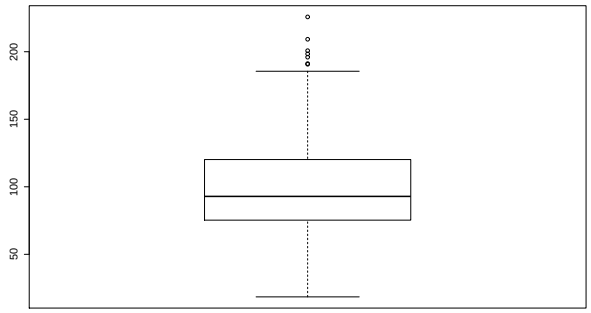


図 4.2: 感情量の要約統計量箱ひげ図

表 4.4: 感情量の要約統計量

	10 分当たり感情量
Min.	18.50
1st Qu.	75.31
Median	92.88
Mean	98.50
3rd Qu.	120.20
Max.	225.80
sd.	46.89

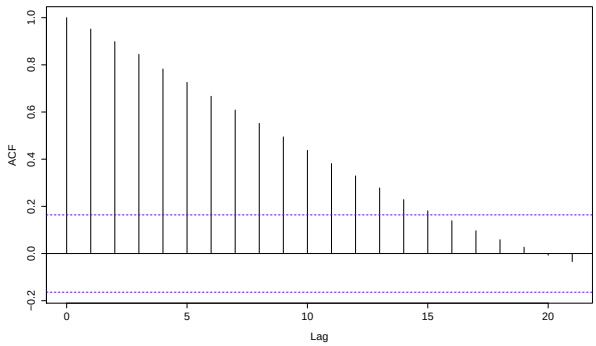


図 4.3: 全感情量の推移の自己相関

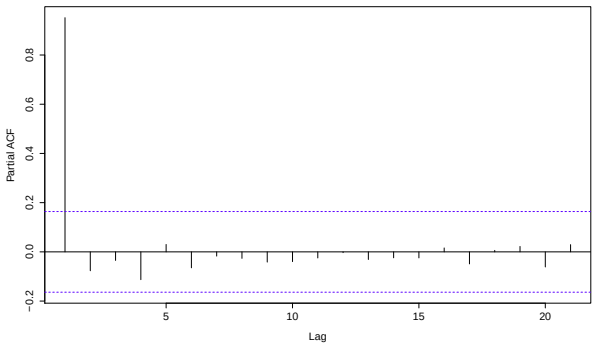


図 4.4: 全感情量の推移の偏自己相関

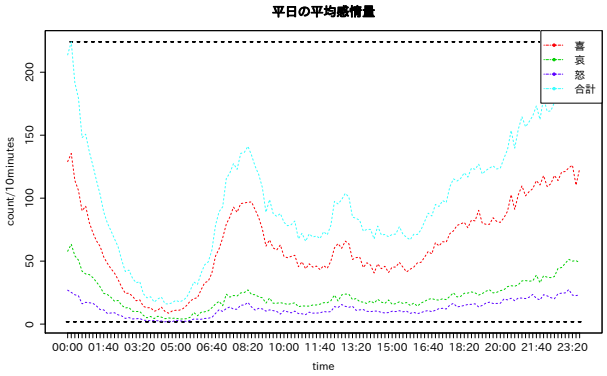


図 4.5: 平日の感情量の推移

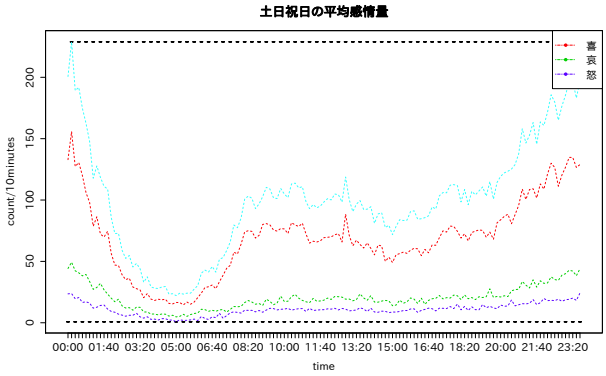


図 4.6: 土日祝日の感情量の推移

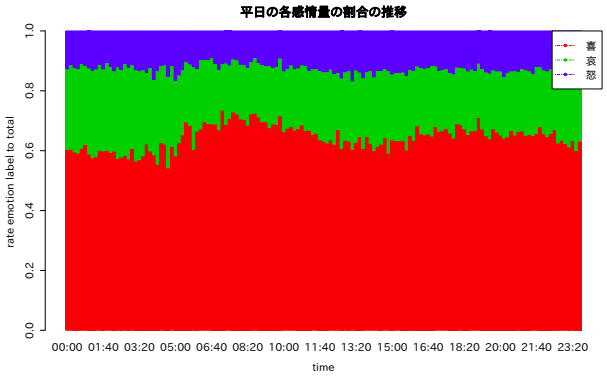


図 4.7: 平日の各感情の割合の推移

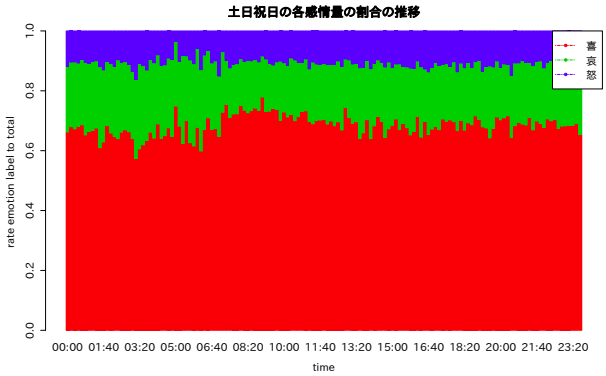


図 4.8: 土日祝日の各感情の割合の推移

いて有意差があることを示す。 図 4.3 から、時差 t < 4 において AIC は 0.8 以上となり、強い時間依存が確認できる。このとき PAIC は低い水準となっており、他の影響は少ないと考えられる。

次に平日と土日祝日における感情量の違いを確認した。この結果を図 4.5, 4.6 に示す。両者とも 0 時以降に感情量が減少し、6 時前後に上昇傾向が見られることが共通しており、就寝時間帯のツイート件数の減少に従い感情量も減少していると考えられる。図 4.5 では 8 時前後、13 時前後に主に”喜”の感情の増加のピークがあり、18 時以降に増加傾向が見られる。これらは通勤・通学時間帯、昼休憩時間帯、帰宅後の時間帯に対応しており、ツイート数の増加に従い感情量が増加していると考えられる。一方、図 4.6 においては平日と比べて昼の時間帯の感情量の減少が緩やかになっていると見られる。このことから、平日の昼の時間帯は仕事、学業などでツイートできない人たちが休日の同時間帯にツイートしていることが推測される。

次に、各感情ラベルの割合の推移を図 4.7,4.8 に示す。両者に特に違いは見られない。

表 4.5: 一日あたり投稿数上位 20 クライアント

クライアント	tweets per day	クライアント	tweets per day
Twitter for Android	2384	jigtwi	297
Keitai Web	1569	モバツイ / www.movatwi.jp	382
web	1445	TweetDeck	273
Twitter for iPhone	1442	Tween	254
twicca	1355	Janetter	254
ついっぷる/twipple	1129	yubitter	170
Twipple for Android	626	Mobile Web	108
ついっぷる for iPhone	522	Saezuri	101
SOICHA	462	Tweetbot for iPhone	98
Echofon	364	HootSuite	92

4.3.2 投稿デバイス依存性

投稿ツールにより感情表現に差があるのかを確認するため、投稿デバイス毎に集計し、考察をおこなった。ツイートの投稿元クライアント情報から PC、携帯電話・タブレットなどのモバイル端末のどちらから投稿されたツイートなのか分類した。Echofon や TweetDeck など PC、モバイルの両方に対応しており、かつ PC、モバイルの判別が不可能なツイートは集計から除外した。また、クライアント情報に”bot”、”ぼっと”、”ボット”のいずれかの文字列が含まれるツイートは自動投稿によるツイートと判断し、集計から除外した。クライアントの一覧を付録 C.1 の表 C.2 に、一日あたり投稿数上位 20 クライアントを表 4.5 に示す。PC、モバイルについて全期間の平均を算出した。この結果を図 4.9, 4.10 に示す。両者を比べると、モバイルのほうが発信される感情量が多くなっていることが分かる。これは一日あたり投稿数が多いクライアントの多くがモバイル用クライアントであるためであり、Twitter ユーザーの多くはモバイル端末から利用していると推定される。次に PC、モバイルについて各感情の割合の推移を図 4.11, 4.12 に示す。PC からの投稿では”喜”の感情の割合が 80%程度と高いのに対し、モバイルからの投稿では 60%程度となっている。

4.3.3 社会イベント等外的要因に対する依存性

外的要因の例として、次の 3 点について比較をおこなった。

1. 株価
2. 天気

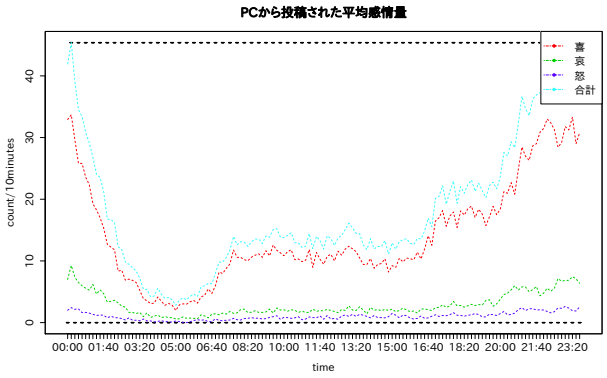


図 4.9: PC から投稿された感情量の推移

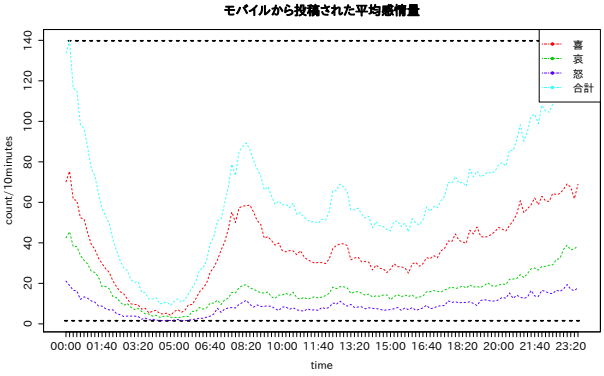


図 4.10: モバイルから投稿された感情量の推移

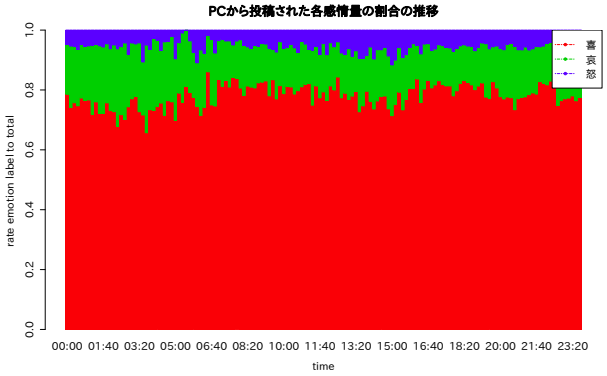


図 4.11: PC から投稿された感情量の割合の推移

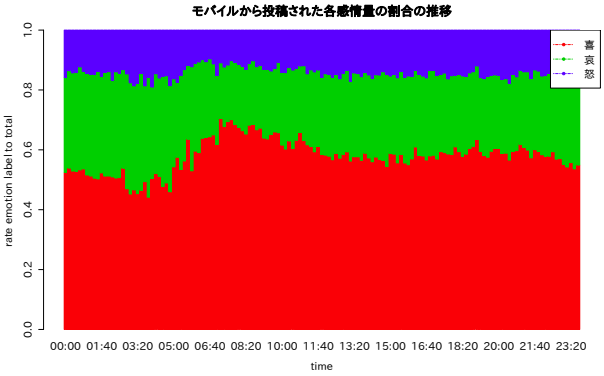


図 4.12: モバイルから投稿された感情量の割合の推移

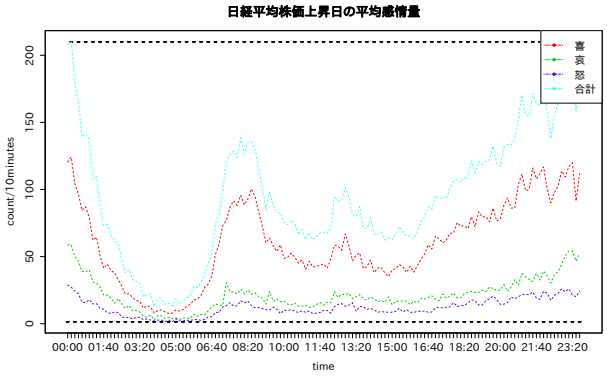


図 4.13: 日経平均株価上昇日の感情量の推移

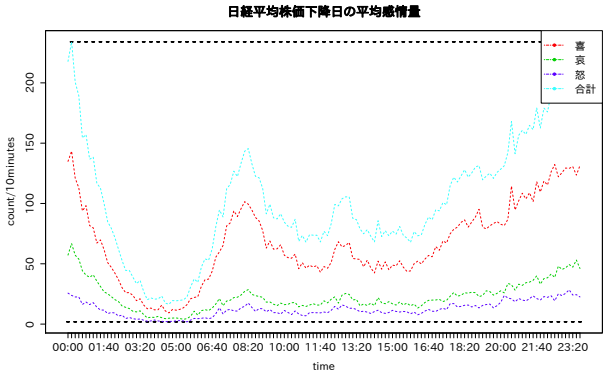


図 4.14: 日経平均株価下降日の感情量の推移

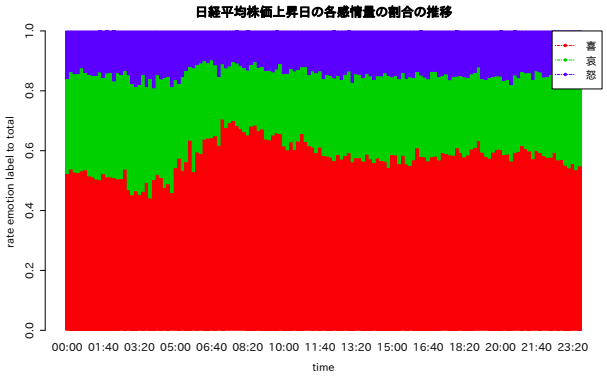


図 4.15: 日経平均株価上昇日の感情量の割合の推移

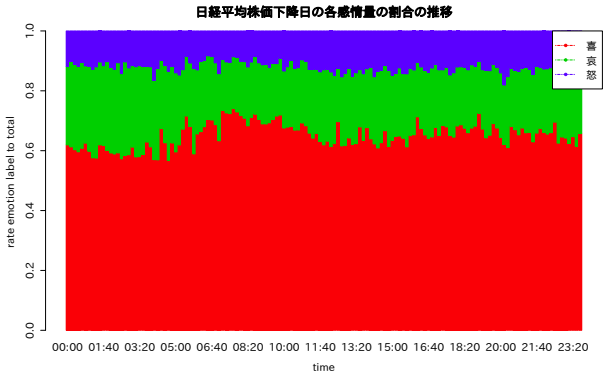


図 4.16: 日経平均株価下降日の感情量の割合の推移

3. イベント

日経平均による感情比較

Johan らの報告により、感情と株価の間に相関関係があることが示唆された [BMZ10]。そこで本研究では日本語のツイートを対象としていることから、事前調査として日経平均株価の変動と感情表現との相関関係を明らかにするため、日経平均の上昇日、下降日について比較をおこなった。YAHOO!ファイナンス[1]から 2011.12.3 ~ 2011.12.28 の期間の中で日経平均株価が前日と比較して上昇した日と下落した日を調査し、集計をおこなった。この結果を図 4.13, 4.14, 4.15, 4.16 に示す。両者を比較した結果、0 時〜6 時の就寝時間帯における”喜”の割合に差があるように見られる。もう少し考察。しかし Johan らの報告は 3 日後の DJIA を予測したものであるため、感情と株価との関係については時差について検討する必要がある。

[1] http://finance.yaho.co.jp/

表 4.6: 晴・曇・雨または雪の該当日数

晴	曇	雨又は雪
20	6	0

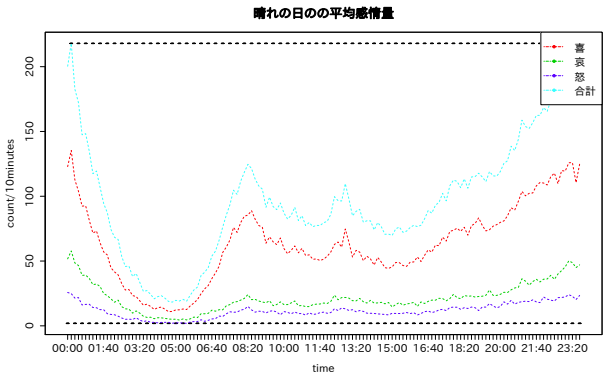


図 4.17: 晴れの日の感情量の推移

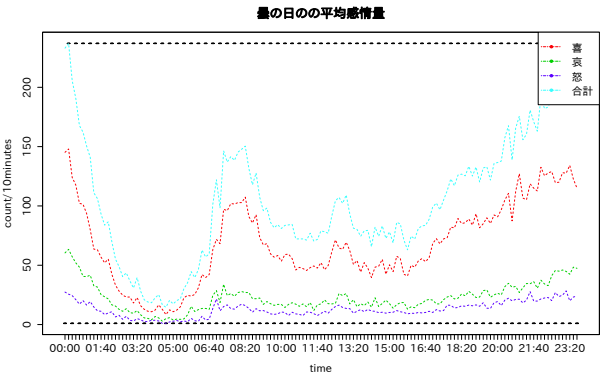


図 4.18: 曇の日の感情量の推移

天気による感情比較

福岡がうつ病を季節性感情障害の一種として捉えた考察をおこなっている [福岡 03]。この中で自殺件数に関して気象と精神状態の関係という視点から考察をしており、気象条件が精神状態に影響を与えることが示唆された。このことから、天気と感情の関係について考察するために東京都の天気に対して、感情と各天気との比較をおこなった。日本気象協会[2]から 2011.12.3 ~ 2011.12.28 の東京都の天気を調査し比較をおこなった結果を図 4.17, 4.18, 4.19, 4.20 に示す。表 4.6 に晴・曇・雨又は雪の各天気の該当数を示す。両者を比較した結果、曇の日は 8 時前後の通勤・通学時間帯においてピークの持続時間が長いという特徴が見られる。今回は日本の人口の分布の観点から東京都の天気について比較をおこなったが、東京都の天気は前日の静岡県の天気と関係がある。そのため、静岡県の前日の天気についても検討する必要があると考えられる。

イベントの有無による感情比較

次に、イベントの有無について感情がどのように反応するのかを明らかにするため、クリスマス (2011.12.24,25) を例として、クリスマスと全期間について比較をおこなった。この結果を図 4.21, 4.22, 4.23, 4.24 に示す。図 4.21, 4.22 を比較すると、クリスマスのイベント時期には全体よりもツイート件数が増加していることが確認される。また、図 4.23,

[2]http://tenki.jp/

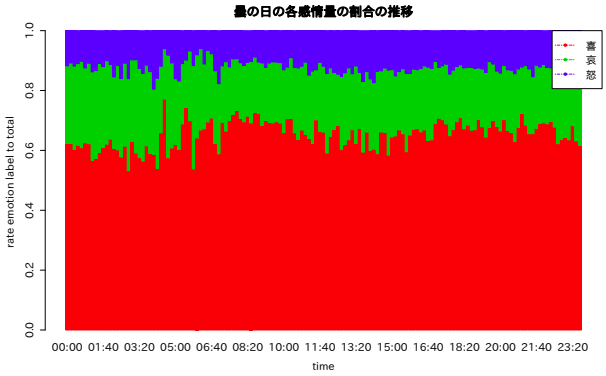


図 4.19: 晴れの日の感情量の割合の推移

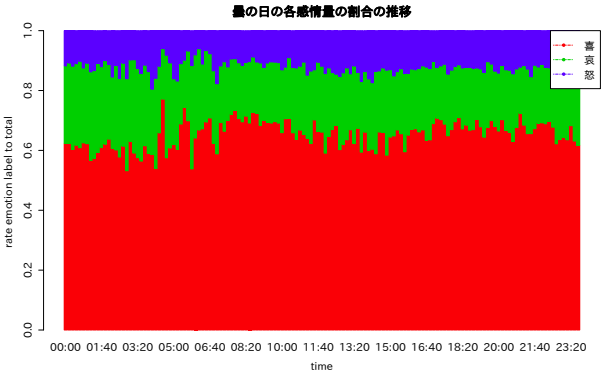


図 4.20: 曇の日の感情量の割合の推移

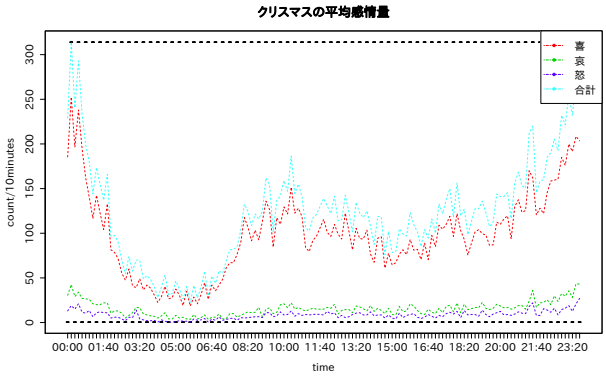


図 4.21: クリスマスの感情量の推移

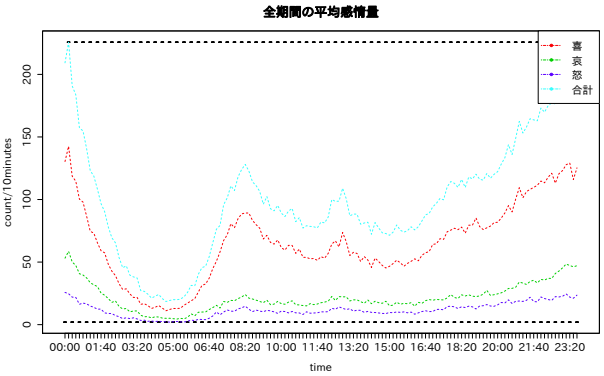


図 4.22: 全期間の感情量の推移

4.24 の比較により、すべての感情が一様に増加するのではなく、”喜”の感情のみが増加していることが確認された。

4.4 結果

Search API により取得したデータから顔文字を抽出し、主観により感情ラベルを割り当て、10 分毎に集計し、各日当たりの感情量の平均と各感情の割合から考察をおこなった。章 4.3.1 で確認したとおり、通勤・通学時間帯や昼休憩時間帯に感情量の増加が確認され、感情量は時間に依存することが示唆された。また、感情量の推移において”喜”、”怒”、”哀”の感情の増減率には差があることが示唆された。

章 4.3.2 では投稿ツールにより表現される感情に違いがあるか調査をおこなった。

章 4.3.3 では外的要因による依存性として、日経平均株価、気象条件、クリスマスを例として比較をおこなった。その結果、日経平均株価と気象条件については顕著な差異は見られなかったが、クリスマスにおいては顕著な差を示した。クリスマスではツイート件数の増加に伴い、表現される感情量がそれ以外の日と比べて増加している。また、”喜”、”

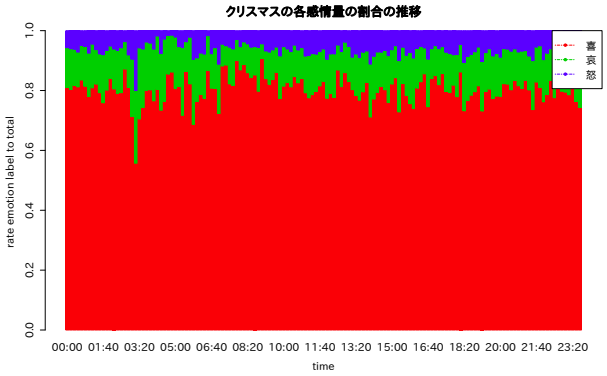


図 4.23: クリスマスの感情量の割合の推移

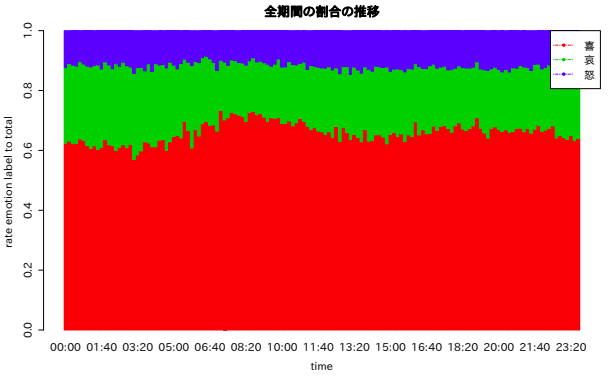


図 4.24: 全期間の感情量の割合の推移

怒”、”哀”の各感情の表現量が一様に増加するのではなく、”喜”の感情のみが顕著な増加を示した。この結果から、クリスマスというイベントでは Twitter 上の感情のトレンドはポジティブな感情となっていることが示唆さ、世間がクリスマスを祝うように、クリスマスはポジティブなイベントであると示唆された。

4.5 課題

感情量の推移について考察をおこなったが、感情量の増減がツイート件数に依存しており、そのため今回の分析結果が得られた可能性がある。故に、ツイート件数を考慮に入れるため Streaming API を用いてデータを収集し、再度分析をおこなう必要がある。

次に、感情ラベルの割り当てが主観的であった点に課題がある。喜の感情を割り当てた顔文字に対し、”怒”、”哀”の感情を割り当てた顔文字の個数が少ないために、分析結果が正しく得られなかった可能性がある。そのため、より客観的な感情ラベルの割り当て手法を適用し、再度分析をおこなう必要がある。

第5章　顔文字解析—決定木による感情推定

5.1　決定木を用いた単一モデルによる顔文字の感情推定

章 4 では感情ラベルの推定が主観的であったところに課題があった。そこで推定を定量的におこなうため、決定木分析により推定のルール作成を試みた。多くの顔文字は目・口・手などを形態素として持っている。そこで、この形態素に着目した分析をおこなう。

5.1.1　利用データ

教師データとして顔文字の館[1]のサイトに掲載されている、”笑う”、”泣く”、”怒る”、”驚く”、”落ち込む”、”照れる”の感情ラベルを持つ顔文字を使用した。検証データとして Streaming API により取得した 2012.7.1 ~ 2012.7.31 までのツイートを使用した。

5.1.2　データ準備

まず教師データについて、顔文字に付随するテキストによって感情を示しているものを削除し、章 2.2.1 で定義した顔文字に当てはまらないものを削除した。その上で、明らかに感情ラベルと合致しないものを削除した。その結果、255 件の教師データを得た。教師データにおける各感情ラベルの内訳を表 5.1 に示す。顔文字が表す感情は目、口、頬、手、イメージ (”∑”マークや”;”など) の各形態素が示す極性に影響を受ける、という仮説をたて、主観により各形態素に Positive/Negative/Neutral のパラメータを設定した。

検証データとして、顔文字を含むツイートをランダムに 6,000 件取得し、ツイートに含まれる各顔文字に上記と同様にパラメータを設定した。

5.1.3　分析

教師データを用いて CART アルゴリズムによる決定木を作成した。この結果を図 5.1 に示す。なお、文字の重なりあいによる視認性の低下を軽減するため分岐条件はラベル化してあるが、詳細を付録 C.2.1 の図 C.1 に示す。図 5.1 において、例えば表 5.2 の (^_^)v の

[1] http://yakata.if.tv/pc/kao/

表 5.1: 教師データにおける各感情ラベルの内訳

笑い	泣く	怒る	驚く	落ち込む	照れる	合計
78	37	79	39	0	22	255

表 5.2: 教師データ一例

顔文字	目	口	頬	イメージ	手	ラベル
(‘- ′メ)	-2	0	0	$xmark$	0	angry
o(¯ ^ ¯ o)	1	0	0	0	hand	angry
ｐ (≧へ≦) ｑ	2	0	0	0	arm	cry
(^_^)v	1	0	0	0	peace	laugh
(⌒∇⌒)V	1	1	0	0	peace	laugh
(*^o^*)	1	1	1	0	0	shy
(*° ◇°)	5	0	1	0	0	surprise

場合、目のパラメータは 1 であり、ルートノードの分岐基準は目 > -6.5 である。よって目 ≧ -6.5 を示す右側へ分岐する。次の分岐では目 ≧ 0.5 を示す右側へ分岐し、目 < 3.5 を示す左側へ分岐する。その結果分類される感情ラベルは”笑い”である。なお、決定木を作成する場合は通常はプルーニングをおこなうが、訓練データの分類精度を向上させるためにプルーニングをおこなっておらず、複雑度は 0.01 と設定した。この結果、作成した決定木の分類精度は 81.2%となった。分類の詳細を表 5.3 に示す。

次に検証データでの精度を確認するため、ツイート本文からパターンマッチングにより顔文字を抽出し、決定木を適用した。分類精度を確認するため筆者ら 6 人により人手で確認をおこなった結果を表 5.4 に示す。”笑う”、”泣く”については良い精度で分類できているが、その他の感情については精度が悪い結果となった。

教師データの”怒る”のラベルを持つ顔文字は”#”を使用した顔文字が多く、特定の特徴を持つ顔文字しか教師データに含まなかったため、精度が低くなってしまったと考えられる。また、”Σ (° д°)”のような”Σ”を含む顔文字は”怒る”、”驚く”の両方の意味を表すことができ、感情を特定するには文脈から判断する必要がある。このような顔文字の多義性により、精度が低くなってしまったと考えられる。

表 5.3: 分類精度

教師データ / 決定木の分類結果	怒る	泣く	笑う	照れる	驚く
怒る	57	0	1	0	0
泣く	1	29	1	0	1
笑う	6	3	71	13	4
照れる	2	1	2	7	0
驚く	13	4	3	2	34
正解率	72.2%	78.4%	91.0%	59.1%	81.2%

表 5.4: 検証データの正解率

	笑う	泣く	怒る	照れる	驚く	合計
正解個数	357	134	13	22	37	563
分類個数	365	136	36	115	470	1122
正解率	97.8%	98.5%	36.1%	19.1%	7.87%	50.2%

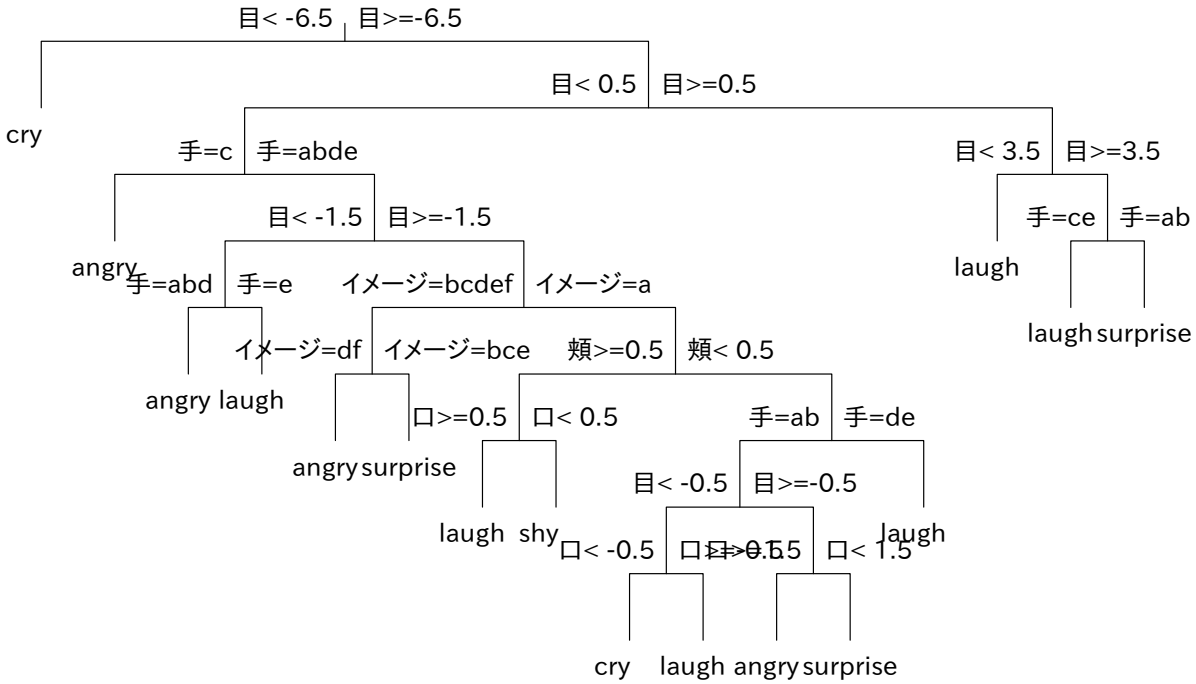

目< -6.5
目>=-6.5
目< 0.5
目>=0.5
cry
手=c
手=abde
目< 3.5
目>=3.5
目< -1.5
目>=-1.5
手=ce
手=ab
angry
手=abd
手=e
イメージ=bcdef
イメージ=a
laugh
laugh
surprise
イメージ=df
イメージ=bce
頬>=0.5
頬< 0.5
angry
laugh
口>=0.5
口< 0.5
手=ab
手=de
angry
surprise
目< -0.5
目>=-0.5
laugh
shy
口< -0.5
口>=0.5
口< 1.5
cry
laugh
angry
surprise

図 5.1: 決定木

5.1.4　結果

定量的な感情ラベル推定手法として、決定木を用いた単一次元モデルによる感情推定手法を試みた。教師データでは良い分類精度を示したが、検証データでは一部の感情ラベルの分類精度が悪い結果となった。これは教師データに十分な種類の顔文字を含められなかったためと考えられる。しかし顔文字の種類は現在なお増加し続けており、これに対し正解ラベルを人手で付与することには限界がある。そのため、顔文字辞書のみを利用した手法では単一モデルによる感情推定は困難であることが明らかになった。また、顔文字が表す感情は文脈に依存することもあり、単一モデルにおいては顔文字辞書のみを用いた感情推定は困難であることが明らかになった。

5.1.5　課題

”笑う”、”泣く”の感情ラベルに対しては良い予測精度であったが、目・口などの形態素のパラメータを主観的に決定していた点に課題が残る。検証用データに対する予測精度の低さが課題である。今回の分析では一つの顔文字に対応する感情を一意に決定しようとしており、顔文字の表す感情は文脈に依存する多様性が考慮されていないという課題がある。

5.2　決定木を用いた多次元モデルによる顔文字の感情推定

章 5.1 では顔文字辞書のみを用いた分析では顔文字と感情ラベルの一対一対応が困難であることが明らかになった。この問題に対応するためには、ツイート本文のテキスト情報から感情を抽出し顔文字解析に活用する方法と、顔文字の感情を一つに特定するのではなく各感情ラベルのもっともらしさ、つまり、各感情をどの程度表すか、という視点から確率で表現する方法がある。

前者にはツイート本文のテキストの字数制限から十分な長さのテキストが得られるとは限らないという懸念があるため、顔文字の感情を確率表現を用いて多次元モデルにより表現する後者の手法を試みた。

5.2.1　利用データ

川上がアンケート調査により顔文字が表す感情を定量的に調査した結果を報告している [川上 08]。まずこのアンケートデータに含まれる 31 個の顔文字について、全角・半角記号を統一した。これにより”(￣□￣;)”の表記揺れを解消し、得られた顔文字を形態素解析により分割した。誤分類のあった 2 つの顔文字を除き、28 個の教師データを得た。

5.2.2 分析

次に各形態素を説明変数、教師データの感情スコアを非説明変数として、決定木による分析をおこなった。枝刈りをおこなうに当たり、分岐数と複雑度の関係を図 5.2、詳細を付録 C.2.2 の表 C.3 に示す。図 5.2 より、cp = Inf を除きいずれも直線 Min + 1SE の上方に点がある。今回は最も Min + 1SE に近い cp = 0.091 を複雑度とし、プルーニングをおこなった。これにより得られた決定木を図 5.3 に、詳細な分類規則を付録 C.2.2 の表 C.2.2 に示す。 作成した決定木により、訓練データの感情スコアを算出した。一例を表 5.5 に

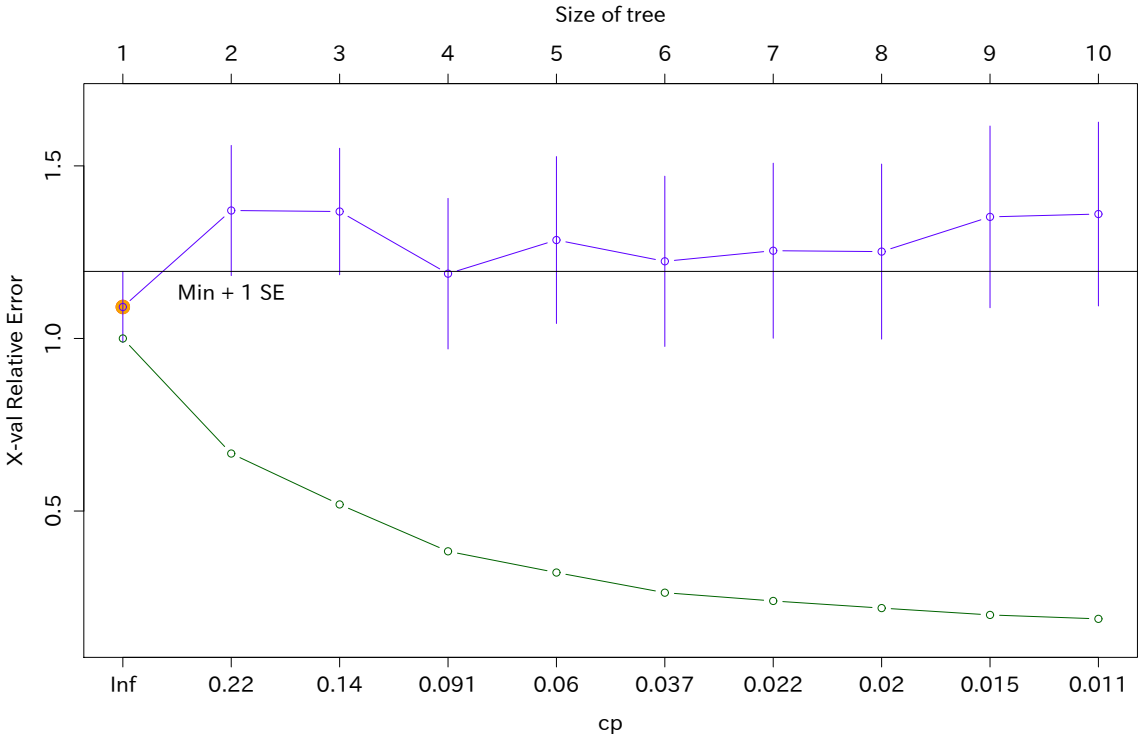


図 5.2: 分岐数と複雑度の関係

示す。

5.2.3 検証

顔文字を多次元モデルにより表現する手法の妥当性を検証する。算出された感情スコアと元の感情スコアとの相関係数を調査した結果を表 5.6 に示す。

相関係数および p 値に着目すると、”喜び”、”哀しさ”、”楽しさ”、”焦り”、”驚き”については元の感情スコアと推定した感情スコアに強い正の相関関係があると認められるが、”怒り”、”強調”については強い相関関係があるとは言えない。次に検定力に着目すると、”強調”では検定力が高くなく、第二種の誤りを考慮する必要がある。それ以外につい

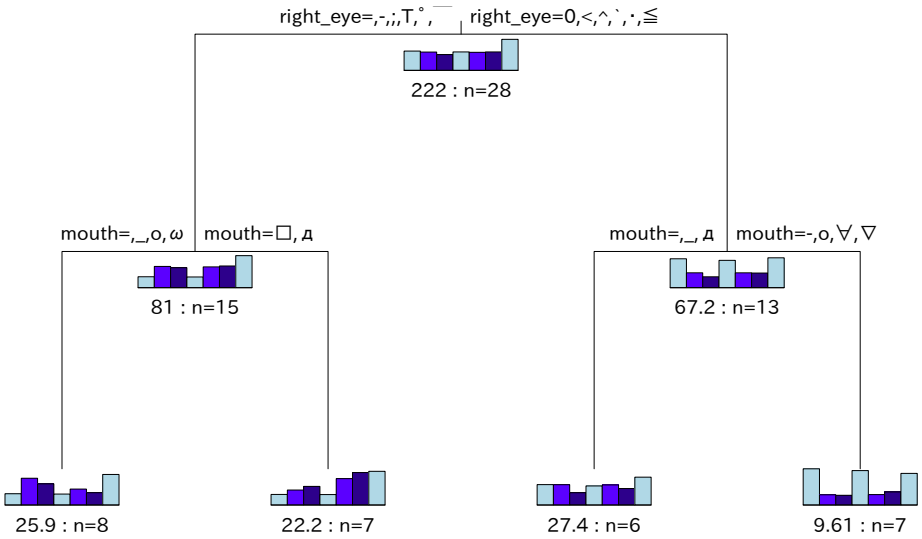


図 5.3: 決定木

表 5.5: 決定木による推測の例

顔文字	喜び	哀しさ	怒り	楽しさ	焦り	驚き	強調
(≧∇≦)	4.147	1.173	1.114	3.941	1.193	1.501	3.621
(>_<)	2.337	2.327	1.410	2.185	2.302	1.885	3.190
(ToT)	1.291	3.066	2.426	1.255	1.826	1.411	3.500

ては検定力が大きく、第二種の誤りの可能性はほぼ無いと言え、訓練用データにおいて正しく感情スコアを推定できたと言える。

また、検証用データで感情の推定をおこなった。Twitter Streaming API を使用して取得したデータから顔文字を抽出し、出現頻度上位 30 件の顔文字について今回作成した決定木で感情を推定できるか調査した結果を表 5.7 に示す。30 件中 15 件において、感情の推定が不可能であった。これは教師データ中の形態素として使用されていない記号が形態素として使用されるために起きる問題であり、教師データの拡充が不可欠であると明らかになった。また、今回使用した訓練データに形態素解析を適用し、各形態素の種類を表 5.8 に示す。これより、$3 \times 12 \times 1 \times 4 \times 9 \times 3 \times 2 \times 13 \times 1 \times 2 \times 2 = 404,352$ 種類の顔文字の表現が可能であり、これが今回作成した決定木により解析可能な顔文字の種類である。しかし、これには右眉が含まれていないため、表現可能な種類が 40 万程度であっても実

表 5.6: 決定木による推定スコアとアンケートによるスコアの相関検定

感情ラベル	相関係数	95%信頼区間	t-value	p-value	検定力
喜び	0.882	.759, .944	9.5635	5.325e-10**	0.9999
哀しさ	0.685	.420, .844	4.8001	0.00005696**	0.9881
怒り	0.556	.239, .770	3.4128	0.002115**	0.8827
楽しさ	0.874	.743, .941	9.1822	1.215e-09**	0.9999
焦り	0.742	.509, .873	5.6345	0.000006353**	0.9977
驚き	0.852	.702, .930	8.2979,	8.852e-09**	0.9999
強調	0.388	.018, .665	2.1476	0.04124*	0.5339

すべてのケースで DF=26

** p < 0.01 * p < 0.5

表 5.7: 出現数上位 30 位の顔文字

顔文字	出現数	推定	顔文字	出現数	推定
(^o^)	744220		(･ω･)	324062	ー
(>_<)	257524		(∇)	217750	
(*^^*)	214605	ー	(^^)	208921	
(^ω^)	176230		o(*∇*)o	172249	ー
(;ω;)	148469	ー	(･_･)	143564	ー
(д)	137336	ー	(∇)	126041	
(;_;)	114943		(･∀･)	112085	
(T_T)	109725		(^-^)	107427	
(ω)	101682	ー	(･ω･)	94735	
(*ω*)	89952	ー	(;∇;)	81952	
(^o^)/	81752	ー	(*∀*)	77188	ー
(*∇*)	76699	ー	(･ω･)	66480	ー
(≧∇≦)	65810		(*∇*)	65152	ー
(〃∇〃)	64054	ー	(∀)	62539	
(д)	61138		(;д)	59211	ー

感情推定が不可であったものには”ー”を記す。
出現数は 2012.10.01~2012.12.28 の合計である。

表 5.8: 訓練データに含まれる形態素の種類

右手	右目	右輪郭	右頬	口	左手	左眉	左目	左輪郭	左頬	動線
3	12	1	4	9	3	2	13	1	2	2

際の解析では解析に適さない顔文字が多くなってしまう。

5.2.4 結果

顔文字を形態素に解析し、決定木分析により各感情ラベルのもっともらしさ利用して感情を多次元モデルで表現する手法を提案した。大学生にアンケートを実施し、顔文字が各感情をどの程度表すか、を調査したデータベースを利用し、相関係数及び検定力分析の観点から手法の評価をおこなった。”怒り”、”強調”の感情スコアの推定において、元の感情スコアとの推定スコアとの小〜中程度の正の相関関係が見られる。”強調”においては検定力が高くなく、第二種の誤りの可能性がある。”喜び”、”哀しみ”、”楽しさ”、”焦り”、”驚き”、の感情については元の感情スコアと推定スコアとの強い正の相関関係が見られ、かつ第二種の誤りの可能性もほぼ無いと言える。このため、提案手法において”喜び”、”哀しみ”、”楽しさ”、”焦り”、”驚き”の感情スコアは正しく推定できると考えられる。

5.2.5 課題

本手法においても章 5.1 と同様に、訓練データに含まれない形態素を持つ顔文字は感情スコアを計算できないという課題が残る。訓練データはデータ数が 30 程度と少なく、感情推定可能な顔文字の種類に大きな課題がある。

また解析可能な顔文字の種類を増やすため、訓練データの拡充が課題となる。しかし顔文字の種類は現在なお増加し続けているため、正解ラベル作成のコストの面から訓練データの収集は今後ますます困難になると考えられる。

5.3 結果

本章では顔文字を形態素に分割し、形態素から感情を推定する単一モデルと多次元モデルを考案し、これらについて評価をおこなった。

単一次元モデルでは、顔文字の感情の特定を試みる場合、顔文字のみを用いた推定には限界があることが明らかになった。今回は利用していないが、顔文字の感情推定にテキストを併用することで、精度が向上する可能性があると考えられる。

多次元モデルでは、顔文字が各感情をどの程度表すか、という視点から確率を用いて感情スコアを表現した。アンケートにより得られた訓練データを用いて、これらの顔文字の形態素から決定木分析により各感情スコアを推測した。元の感情スコアと算出した感情スコアの相関係数、検定力分析による考察をおこない、多次元モデルにおいて”喜び”、”哀しみ”、”楽しさ”、”焦り”、”驚き”の感情スコアは正しく推定可能であると結論づけた。

次章の分析においては、この感情スコアが正しいものと仮定しておこなう。

5.4 課題

顔文字の種類は現在なお増加しており、正解ラベルを含む訓練データを用意することが非常に困難である。そのため、正解ラベルの含まれないデータにおいても正しく予測・分類が行える分析手法の適用が求められる。

第6章　SNS感情トレンドと社会トレンドとの関係

6.1　利用データ

Streaming API により収集したデータを利用した。ツイートを一意に示すツイート ID、発信された時間 (世界標準時)、ツイート内容を収集し、世界標準時から日本標準時への変換、全角・半角記号の統一を行い、ツイート ID の比較から重複するデータのないようにした。収集したデータの一例を表 6.1 に示す。

取得したデータから顔文字を抽出し、URL や”(笑)”といった文字列を取り除いた。考察を行うにあたり、前章で提案した決定木を用いた多次元モデルによる顔文字の感情推定手法を利用した。以降でおこなった比較に際しては、2012.10.01~2012.12.26 までのデータを使用している。

表 6.1: twitter から収集するデータの概要

ツイート ID	266482307072937984
時間	2012-11-08 19:08:27+09
ツイート	帰宅 (′ω‘)

6.2　データ要約

全期間、平日・土日祝日、イベントでデータを分類し、各日 10 分毎のツイート数の平均値を集計した。

6.2.1　全期間

全ツイート

ツイート数の要約統計量を表 6.2 に、箱ひげ図を図 6.1 に、ヒストグラムを図 6.2 に示す。これらの結果より、ツイート数は平均値未満ではばらつきが小さく、平均値以上では

表 6.2: ツイート数の要約統計量 (tweets/10minutes)

Min.	708.6
1st Qu.	2285.0
Median	2888.0
Mean	3344.0
3rd Qu.	4391.0
Max.	7124.0
sd	1746.559

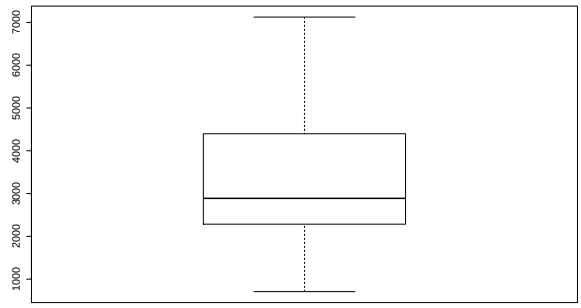


図 6.1: ツイート数の箱ひげ図

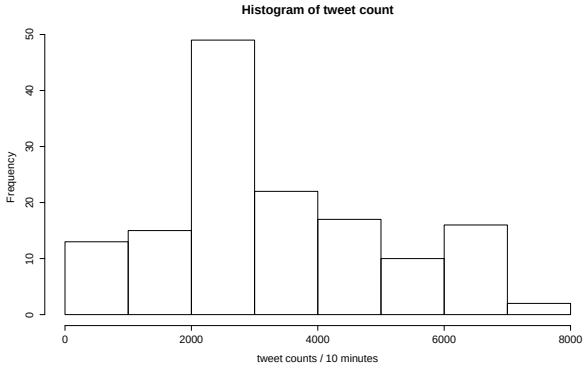


図 6.2: ツイート数の分布

ばらつきが大きいことが分かる。つまり、ツイート数の減少方向には振れにくいが増加方向には振れやすいことが分かる。

顔文字を含むツイート

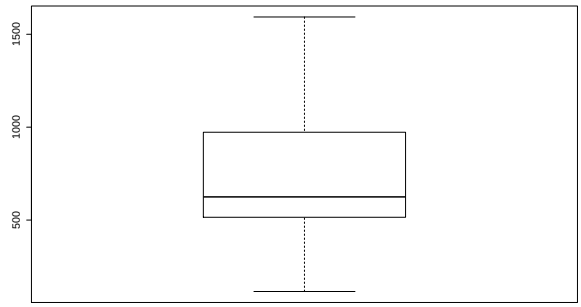
顔文字を含むツイート数の要約統計量を表 6.3 に、箱ひげ図を図 6.3 に、ヒストグラムを図 6.4 に示す。これらの結果より、顔文字を含むツイート数は前述の全ツイートの場合と同様に平均値未満ではばらつきが小さく、平均値以上ではばらつきが大きいことが分かる。

顔文字を含むツイートの割合

顔文字を含むツイートの割合の要約統計量を表 6.4 に、箱ひげ図を図 6.5 に、ヒストグラムを図 6.6 に示す。これらの結果より、全ツイート数に対する顔文字を含むツイート数の割合は平均して 21.0%であり、ヒストグラム及び標準偏差から、変動幅は大きくないと言える。

表 6.3: 顔文字を含むツイート数の要約統計量 (tweets/10minutes)

Min.	115.6
1st Qu.	515.8
Median	624.7
Mean	722.6
3rd Qu.	972.0
Max.	1593.0
sd	402.993



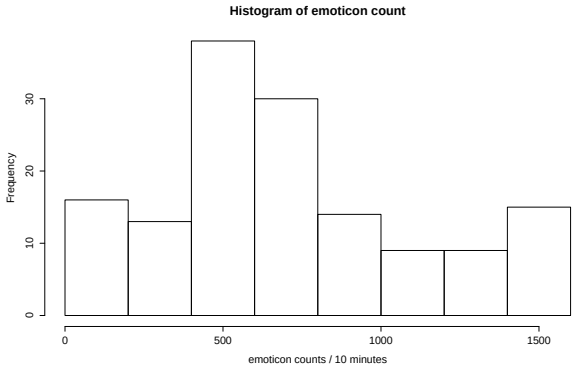


図 6.3: 顔文字を含むツイート数の箱ひげ図　図 6.4: 顔文字を含むツイート数の分布

表 6.4: 顔文字を含むツイート数の割合の要約統計量

Min.	0.1494
1st Qu.	0.2056
Median	0.2150
Mean	0.2100
3rd Qu.	0.2242
Max.	0.2547
sd	0.02329

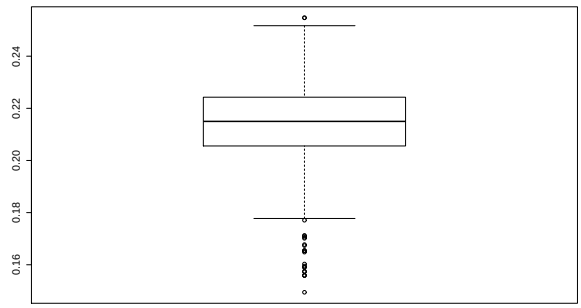


図 6.5: 顔文字を含むツイートの割合の箱ひげ図

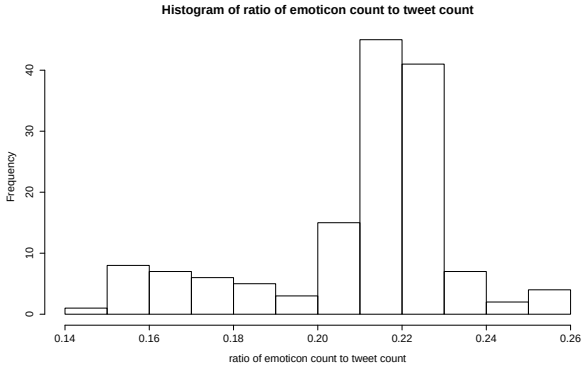


図 6.6: 顔文字を含むツイートの割合の分布

まとめ

全ツイート数、顔文字を含むツイート数、割合を図6.7に示す。この図より、7時前後に顔文字を含むツイート数の割合の増加のピークがみられる。増加率という観点では5:30~7:20にピークがあり、起床時間帯に合致している。

6.3 解析

解析は以下の手順でおこなった。

- 1分毎にデータを区切り、出現した顔文字を集計する
- 各顔文字に章5.2で作成した決定木のルールに基づき、感情スコアを付与する
- 1分毎に顔文字の感情スコアを集計、平滑化し、感情量とした
- 感情量の時間推移から考察をおこなう

なお、平滑化はtを現在の時刻、$\Delta t = 10minutes$として、以下の式によりおこなった。

$$score_{label} = \frac{N_{t-\Delta t} + N_{t-(\Delta t+1)} + ... + N_t + N_{t+1} + ... + N_{t+\Delta t}}{2\Delta t} \quad (6.1)$$
$$label : joy, sad, angry, fun, hurry, surprise$$

なお、Nは1分間の間に出現した顔文字の各感情ラベルのスコアを合計したものである。

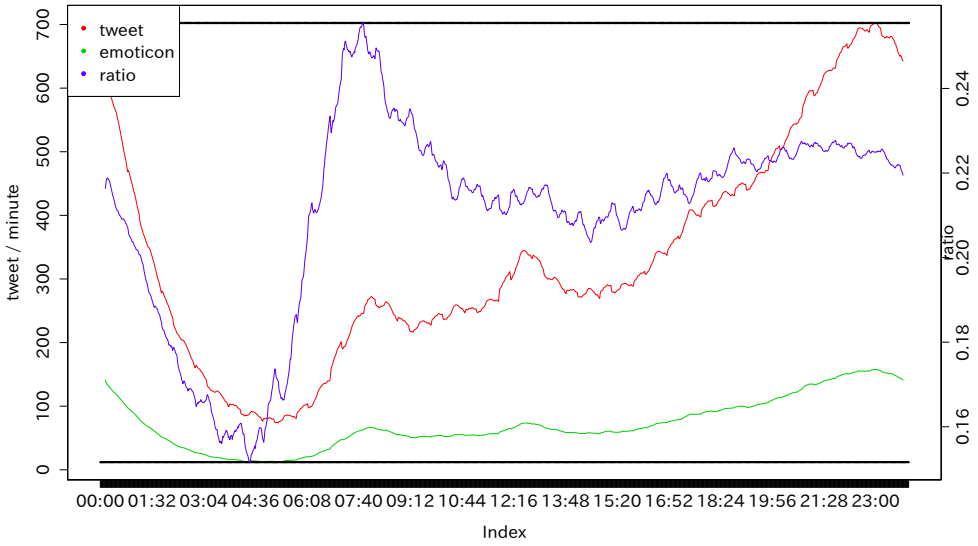


図 6.7: 全ツイート数、顔文字を含むツイート数、割合
第一 Y 軸に全ツイート数 (赤)、顔文字を含むツイート数 (緑) を、
第二 Y 軸に全ツイート数に占める顔文字を含むツイート数の割合 (百分率)(青) を示す

6.3.1 時間帯、曜日単位での感情比較

日本時間 (JST) を基準とし、24 時間における感情量の推移を曜日ごとに考察をおこなった。

図 6.7、表 6.6 より、顔文字を含むツイートの割合は 21%程度で安定しており、顔文字のみを使用した感情解析においても安定した結果を期待できる。平日と土日祝日の感情推移を図 6.8,6.9 に示す。平日と土日祝日には、まず感情量の最大値に違いがみられる。土日祝日の方が最大値が高く、また昼の時間帯の感情量の推移も高い状態を維持している。平日では 8 時前後の通勤・通学時間帯、12 時前後の昼休憩の時間帯にピークがあり、帰宅後の 18 時以降に感情量の継続的な増加傾向がみられるのに対し、土日祝日では 7 時前後の起床時間後はほぼ一貫して継続的な増加傾向がみられる。

章 4 での結果と同様に、平日は通勤通学時間・昼休憩・帰宅後の時間帯に特徴がある。

6.3.2 社会イベント等、外的要因に対する感情表現の依存性

4 章から、ポジティブな外的要因と感情の間に関係がある可能性が示唆された。そこでここではポジティブ/ネガティブな外的要因によって感情がどのように反応するのかについて考察した。なおそれぞれの要因に、世間一般でポジティブなイベントと捉えられるクリスマスを、ネガティブなイベントとして捉えられる地震発生日を例として比較をおこ

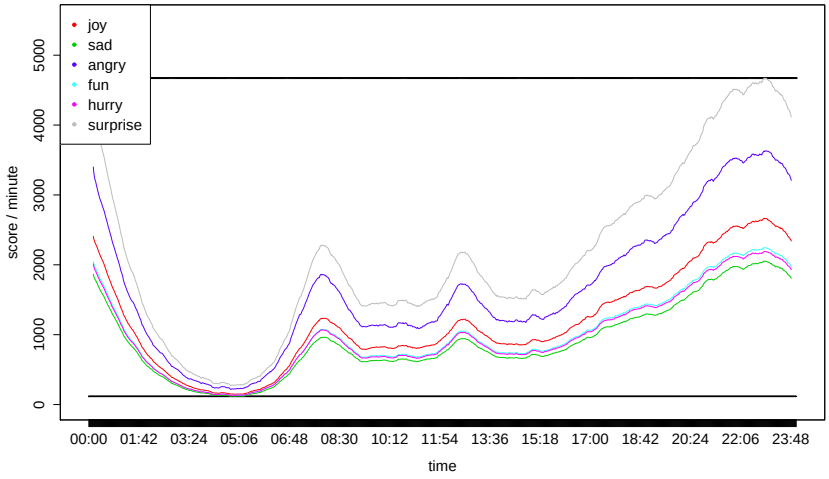


図 6.8: 平日の感情の推移

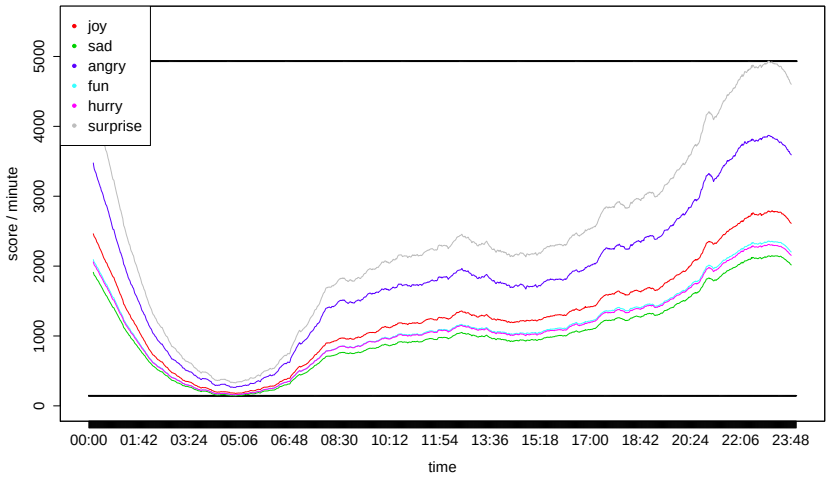


図 6.9: 休日の感情の推移

なった。なお、地震発生日はYahoo!Japanが提供する地震速報[1]を参考にし、東北地方でマグニチュード5.0、最大震度4以上の地震が発生した2012年10月3日、25日、11月3日、9日、24日、12月7日とした。図6.10にクリスマスの感情推移を、図6.11に地震発生日の感情推移を示す。

クリスマスは地震発生日だけでなく平日・休日と比較しても、感情量の総量が増加している。クリスマス以外では、23時頃まで感情量の増加傾向がみられ、その後減少傾向に変化するのに対し、クリスマスではこの変化は見分けられない。

地震発生日については、17時から18時にかけて感情量が急増している。地震速報によると、12月7日17時29分に三陸沖にて最大震度5弱、マグニチュード7.3の地震が発生し、その後20分の間にマグニチュード5以上の地震が二度発生している。大規模な地震のような突発的な出来事に対しては、情報の不足から事実や推測が普段より多く流通すると考えられ、実際に2011年に発生した東日本大震災時にも地震発生前後でツイート数に大きな違いがあることが確認されている[榊12]。官公庁(首相官邸[2]、防衛省[3]など)においては災害関連の情報提供を目的としてTwitterアカウントを開設しており、これらから発信される公的で信頼度の高い情報が多くリツイートなどを通じて発信されると考えられる。分析をおこなった日においては、これらの情報の増加と共に、顔文字を含んだツイートの件数も増加していると考えられる

6.4 結果

Streaming APIにより取得したデータを用い、5章で提案した多次元モデルにより、SNS感情の推移の定量化を試みた。データの要約により、全ツイート中に顔文字は常に21%程度の割合でみられることが明らかになった。平日と土日祝日の比較から、土日祝日の方が表現される感情量は多いことが明らかになった。昼の時間帯の感情量の比較から、この違いは投稿するユーザーの多さによるものだと考えられる。クリスマスを対象におこなった分析では、感情量が他の日に比べて多いことが確認された。また、地震発生日を対象におこなった分析では、地震発生後に感情量の急増がみられ、多くの人が地震発生について各自の感情を示しているして、SNS上の感情の様相がみられた。

しかし前節5.2で述べたように、高頻度で出現する顔文字が解析対象から外れていることや、本章と章4の結果を比較すると、分析に使用している感情ラベルの種類が違うものの、結果に大きな差異がみられる。前節で提案した多次元モデルによる感情表現の推定スコアは正しいと仮定したが、主観による顔文字分類にも一定の評価を与えるならば、両者の分析結果が異なる点に提案手法の改良の余地がみられる。

[1] http://typhoon.yahoo.co.jp/weather/jp/earthquake/list/
[2] https://twitter.com/Kantei_Saigai
[3] https://twitter.com/bouei_saigai

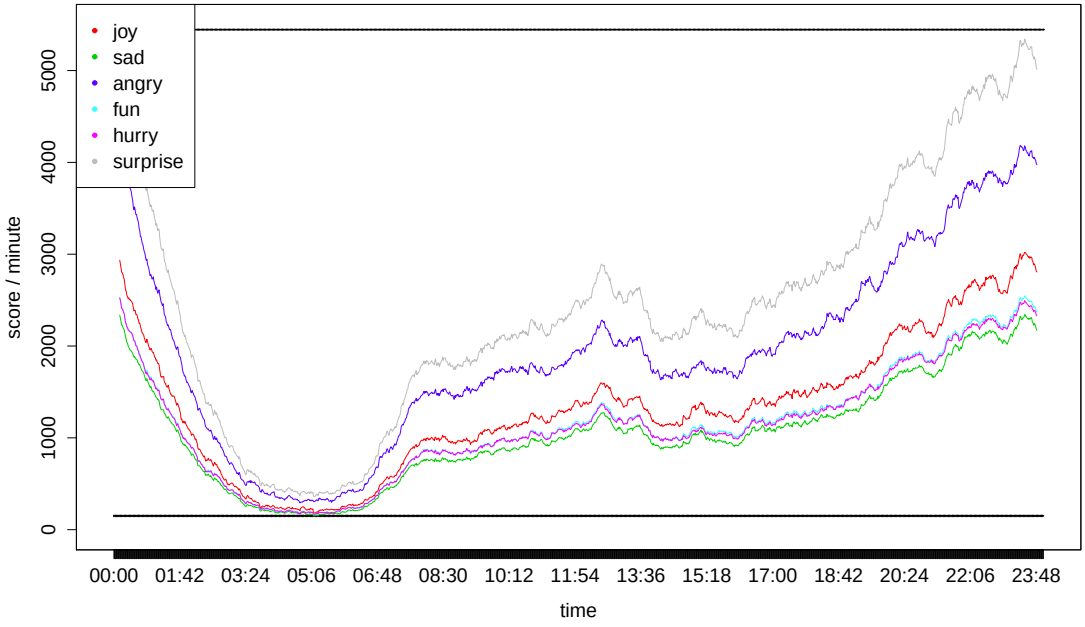


図 6.10: クリスマスの感情推移 (ポジティブな外的要因の例)

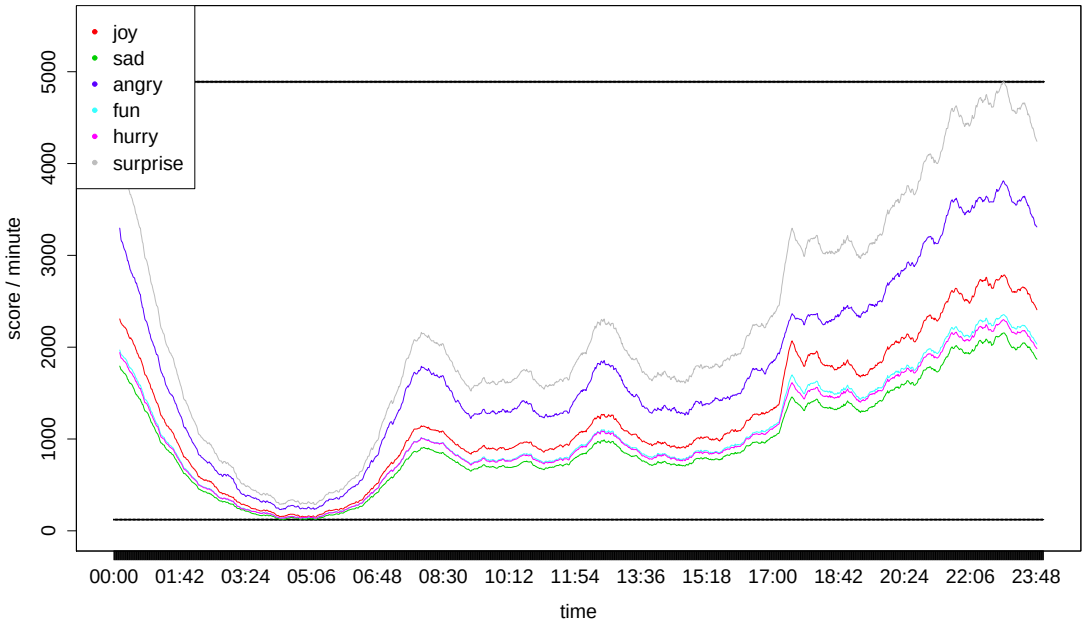


図 6.11: 地震発生日の感情推移 (ネガティブな外的要因の例)

6.5 課題

SNS の感情と社会トレンドとの関係を明らかにするために分析をおこなったが、分析がツイート数のみによっている点に課題がある。単純なツイート数のみではなく、ユーザー数やユーザーの分類、ツイートの拡散という観点も合わせた複合的な解析が必要と考えられる。

また、地震に着目した解析においては地理情報を考慮に入れていない点にも課題が残る。地理情報を考慮することにより、震源地付近でのツイートと震源地から離れた地域でのツイートを比較することでより詳細な地震の影響を測ることができると考える。地震発生時には自身が感知した揺れの程度を報告するユーザーもおり、地震発生時刻のデータとこの情報、ユーザーが地震発生時に居た地理情報を収集し、分析することで、災害マップや防災に大きな貢献をすることができると考える。

第7章　結論・今後の展望

7.1　結論

本研究では、顔文字による感情解析手法の確立、SNS 感情と社会トレンドの関係について分析した。

テキストに依らない感情表現として顔文字に着目した感情推定手法を作成した。顔文字を目、口といった形態素に分割し、形態素から感情を推定する手法として、単一次元モデル、多次元モデルを作成し決定木分析によりそれぞれのモデルを用いて形態素から感情を推定した。その結果、単一次元モデルでは十分な精度が得られず、顔文字のみを用いて感情を推定することは困難であると分かった。

多次元モデルにおいては、アンケート手法により作成された顔文字とそれが表す感情についてのデータベースを訓練データとし、決定木分析により感情を推定した。アンケートによるスコアと推定されたスコアの相関係数、検定力分析により手法の評価をおこなった。その結果、”喜び”、”哀しみ”、”楽しさ”、”焦り”、”驚き”の感情について有意に強い正の相関関係があると認められ、検定力も十分であった。しかし訓練データとして使用した顔文字の種類が少なく、実際のツイートデータから取得した出現頻度上位 30 個の顔文字中 15 個について感情を表現することが出来なかった。

Twitter Streaming API により収集したツイートデータを使用し、多次元モデルを用いて顔文字の感情を表現し、単位時間あたりの集計を感情量とみなして社会トレンドとの関係について分析をおこなった。その結果、顔文字を含むツイートが全ツイートに占める割合は 21%程度で安定しており、顔文字を対象とした分析でも安定したデータ量が確保できることが明らかになった。多次元モデルにより顔文字の感情スコアを算出し、時系列方向の視点から分析をおこなった。その結果、平日と土日祝日の比較では表現される感情量の推移に違いがあることが明らかになった。また、クリスマスとクリスマス以外の日 (非イベント時期) の比較では、クリスマスについては非イベント時期よりも表現される感情量の最大値が大きいことがみられ、また 23 時以降における感情量の減少傾向が明らかでないことから、非イベント時期とは異なる傾向を示していた。地震発生日では、地震発生直後に感情量の急激な増加が確認さ、社会で発生した突発的な出来事による影響が SNS 上で確認された。

顔文字を利用した感情推定手法には改良の余地が大きくみられるが、平日と土日祝日の比較、クリスマスと非イベント時期の比較、地震発生日においてそれぞれの特徴を見ることができた。そのため、SNS 感情と社会トレンドとの解析においては有用な比較指標と

なりうることが確認された。

7.2 展望

本研究では顔文字について形態素解析をおこない、形態素を用いた決定木分析により顔文字の感情スコアを推定する手法を提案した。しかし実際に Twitter 上で使用されている顔文字においては、解析の対象とならない顔文字が多い点に課題がある。

顔文字の種類は現在なお増加傾向にあり、日本語や英語以外の言語を使用した顔文字も出現していることから、今後十分な数の訓練データを収集することには限界がある。そのため、教師なしデータを用いて予測精度・分類精度を向上させる半教師あり学習の手法を適用し、増え続ける顔文字の感情を解析できる頑強なシステムの作成に取り組みたい。

また顔文字は記号列であり、その並び方により感情が表現されるのではない。目がうれしそうだ、口が悲しんでいそうだ、といった様態を判断して、人が感情を推定している。このような様態は人の主観により判断されるため、文字を記号として扱うテキスト解析では、様態をコンピュータで解釈するのは困難である。そこで画像処理の手法を適用することにより、様態をコンピュータで解釈することが可能となれば、訓練データの収集困難性や未知語への対応にも大きな貢献がされるものと考える。

参考文献

[A80] Russell James A. A circumplex model of affect. *Journal of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161–1178, Dec 1980.

[AH10] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pp. 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[BMZ10] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2010.

[DBvG07] Daantje Derks, Arjan E.R. Bos, and Jasper von Grumbkow. Emoticons and social interaction on the internet: the importance of social context. *Computers in Human Behavior*, Vol. 23, No. 1, pp. 842 – 849, 2007.

[Gil12] Jim Giles. Computational social science: Making the links. *nature International weekly Journal of science*, Vol. 488, No. 488, pp. 448–450, 2012.

[Man08] Jack M. Maness. A linguistic analysis of chat reference conversations with 18-24 year-old college students. *The Journal of Academic Librarianship*, Vol. 34, pp. 31–38, 2008. 1.

[PDRA10] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, and Kenji Araki. Towards fully automatic emoticon analysis system (^o^). *In Proceedings of The Fifteenth Annual Meeting of The Association for Natural Language Processing (NLP-2010)*, pp. pp. 583--586, 2010.

[PDS+09] Michal PTASZYNSKI, Pawel DYBALA, Wenhan SHI, Rafal RZEPKA, and Kenji ARAKI. A system for affect analysis of utterances in japanese supported with web mining. 知能と情報, Vol. 21, No. 2, pp. 194--213, 2009.

[PMD+10] M. Ptaszynski, J. Maciejewski, P. Dybala, R. Rzepka, and K. Araki. Cao: A fully automatic emoticon analysis system based

on theory of kinesics. *Affective Computing, IEEE Transactions on*, Vol. 1, No. 1, pp. 46 --59, jan. 2010.

[TTO05] Yuki Tanaka, Hiroya Takamura, and Manabu Okumura. Extraction and classification of facemarks with kernel methods. *International Conference on Intelligent User Interfaces*, 2005.

[井上 97] 井上みづほ，藤巻美菜子，石崎俊．電子メール文における感情表現の解析システムについて ： 感情表現の収集・分類・解析．電子情報通信学会技術研究報告．TL，思考と言語，Vol．96，No．608，pp．1--8，mar 1997．

[加藤 05] 加藤由樹，杉村和枝，赤堀侃司．電子メールを使ったコミュニケーションにおいて生じる感情への電子メールの内容の影響．日本教育工学会，Vol．29，pp．93--105，2005．

[加藤 08] 加藤由樹，加藤尚吾，杉村和枝，赤堀侃司．テキストコミュニケーションにおける受信者の感情面に及ぼす感情特性の影響-電子メールを用いた実験による検討-．日本教育工学会，Vol．31，pp．403--4145，2008．

[荒川 04] 荒川歩，鈴木直人．謝罪文に付与された顔文字が受け手の感情に与える効果．対人社会心理学研究，Vol．4，pp．128--133，2004．

[荒牧 12] 荒牧英治，増川佐智子，森田瑞樹．文章分類と疾患モデルの融合によるソーシャルメディアからの感染症把握．言語処理学会誌，Vol．Vol.19 No.5，pp．pp.419--435，2012．

[榊 12] 榊剛史，丸井淳己，松尾豊，鳥海不二夫，篠田孝祐，風間一洋，栗原聡，野田五十樹．大規模災害時におけるソーシャルメディアの変化．言語処理学会第 18 回年次大会，2012．

[川上 08] 川上正浩． 顔文字が表す感情と強調に関するデータベース． *The Human Science Research Bulletin*, Vol．No.7，pp．67--82，2008．

[中村 93] 中村明．感情表現辞典．東京堂出版，1993．

[登美 04] 登美原田．「顔文字」による日本語の円滑なコミュニケーション ： 「配慮」と「ポライトネス」の表現機能．言語と文化，Vol．8，pp．205--224，mar 2004．

[福岡 03] 福岡義隆．気象・季節の感情障害への影響．国際環境研究協会，2003．

発表論文

データマイニングを用いた顔文字表現の定量的評価による感情解析 山口 和宏, 杉山 歩, 鈴木 健之, 藤田 哲也, Ho Bao Tu, Dam Hieu Chi 言語処理学会第 18 回年次大会 2012 年

SNS 上に表れる個人感情を用いた社会トレンドについての研究 山口 和宏, 杉山 歩, Ho Bao Tu, Dam Hieu Chi 言語処理学会第 19 回年次大会 2013 年

謝辞

本研究を進めるにあたり、要所で的確なご助言を頂いた北陸先端科学技術大学院大学 知識科学研究科 Dam Hieu Chi 准教授に感謝いたします。普段の研究では示唆に富んだご指導・ご助言により、熱意を失うことなく、より多角的な視点を意識して活動できました。分析手法の議論の際には、私の理解不足から招いた間違いをご指摘いただき、丁寧な説明により軌道修正することが出来ました。特に発表用資料作成時には内容へのご指摘のみならず、聴衆にとってより魅力的な発表にするために多大なご助力をいただきましたことを感謝いたします。

日頃の研究の進め方に対する丁寧で分かりやすいアドバイスを頂いた北陸先端科学技術大学院大学 知識科学研究科 杉山 歩 助教授に感謝いたします。研究活動のみならず、就職活動においても様々なご助言をいただきました。本論文をまとめることが出来ましたのも、論文の推敲についての多くのご助言と、また目先の事物に集中しがちであった私に全体を俯瞰した上でのご指摘を頂いたおかげと感謝いたします。

研究をおこなうに当たり、心構えや考え方などの基礎的な部分をご指導いただいた北陸先端科学技術大学院大学 マテリアルサイエンス研究科 水上 卓 助教授に感謝いたします。中間審査や修士論文執筆などの繁忙期には、研究室が異なるのにも関わらず多大なお気遣いとご助力をいただきました。

本論文を執筆するにあたり、屈託のない意見を頂いただけでなく、顔文字の解析結果の精度確認で地道な作業にご助力いただいた研究室の皆様に感謝いたします。

分析や論文執筆が思うように進まず根を詰めていたときに、何気ない談笑や食事に連れ出し、または (半ば強引に) テニスへ誘ってくれ、気分転換の機会を頂いた友人諸氏、研究室の皆様、テニスサークルの皆様に感謝いたします。

最後に、大学院での研究生活を支援してくれた家族に感謝します。

付 録A　データベース概要

PostgreSQL 9.1 を使用し、データベースを作成した。この概要を図 A.1 に示す。各カラム名は付録 B にある項目および公式サイト[1]を参照されたい。CONVERTED_TEXT カラムについては、ツイート本文について全角・半角記号の統一をしたデータを保存するカラムである。なお、薄紫の背景はスキーマを、赤字のカラムは主キーを、緑字のカラムは外部キーを示す。

[1]https://dev.twitter.com/docs/platform-objects/tweets

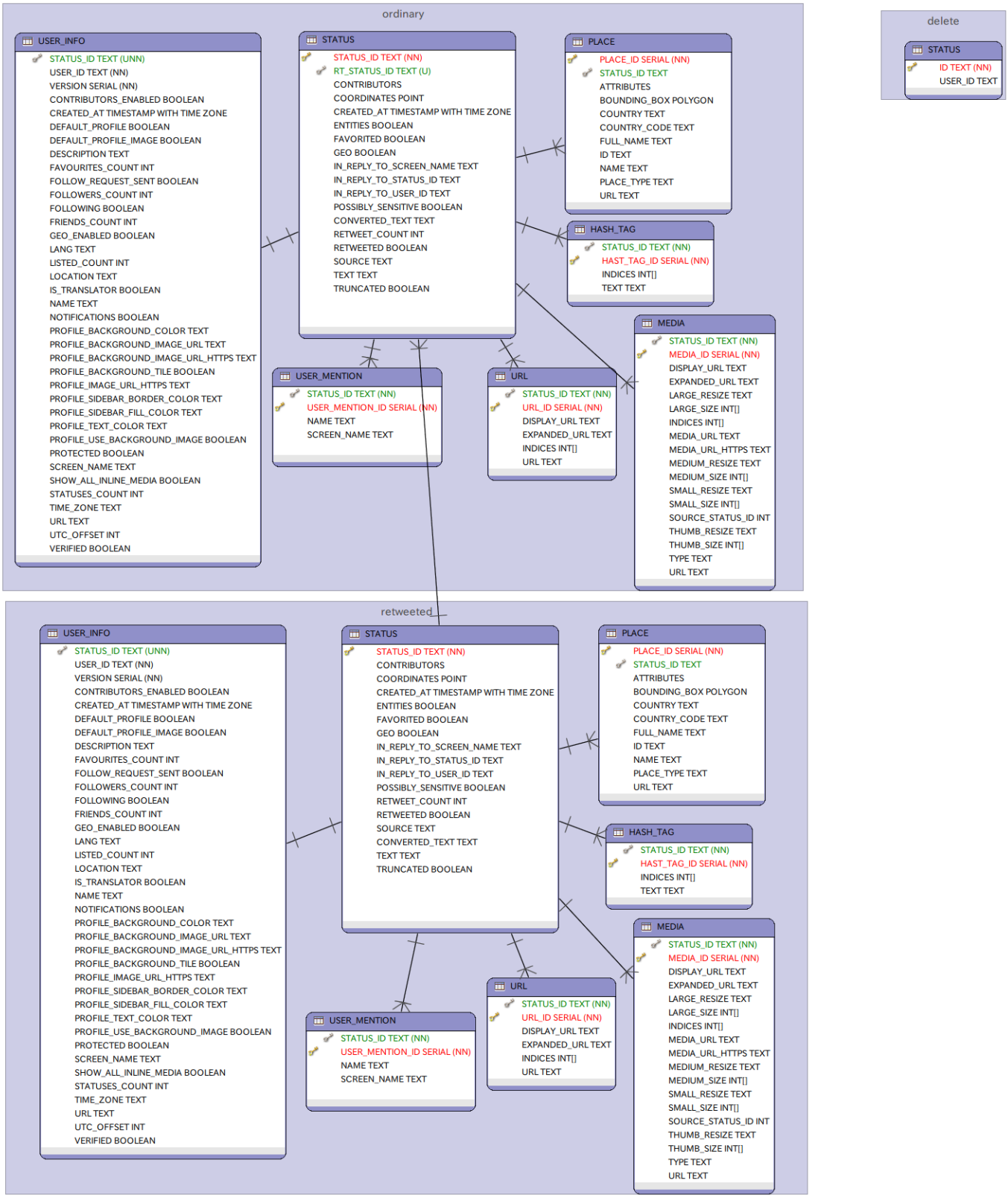


図 A.1: データベース概要図

表 B.1: ツイート情報

field	description
contributors	信頼されたユーザー
coorinates	地理情報
created_at	ツイート発信時刻 (世界標準時)
current_user_retweet	リツイートしたユーザに関する情報
entities	ハッシュタグ、URL、メンションなどに関する情報
favorited	お気に入り登録
id	ツイートを一意に示す id
in_reply_to_screen_name	リプライしたユーザーのスクリーンネーム
in_reply_to_status_id	リプライツイートの id
in_reply_to_user_id	リプライしたユーザーの id
place	地理情報
possibly_sensitive	リンクフラグ
retweet_count	リツイートされた回数
retweeted	リツイートフラグ
source	投稿元クライアント
text	ツイート本文
truncated	切り詰めフラグ
user	ツイートを発信したユーザー情報

付 録B　Twitterデータ概要

取得可能な属性一覧を表 B.1,B.2,B.3,B.4 に示す。各属性の詳細は公式サイトを参照されたい。

表 B.2: ユーザー情報

field	description
contributors_enabled	Contributor モードフラグ
created_at	ユーザー登録した時刻 (世界標準時)
default_profile	デフォルトのテーマ使用フラグ
default_profile_image	デフォルトのユーザーアイコン使用フラグ
description	アカウントの説明
favourites_count	お気に入り数
follow_request_sent	フォローリクエスト送付フラグ
following	フォローフラグ
followers_count	フォロワー数
friends_count	フォロー数
geo_enabled	地理情報使用許可
id	ユーザーを一意に示す id
lang	使用言語
listed_count	リストに追加された回数
location	地理情報
name	アカウント名
notifications	通知フラグ
profile_background_color	プロフィールの背景色
profile_background_image_url	プロフィールの背景画像 URL
profile_background_tile	プロフィールの背景繰り返しフラグ
profile_image_url	プロフィール画像 URL
profile_link_color	プロフィールリンク色
profile_sidebar_border_color	サイドバー色
profile_sidebar_fill_color	サイドバー背景色
profile_text_color	テキスト色
profile_use_background_image	背景画像使用フラグ
protected	保護フラグ
screen_name	ユーザー名
show_all_inline_media	ツイートに付加されたメディア情報表示フラグ
status_count	ツイート数
time_zone	タイムゾーン
url	URL
utc_offset	世界標準時との差

表 B.3: エンティティ情報

field	description
hashtags	ハッシュタグ
media	画像
urls	URL
user_mentions	メンション

表 B.4: 場所情報

field	description
attributes	場所
bounding_box	4 点の緯度経度で表現される位置情報
country	国
country_code	国の識別コード
full_name	場所の正式名称

付 録C　分析結果

C.1　顔文字の感情分類

主観による感情割り当ての結果を表 C.1 に示す。

C.1.1　利用クライアント一覧

投稿クライアント一覧を表 C.2 に示す。ただし、一日あたり投稿数が 1 回に満たないクライアントは省略した。

C.2　決定木による感情推定

C.2.1　単一次元モデル

章 5.1 により作成した決定木の分類規則を図 C.1 に示す。出力結果の見方は、例えば”6) 目 0.5 140 69 怒る (0.51 0.064 0.16 0.064 0.2)”であれば、第 6 ノードに該当するデータは 140 件あり、分類ラベルは”怒る”である。そのうち誤分類のデータは 69 件ある。(0.51 0.064 0.16 0.064 0.2) は、140 件中に占める怒る、泣く、笑う、照れる、驚く、の各教師データラベルの割合である。つまり 140 件中 51%は怒る、16%は笑うのラベルを持つデータである、と解釈できる。末尾に”*”が付与されている場合は、そのノードがターミナルノードであることを示す。

C.2.2　決定木による多次元モデルによる感情推定

分岐数と複雑度の関係

章 5.2 において決定木を作成するに当たり複雑度を算出した。この詳細を表 C.3 に示す。

分類規則

章 5.2 により作成した決定木の分類規則を表 C.2 に示す。

表 C.1: 顔文字の分類

	喜		怒	哀
	(*^^*)	(●>U<●)	(‘_ ′)	(・ω・‘=)
	(^o^)	(。・ω・。)	(‘д ′)	(′-ω-‘;)
	(^-^)	(≧∇≦)	(¯ ^ ¯)	(′・ω ゞ)
	(^_^)	(^ω^)	Σ(° д°)	(′・ω・ゞ)
	(^^)	(〃 ′・ω・‘)	(° д°)	(T_T)
	(*‘ω ′)	(*・ω・)	(;° д°)	(′・ω・‘)
	(‘ω ′)	(* ′∀ ‘ゞ)	(‘・ω ′・c)	(; ′д ‘)
	(‘・ω・′)	(’ω’)	(つ‘ω ′♂c)	(′; ω;‘)
	(‘-ω- ′)	(≧ε≦●)	(・д ′・+)	(′・_・‘)
	(*‘ω ′*)	(^д^)	(‘o ω n ′)	(′ε ‘)
	(*‘・ω・)	(。・ω・)		(^_^;)
	(*‘ω ′)	(#⌒∇⌒#)		(′ε ‘●)
	(¯ ∇ ¯ *)	(* ′∀ ‘*)		(′д ‘)
	(¯ ∇ ¯)	(・∀・)		(>_<)
	(¯ ∀ ¯)	(^∇^)		(′ω ‘)
	(¯ ∀ ¯ *)	(* ′ω ‘*)		(・ω・ゞ)
	(・ω・)	(‘・ω・′)		(● ′ε ‘●)
	(′∇ ‘)	(・ω・)		(¯ ∇ ¯ ;)
	(′∇ ‘*)	(*‘ω ′)		(′ë ‘)
	(● ′ω ‘●)	(¯ ∇ ¯ *)		(ノд ‘)
	(* ′∇ ‘*)	(‘・ω・)		(* ′ω ‘)
	(′∇ ‘)	(‘ω ′)		(* ′д ‘*)
	(・ω・*)			(*>д<*)
合計	45		10	24

表 C.2: クライアント一覧

client name	tweet per day	client name	tweet per day
Twitter for Android	2384	Twitpic	4
Keitai Web	1569	Colotwi	4
web	1445	berSocial for BlackBerry	4
Twitter for iPhone	1442	メルマガ執筆中！	4
twicca	1355	ガチャツイ	4
ついっぷる/twipple	1129	fuyutiger	3
Twipple for Android	626	TweetParakeet	3
ついっぷる for iPhone	522	SimplyTweet	3
SOICHA	462	NHK	3
Echofon	364	Facebook	3
jigtwi	297	サンシャイン栄 SKE48 劇場	3
モバツイ / www.movatwi.jp	382	ふぐりくん bot	3
TweetDeck	273	そぐねふぃよるど	3
Tween	254	Twidroyd for Android	3
Janetter	254	Realtime love	3
yubitter	170	pochitter	2
Mobile Web	108	okinawa_jp_net	2
Saezuri	101	albo	2
Tweetbot for iPhone	98	TwitPal	2
HootSuite	92	Rewit	2
TwitBird	85	Hatena	2
ついっぷる Pro for iPhone	81	ERIWACS	2
YoruFukurou	78	棒読みちゃん	2
Tweet Button	72	森林	2
twittbot.net	66	ニコニコ生放送	2
ツイタマ	60	Photos on iOS	2
TweetCaster for Android	58	Janetter for Mac	2
Seesmic	57	Get! PocketVegas	2
Tweetlogix	50	FC2 Blog Notify	2
Twittelator	42	yoono	1
mixi ボイス	40	twksr	1
ニコニコ動画	38	twicli	1
Teewee	36	tweetz	1
TweetList!	35	tGadget	1
twitbeam[]	34	ra	1
Twil2 (Tweet Anytime Anywhere by Mail)	32	hamooooooon	1
Twipple Pro for Android	31	V2C	1
Tabtter	29	Twitterrific	1
ついっぷる for iPad	28	TkMixiViewer	1
Tweet ATOK	25	T3:TonyTonyTweet	1
TweetList Pro	21	SoraUsagi	1
twitterfeed	20	REALWORLD	1
Plume	18	Naonatter	1
Movatter	17	MySweetBot	1
Twitter for iPad	17	Minkara	1
Osfoora for iPhone	16	Google	1
モバツイ	15	Frother	1
Twitter for Mac	13	DECOLOG	1
ついっぷる Pro for iPad	11	Azurea	1
Twit for Windows	11	忘却の城 地上の 12 階	1
Ustream.TV	10	ロケタッチ（loctouch）	1
TwitCasting	9	ユニリーバ ツイッター募金	1
P3:PeraPeraPrv	9	メガネケエス	1
Krile2	9	デスティニーアイランドの浜辺	1
Easybotter	9	みについ	1
モバツイ smart / www.movatwi.jp	9	にせほ脳	1
Silver Bird	9	とある美琴の目覚時計	1
dlvr.it	8	たぁbot / tacha_bot	1
TweetMe for iPhone	7	たぁbot	1
Feel on! iPhone 版	7	えるえーにたんしんふにんなう。	1
NatsuLiphone	6	あーりんの部屋	1
Crowy	6	twtkr for iPhone	1
Twitter for BlackBerry@	6	Twitter for iAppli	1
instagram	5	TwitBird iPad	1
guppi	5	TweetMe for S!アプリ	1
TwitShepherd	5	Sonic ★ Tweet	1
モバツイ touch	5	Seesmic twhirl	1
au one Friends Note	5	Samsung Mobile	1
UberSocial for Android	5	SKE48 オフィシャルブログ	1
Jibe Mobile App	5	New さたん bot	1
Janetter Classic	5	MultiTweet for iPhone	1
yabmin	4	Feel on! Android 版	1
twippa	4	Camera on iOS	1
mixvTwee	4	CROOZ blog	1
foursquare	4	2ch まとめサイトリーダー	1
HTC Peep	1		

図 C.1: 決定木の分類規則

```
n= 255

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 255 176 怒る (0.31 0.15 0.31 0.086 0.15)
   2) 目< -6.5 28   2 泣く (0.036 0.93 0.036 0 0) *
   3) 目>=-6.5 227 149 怒る (0.34 0.048 0.34 0.097 0.17)
     6) 目< 0.5 140  69 怒る (0.51 0.064 0.16 0.064 0.2)
       12) 手=fxxk 26   0 怒る (1 0 0 0 0) *
       13) 手=0,arm,hand,peace 114  69 怒る (0.39 0.079 0.2 0.079 0.25)
         26) 目< -1.5 26   3 怒る (0.88 0 0.12 0 0)
           52) 手=0,arm,hand 24   1 怒る (0.96 0 0.042 0 0) *
           53) 手=peace 2   0 笑う (0 0 1 0 0) *
         27) 目>=-1.5 88  60 驚く (0.25 0.1 0.23 0.1 0.32)
           54) イメージ=phew,sigma,sigma and sweat,
                       sweat,xmark 27  14 怒る (0.48 0.037 0 0.037 0.44)
             108) イメージ=sigma and sweat,xmark 6   0 怒る (1 0 0 0 0) *
             109) イメージ=phew,sigma,sweat
                          21   9 驚く (0.33 0.048 0 0.048 0.57) *
           55) イメージ=0 61  41 笑う (0.15 0.13 0.33 0.13 0.26)
             110) 頬>=0.5 15   8 照れる (0.13 0.067 0.33 0.47 0)
               220) 口>=0.5 3   0 笑う (0 0 1 0 0) *
               221) 口< 0.5 12   5 照れる (0.17 0.083 0.17 0.58 0) *
             111) 頬< 0.5 46  30 驚く (0.15 0.15 0.33 0.022 0.35)
               222) 手=0,arm 37  22 驚く (0.19 0.19 0.22 0 0.41)
                 444) 目< -0.5 11   6 笑う (0 0.36 0.45 0 0.18)
                   888) 口< -0.5 4   1 泣く (0 0.75 0 0 0.25) *
                   889) 口>=-0.5 7   2 笑う (0 0.14 0.71 0 0.14) *
                 445) 目>=-0.5 26  13 驚く (0.27 0.12 0.12 0 0.5)
                   890) 口>=1.5 2   0 怒る (1 0 0 0 0) *
                   891) 口< 1.5 24  11 驚く (0.21 0.12 0.12 0 0.54) *
               223) 手=hand,peace 9   2 笑う (0 0 0.78 0.11 0.11) *
     7) 目>=0.5 87  33 笑う (0.08 0.023 0.62 0.15 0.13)
       14) 目< 3.5 73  21 笑う (0.068 0.027 0.71 0.16 0.027) *
       15) 目>=3.5 14   5 驚く (0.14 0 0.14 0.071 0.64)
         30) 手=fxxk,peace 3   1 笑う (0.33 0 0.67 0 0) *
         31) 手=0,arm 11   2 驚く (0.091 0 0 0.091 0.82) *
```

図 C.2: 決定木の分類規則

```
n= 28

node), split, n, deviance, yval
      * denotes terminal node

    1) root 28 222.3500 2.2798
      2) right_eye=,-,;,T,^^ef^^be^^9f, ¯ 15  80.9860 2.2495
        4) mouth=,_,o, ω 8  25.9470 2.1109 *
        5) mouth=□, д 7  22.2380 2.4079 *
      3) right_eye=0,<,^,‘,^^ef^^bd^^a5, ≦ 13  67.2350 2.3148
        6) mouth=,_, д 6  27.4320 2.2336 *
        7) mouth=-,o, ∀, ∇ 7   9.6139 2.3845 *
```

表 C.3: 分岐数と複雑度の関係

CP	nsplit	rel error	xerror	xstd
0.33339	0	1.0000	1.091	0.1031
0.14752	1	0.6666	1.371	0.1886
0.13577	2	0.5191	1.368	0.1833
0.06152	3	0.3833	1.188	0.2182
0.05835	4	0.3218	1.285	0.2418
0.02396	5	0.2634	1.224	0.2467
0.02068	6	0.2395	1.254	0.2536
0.01980	7	0.2188	1.252	0.2536
0.01138	8	0.1990	1.352	0.2633
0.01000	9	0.1876	1.360	0.2663