

マイクロブログの極性の差に着目した期待を裏切るスポットの発見

豊島 美穂[†] 加藤 大受^{††,†††} 遠藤 雅樹^{††††} 莊司 慶行[†] 廣田 雅春^{†††††}

石川 博^{††}

[†] 首都大学東京 システムデザイン学部 〒191-0065 東京都日野市旭が丘 6-6

^{††} 首都大学東京大学院 システムデザイン研究科 〒191-0065 東京都日野市旭が丘 6-6

^{†††} ウイングアーク1st株式会社 〒150-0031 東京都渋谷区桜丘町 20-1 渋谷インフォスター

^{††††} 職業能力開発総合大学校 基盤ものづくり系 〒187-0035 東京都小平市小川西町 2-32-1

^{†††††} 大分工業高等専門学校 情報工学科 〒870-0152 大分県大分市大字牧 1666

E-mail: [†]toyoshima-miho@ed.tmu.ac.jp, ^{††}kato.d@wingarc.com, ^{†††}tendou@uitech.ac.jp, ^{††††}y_shoji@tmu.ac.jp,

^{†††††}m-hirota@oita-ct.ac.jp, ^{††}ishikawa-hiroshi@tmu.ac.jp,

あらまし 本論文では、Twitter に投稿されたテキスト情報、位置情報、および時間情報を利用して「多くの人が良くも悪くも期待を裏切られたと感じるスポット」を発見する手法を提案する。そのスポットを発見するために「ユーザのあるスポットを訪れる前と後の Tweet の感情スコアの差は、そのユーザの期待と実際の感想のギャップを表す」という仮説を立て、モデルを構築した。東京都全体と一部の周辺地域を対象とし、多くのユーザが期待よりも悪かったと感じたスポットと、期待よりも良かったと感じたスポットをランキング化した。被験者が作成したランキングと比較することで、評価を行なった。評価実験の結果、ある地点での Tweet は直後の Tweet の感情スコアに影響を及ぼすという仮説のもと作成されたランキングと被験者が作成したランキングに弱い相関が見られた。

キーワード 感情分析, Twitter, 位置情報, 観光プロモーション

1. はじめに

世の中には数多くの観光スポットが存在するが、その中にはがっかりな観光スポットと揶揄される場所も多い。札幌の時計台や、シンガポールのマラーイオンなどはその代表格と言える。また、一般的にがっかりのイメージの少ないローマのトレヴィの泉や、京都の鹿苑寺金閣に対してがっかりしたという声も少なくない^(注1)。また、観光客をがっかりさせるこれらの観光スポットとは逆に、訪れた人の多くが訪れる前の予想よりも魅力的だと感じるスポットも存在していると考えられる。そのようなスポットは、まだ観光スポットとして確立していないスポットとされ、将来的には観光スポットになる可能性がある。

本研究では「訪れた多くの人が良くも悪くも期待を裏切られたと感じること」をギャップと呼称し、そのようなギャップを感じるスポットをギャップスポットと定義する。観光スポットの著名・無名に関わらずギャップスポットを発見することを最終目標とする。ここで、「期待を裏切られた」とは、訪問する前に予想していたものと実際に訪れた際の感想が大きく異なった場合である。期待が大きい状態であるスポットに訪れ、実際の感想がその期待を大きく下回った時、そのスポットは期待を裏切るスポット（以下、がっかりスポット）となりうる。逆に、期待が小さい状態であるスポットを訪れ、その期待を実際の感想が大きく上回った時、そのスポットは新しい観光スポット（以

下、掘り出し物スポット）となりうる可能性を秘めている。

がっかりスポットを発見することで、旅行者はあらかじめその情報を知り、観光スポットを選択する基準のひとつとすることができる。また、観光スポット側はがっかりを集客に繋げることもできる。一方、掘り出し物スポットを発見することで、新しい観光スポットとして集客につなげることができる。このように単純な知名度や人気によらない観点から新しい観光スポットを提示することで、観光産業の活性化が期待される。

近年、マイクロブログのひとつである Twitter^(注2) の投稿からユーザの感情を推定する研究 [1] [2] [3] が盛んに行なわれている。これは Twitter が広く普及しているツールであることと、手軽に今の気持ちや状態を投稿することができるためである。本研究では、観光中でも容易に投稿できる Twitter に対して感情分析を行うことで、スポットが極性値に与える影響を調査する。

本研究では、がっかりスポットと掘り出し物スポットを発見するため、Twitter のテキスト情報から感情分析を行う。あるスポットで投稿された時系列順に隣接する Tweet の感情極性値の変化を調査した。最終的にはギャップスポットの発見により、旅行者の観光スポットの選択の幅が広がることが期待される。

以下に本稿の構成を述べる。2 節でテキストの感情分析に関する研究と地域に対する感情分析の研究について述べる。3 節で提案手法と実験方法について述べ、それらの結果を 4 節に示

(注1)：しらべえ

<http://sirabee.com/2016/08/04/148565/>

(注2)：Twitter

<https://about.twitter.com/company>

す。5節で考察し、6節で本稿のまとめを述べる。

2. 関連研究

2.1 テキストの感情分析に関する研究

菅原ら[4]は日本語テキストから感情を抽出する手法を提案した。彼らの研究では、感情スコアの算出に必要な語を収集し、語句レベルで解析した後、文全体での感情スコアを算出した。Goら[1]はTwitter上の顔文字等のノイズを多く含むテキストを、自動的に感情分類する手法を提案した。複数の機械学習と複数の特徴分類器を様々な組み合わせで試した。結果として、機械学習は感情分析に効果的な方法であることを示した。Kouloumpisら[5]は、従来、ニュースなどの感情分析に多く利用されていた自動品詞分析や感情分析がTwitterのようなマイクロブログにも有効に利用できるかを調査した。村上ら[2]は、類似する感情が類似する時間に出現するという仮説から、災害時に対象地域の住民が感じる不安や要望をTwitterから抽出した。また、時間情報とランドマークによる位置情報を利用し、時間変化する不安度の可視化を行った。三好ら[6]は顔文字を構成する文字や記号に着目しPN分類を行う手法を提案した。

本研究は、極性値や記号のみに着目したのではなく、あるスポットにおいて時間に伴って変化する極性値の急低下と急上昇に着目することにより、場所と時間による人々への感情的な影響を分析し、新たな観点からの観光スポットへの発見に取り組む。

2.2 地域に対する感情分析の研究

地域ごとの感情を分析する研究として、Lermanら[3]は、Twitterのテキストデータを用いて、社会的つながりや住民の感情分析を行い、それらを従来の統計データと比較する研究を行った。結果として、抽出できた感情と社会性は地域毎に差が見られた。Mitchellら[7]は、位置情報が付与されたTweetを使用し、アメリカ全土の州でそれぞれの幸福度を推定した。これらを推定するにあたり、どのような言葉で差が出るのか、単語の使用頻度を調査した。また、これらを統計データと比較を行うことで様々な項目との相関をとった。小原ら[8]はTwitterに位置情報が付与されているTweetが圧倒的に少ないという問題点に対して、Tweet本文からユーザの居住地や位置情報を推定する方法を用いた。Twitterからの地域を連想させる言葉やパターンマッチングを行い、ユーザの居住地推定と地域の観光情報を抽出する手法を提案し、システムを構築した。観光スポットの分析について、奥村ら[9]らは口コミなどのない非施設の長所情報をブログから抽出し、それらの長所情報を分類する手法を提案した。杉本ら[10]は口コミに含まれる感情語を利用し、観光スポットの分類を検討した。村上ら[11]は国外のユーザのブログを対象に、日本が外国人の旅行者からどのようなイメージや印象を持たれているかを調査した。ブログに出現する単語の回数によって、地域別の観光地としての特色を大まかに抽出し、分析を行った。

本研究では、地域の分析の中でも観光に着目して、観光に伴う感情の変化に着目して、観光地を特徴づけることに取り組んでいる。

3. 提案手法

本章では、ギャップスポットを発見するための手法について述べる。マイクロブログのひとつであるTwitterを用い、以下の2つの仮説を立て、検証を行った。

- ある地点でのTweetと直前のTweetの感情スコアの差は期待とのギャップを示す（以下、期待仮説）
- ある地点でのTweetは直後のTweetの感情スコアに影響を及ぼす（以下、引きずり仮説）

はじめに、手法の概要について3.1節に示す。感情スコアの計算方法を3.2節に、データセットの作成方法を3.3節に、ランキング作成にあたっての処理に関して3.4節にそれぞれ示す。

3.1 手法全体の流れ

初めに、期待仮説の検証方法を述べる。まず、Tweet群から、直前のTweetより感情スコアの差が閾値以上のTweetを抽出する。その後、対象地域を任意の正方形のグリッドに分割し、抽出されたTweetがそれぞれのセルに属するかを判定する。それぞれのセルにおいて、抽出されたTweetのうち感情スコアが閾値上がったTweet、下がったTweetの割合を調査しそれぞれ割合の大きい順にランキングを作成する。

また、引きずり仮説の検証を行うために、取得するTweetを、閾値以上の感情スコアの差があったTweetの直前のTweetに変更して同様の実験を行なった。

3.2 Tweet本文に対する感情スコアの計算方法

本節では、Tweetに感情スコアを付与する方法について述べる。Tweetから得られる形態素ごとの感情スコアを合算することで、Tweetの感情スコアを求める。ここで、形態素の感情スコアとは、その語がどの程度ポジティブまたはネガティブな意味合いを持つのかを数値で表したものである。また、それぞれの数値は、単語と感情スコアの対応表の辞書から求める。

まず、Tweetのテキストを形態素解析する。得られた形態素の中で、辞書に含まれていた数を num_{dict} 、そのスコアを sc_{word} としたとき、Tweetの感情スコア sc_{text} を

$$sc_{text} = \frac{\sum_{i=0}^{num_{dict}} sc_{word,i}}{num_{dict}} \quad (1)$$

と定義する。この際、テキスト内の単語が全て辞書に含まれていなかった場合、感情スコアが計算できなかったTweetとし、以後の処理から除外する。

3.3 Tweetのペアの感情スコアの差の計算

本節では、手法に必要なデータセットの作成方法について説明する。初めに、各ユーザ毎のTweetを投稿時刻に基づいて昇順にソートする。ソートした各Tweet群から時間間隔が1日以内のTweetのペアを抽出する。これは、前後でのTweetが時間的に離れすぎていると、Tweetの感情スコアの差が期待と現実のギャップを表していないと考えられるためである。また、ペア同士が同一セル内にある場合のTweetのペアを除外した。これは同一セル内で投稿されたTweetでは、スポットに対する感情変化ではないTweetや、自宅や職場での連続したTweetが存在すると考えられ、それらをあらかじめ除外するためである。残ったデータを本研究で扱うデータセットとする。



図1 Tweet ペア群抽出の例

Tweet のペアは、時系列順に並べて、前の Tweet b 後の Tweet a の二つで構成されており、これを T_{pairs} と定義する。これら抽出の例を図1に示す。また、各 Tweet に付随する感情スコアの集合を SC_{pairs} で定義する。 T_{pairs} を式 (2) に示し、 SC_{pairs} を式 (3) にそれぞれ示す：

$$T_{pairs} = \{(b_0, a_0), (b_1, a_1), \dots, (b_n, a_n)\}, \quad (2)$$

$$SC_{pairs} = \{(sc(b_0), sc(a_0)), (sc(b_1), sc(a_1)), \dots, (sc(b_{n-1}), sc(a_{n-1})), (sc(b_n), sc(a_n))\}. \quad (3)$$

これら集合の中から、仮説を検証するための集合を抽出する。直前の Tweet より閾値以上の感情スコアが上がった Tweet の集合 T_a^{pos} を式 (4) に、直前の Tweet より閾値以上の感情スコアが下がった Tweet の集合 T_a^{neg} を式 (5) に、それぞれ定義する。また、閾値以上の感情スコアが上がった Tweet の直前の Tweet の集合 T_b^{pos} を式 (6) に、閾値以上の感情スコアが下がった Tweet の直前の Tweet の集合 T_b^{neg} を式 (7) に同様に定義する：

$$T_a^{pos} = \{a_n | sc(a_n) - sc(b_n) > threshold\}, \quad (4)$$

$$T_a^{neg} = \{a_n | sc(b_n) - sc(a_n) > threshold\}, \quad (5)$$

$$T_b^{pos} = \{b_n | sc(a_n) - sc(b_n) > threshold\}, \quad (6)$$

$$T_b^{neg} = \{b_n | sc(b_n) - sc(a_n) > threshold\}. \quad (7)$$

また、この時の閾値は Tweet 間の感情スコアの差の閾値を表し、閾値以上の感情スコアがあった時にそのスポットに対して期待とのギャップがあったとする。取得した Tweet の集合のイメージを図2に示す。ある一人のユーザが図2のように Tweet を投稿したとき、閾値以上に極性値が上がった Tweet のペア (b_n, a_n) の要素 b_n, a_n は、それぞれ式 (6) と式 (4) に示される T_b^{pos} と T_a^{pos} に属する。閾値以上極性値が下がった Tweet のペア (b_{n+1}, a_{n+1}) の要素 b_{n+1}, a_{n+1} は、それぞれ式 (7) と式 (5) に示される T_b^{neg} と T_a^{neg} に属する。

3.4 セル内で各 Tweet 集合の割合の比較とランク付け

本節では、抽出した Tweet 群を用いて「がっかりスポット順」と「掘り出し物スポット順」のランキングを作成する手順について述べる。提案手法を適用する地域を任意の大きさのグ

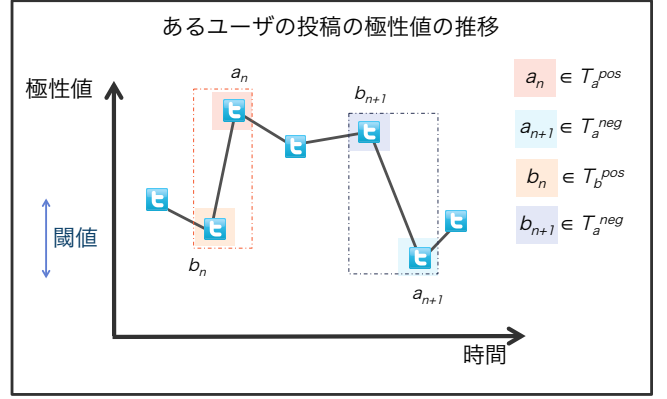


図2 手法で使用する抽出 Tweet のイメージ

リッドに区切り、そのセル内の T_a^{pos} と T_a^{neg} の割合、 T_b^{pos} と T_b^{neg} の割合をそれぞれ比較し、ランキング化する。この時、セル g_n 内で投稿された全 Tweet を $T(g_n)$ と定義する

以下に、期待仮説の検証方法について述べる。期待仮説は直前の Tweet との感情スコアの差異を利用するため、 T_a^{pos} と T_a^{neg} より計算する。セル g_n におけるプラス変化の度合い $SC_a^{pos}(g_n)$ は式 (8)、マイナス変化の度合い $SC_a^{neg}(g_n)$ は式 (9) にそれぞれ示される：

$$SC_a^{pos}(g_n) = \frac{|T_a^{pos} \cap T(g_n)|}{|T_a^{pos} \cap T(g_n)| + |T_a^{neg} \cap T(g_n)|}, \quad (8)$$

$$SC_a^{neg}(g_n) = 1 - SC_a^{pos}(g_n). \quad (9)$$

次に、引きずり仮説の検証方法について述べる。引きずり仮説は直後の Tweet との感情スコアの差異を利用するため、 T_b^{pos} と T_b^{neg} より計算する。期待仮説と同様にプラス変化の度合い $SC_b^{pos}(g_n)$ は式 (8)、マイナス変化の度合い $SC_b^{neg}(g_n)$ は式 (11) にそれぞれ示される：

$$SC_b^{pos}(g_n) = \frac{|T_b^{pos} \cap T(g_n)|}{|T_b^{pos} \cap T(g_n)| + |T_b^{neg} \cap T(g_n)|}, \quad (10)$$

$$SC_b^{neg}(g_n) = 1 - SC_b^{pos}(g_n). \quad (11)$$

式 (8)、式 (9)、式 (10)、式 (11) の値が大きい順に、それぞれランキング化する。

4. 実験結果と評価

本章では、期待仮説と引きずり仮説の2つの仮説を検証実験とその評価結果について述べる。また、提案手法で作成した各ランキングとベースラインを、アンケート結果との順位相関係数で比較した。このときのベースラインとは、グリッド内の感情スコアの平均値である。これらの比較を行うことで、時間的に変化する感情スコアの差は期待を裏切られたスポットを抽出する際に有効であるかを確認する。4.1 節に今回使用したデータセットについて示す。閾値の決定について4.2 節に示す。具体的な実験手順について4.3 節に、評価実験の手順について4.4 節にそれぞれ示す。また、実験結果を4.5 節に示す。

表 1 単語極性対応表に含まれる単語とスコアの例

単語	スコア	単語	スコア
優れる	1.00	悪い	-1.00
嬉しい	1.00	悲しい	-1.00
好評	1.00	悪評	-0.99
よろしく	0.68	油断	-0.60
健康	0.44	地滑り	-0.40
天使	0.37	消極	-0.35

4.1 データセット

本節では、今回の手法で使用したデータの詳細について述べる。はじめに、Twitter から取得した東京都で投稿された位置情報付き Tweet を 2014/6/11～2016/5/22 の期間で抽出した。この中から、「I'm at ×××」といった位置情報から現在地のみを自動的に投稿する Tweet や、「@×××」といったリプライに使用される表現を含んだ Tweet も今回の予想と実際の感情のギャップを表す実験には不適切であるとして、データセットから除外した。上記の表現を含まない Tweet の総数は 6,827,317 件で、Tweet のペアは 4,727,263 組であった。

これら Tweet の感情スコアを 3.2 節の手順を用いて計算したところ、ペアのどちらの Tweet でも感情スコアが計算できたペアは 4,220,283 組であった。本研究では、辞書に高村らによる単語感情極性対応表 [12] を使用した。この辞書において、単語がポジティブな意味合いをもつときスコアは 1 に近づき、ネガティブな意味合いをもつときスコアは -1 に近づく。この辞書には 55,125 個の動詞、名詞、形容詞が収録されている。対応表の一部を表 1 に示す。また、形態素解析には MeCab^(注3) を使用した。

4.2 閾値の決定

本節では感情スコアの変化の差の閾値について説明する。本手法では、期待と現実のギャップ差があるとみなす閾値を設定し、閾値以上の差がある時にギャップがあると見なした。今回のデータセットを使用し、感情変化の差を求めたところ、感情スコアの差の平均は約 0.14 であった。このことから、本実験では閾値を平均値以上の 0.2, 0.3 に定めそれぞれ実験を行なった。

4.3 実験手順

本節では、評価実験の具体的な手順について述べる。データセットの Tweet のペアから、感情スコアが閾値以上の Tweet をそれぞれ取得し、それらが T_a^{pos} , T_a^{neg} , T_b^{pos} , T_b^{neg} のどれに該当するかを調べた。データセットから集合を取得した。結果を表 2 に示す。次に、得られたそれぞれの Tweet 集合に含まれる Tweet の位置情報を調べ、東京都内のどのセルに属するかを調べた。この時、セルに属する Tweet が十分に少ない場合は分析が難しいと考え、Tweet が 50 件以下のセルを除外した。セルごとに $SC_a^{pos}(g_n)$, $SC_b^{pos}(g_n)$, $SC_a^{neg}(g_n)$, $SC_b^{neg}(g_n)$ を計算し、その値でランキングを作成した。

4.4 評価方法

本節では評価方法とその結果について示す。今回はフォート

表 2 閾値別の Tweet 取得件数

	T_a^{pos}	T_a^{neg}	T_b^{pos}	T_b^{neg}
0.2	154,886	152,005	154,781	151,904
0.3	87,419	85,478	87,303	85,578

ラベルの東京の観光スポットの人気ランキング TOP100^(注4) より、上位 50 件の著名な観光スポットを抽出し、被験者 15 人に対してアンケートをとった。アンケートでは被験者が訪れたことがある観光スポットに対し、以下の 4 段階の評価で、訪れる前のイメージと訪れたあとの感想をそれぞれスコア付けしてもらった。

- とてもつまらなさそう (つまらなかった)
- まあまあつまらなさそう (つまらなかった)
- まあまあ楽しそう (楽しかった)
- とても楽しそう (楽しかった)

観光スポット P に対し、 P に訪れたことがある n 人の内、被験者 i が訪れる前に持っていたイメージのスコアを $SC_{pre}(i)$ 、訪れた後の感想のスコアを $SC_{post}(i)$ とそれぞれ定義すると、観光スポットの P に対するイメージと感想のギャップ $G(P)$ は式 (12) で定義される：

$$G(P) = \frac{\sum_{i=0}^n SC_{post}(i) - SC_{pre}(i)}{n}. \quad (12)$$

式 (12) で得たスコアを元に、著名の観光地を降順にランキング化した。この時、訪れたことがあると回答した人数が 5 人以下だった観光スポットはあらかじめ除外した。このアンケート結果から作成したランキングと作成したベースライン、掘り出し物スポットランキング・がっかりスポットランキングとそれぞれスピアマンの順位相関係数を求めた。この時、スコアを降順にランキングしたため、掘り出し物スポットランキングで完全に相関がある時の順位相関係数は 1 になり、がっかりスポットランキングで完全に相関がある時の順位相関係数は -1 になる。

4.5 実験結果

本節では、得られたランキングと評価実験の結果について示す。

はじめに、期待仮説と引きずり仮説に基づいた 2 つのランキングを、掘り出し物スポットとがっかりスポットのそれぞれで作成した。表 3 は期待仮説に基づいて作成した掘り出し物スポットのランキングを、表 4 は引きずり仮説に基づいて作成した掘り出し物スポットのランキングである。それぞれの表において、閾値 0.2, 0.3 とした場合の結果を並べて示す。表中の PLACE は、Google Maps^(注5) から人手で作成し、セルの範囲に含まれる主要な建物やスポットの名前である。 $SC_a^{pos}(g_n)$, $SC_b^{pos}(g_n)$, $SC_a^{neg}(g_n)$, $SC_b^{neg}(g_n)$ はそれぞれ 3.4 節で示した式 (8), 式 (9), 式 (10), 式 (11) をそれぞれ計算した値である。

(注4)：フォートラベル 東京 観光 人気ランキング

<http://4travel.jp/domestic/area/kanto/tokyo/kankospot/>

(注5)：Google Maps

<https://www.google.co.jp/maps/>

(注3)：MeCab <http://taku910.github.io/mecab/>

表 3 期待仮説に基づいた掘り出し物スポットのランキング

閾値 0.2		閾値 0.3	
PLACE	$SC_b^{pos}(g_n)$	PLACE	$SC_b^{pos}(g_n)$
神谷町駅 虎ノ門付近	0.90	京急川崎～港町間にあるマンション ラーメン二郎川崎店付近	0.90
京急川崎～港町間にあるマンション ラーメン二郎川崎店付近	0.89	渋谷駅南口付近	0.66
両国二丁目 回向院付近	0.81	東京ドーム	0.65
道玄坂上交差点付近 東急ホームズ前 道玄坂ヒミツキチラボ	0.81	けやきひろば (さいたまスーパーアリーナ前の広場)	0.64
新横浜公園	0.79	万世橋交差点 セガ前 秋葉原から南西に約 100m	0.64
御茶ノ水駅から北東約 200m のビル街	0.75	ZARA 渋谷店前 井の頭通り 渋谷ロフト付近	0.63
神田明神付近 江戸神社付近 御茶ノ水駅から北東に約 200m の宮本公園	0.65	東京ビックサイト	0.62
大塚駅付近	0.65	渋谷駅から北に約 100m レストラン街	0.62
明治神宮野球場 ゴルフプラザ神宮	0.65	五反田駅西口	0.62
中野サンモール (商店街)	0.64	目黒駅から西北西に約 500m の住宅街	0.62

表 4 引きずり仮説に基づいた掘り出し物スポットのランキング

閾値 0.2		閾値 0.3	
PLACE	$SC_b^{pos}(g_n)$	PLACE	$SC_b^{pos}(g_n)$
両国駅から南西に約 200m のビル街	0.98	五反田駅前付近	0.77
ガイア分倍河原店 (パチンコ屋) 分倍河原駅付近 飲屋街	0.81	中目黒駅から南南西に約 1km の住宅街	0.64
新宿トヨタカローラ前 au ショップ前 新宿 4 丁目交差点付近	0.70	BAKE 前 サイゼリヤ前 自由が丘駅付近	0.64
品川駅 (東海道新幹線乗り換え付近)	0.68	立川駅 (南口周辺)	0.63
相模大野駅	0.68	上野駅	0.61
五反田駅前付近	0.67	自由が丘駅前周辺	0.61
池袋サンシャインシティ	0.67	新宿 LIMINE EST	0.61
新宿区役所本庁舎	0.66	歌舞伎町 (HUB 新宿歌舞伎町店付近) 新宿東方ビル周辺	0.61
横浜アリーナ	0.66	明治神宮野球場	0.60
中野サンプラザ	0.64	東京ディズニーシー (リドアイル付近)	0.60

表 3 の閾値 0.2 について、期待仮説を用いた時の掘り出し物スポットのランキングは、上位に住宅街や駅が多く、観光スポットはほとんどランクインしなかった。一方、閾値 0.3 について、期待仮説を用いた時の掘り出し物スポットのランキングは、上位に住宅街や駅が含まれるものの、東京ドームやさいたまスーパーアリーナ近くなど観光スポットがランクインした。表 4 の閾値 0.2 について、引きずり仮説を用いた時の掘り出し物スポットのランキングは、上位 5 位までは駅付近や雑居ビルが、6～10 位は著名な観光スポットがランクインした。また、閾値 0.3 について、引きずり仮説を用いた時の掘り出し物スポットのランキングは、駅やその周辺が多くランクインした。東京ディズニーシーなどの観光スポットも少数ではあるが見受けられた。

がっかりスポットのランキングについても同様に作成した。表 5 は期待仮説に基づいて作成したがっかりスポットのランキングを、表 6 は引きずり仮説に基づいて作成したがっかりスポットランキングである。表 5 の閾値 0.2 について、期待仮説を用いた時のがっかりスポットのランキングは、上位 10 件に駅や住宅街がほとんどで観光スポットを含むようなセルはなかった。同様に、閾値 0.3 について、期待仮説を用いた時のがっかりスポットのランキングも、上位の多くは駅や住宅街であった。表 6 の閾値 0.2 について、引きずり仮説を用いた時のがっかりスポットのランキングは、上位 10 件に駅や交差点が見受けられる。一方、寺やショッピングモールもランクインした。閾値 0.3 について、引きずり仮説を用いた時のがっかりスポットのランキングは、上位 10 件では駅や周囲に店舗の多い交差点や道路が主にランクインした。

次に評価実験の結果を示す。表 7 は、表 3～表 6 のランキングそれぞれに対応する順位相関係数をそれぞれ示している。表

7 が示すように、引きずり仮説を用いた時のがっかりスポットランキングにおいて閾値 0.2 のとき順位相関係数が-0.367、閾値 0.3 のとき順位相関係数が-0.333 とそれぞれ弱い相関を示した。一方で、期待仮説を用い閾値 0.2 の時の掘り出し物スポットランキングでは弱い逆相関を示した。また、それ以外のランキングはベースラインを含め相関がなかった。

5. 考 察

表 3～表 6 より、上位には観光スポットだけでなく住宅街や駅が多く現れた。住宅街が多いのは、ある一人のユーザが数多く Tweet を投稿しているためだと推察される。また、駅が多いことも、駅は待ち時間が多いスポットであることや移動手段に利用する人が多いこともあり、Tweet が多く投稿されるためであると推察される。他にも東京ドームや横浜アリーナなどの球場やイベント会場もランキング上位に現れている。これは規模の大きいイベント会場であるため、感情が大きく変化することが多いスポットであるためと推察される。

表 7 において、引きずり仮説の閾値が 0.2, 0.3 とした場合のそれぞれのがっかりスポットのランキングから弱い負の相関が得られた。この結果より、引きずり仮説を用いて、がっかりスポットランキングを作成した時が、一番目的に近いランキングを作成できた。しかし、他のランキングは無相関、逆相関であった。そのため、掘り出し物スポット・がっかりスポットを発見するためには別の手法を考える必要があると推察される。これらの手法で相関が得られなかった原因として、観光地に対しての予想や期待感が多くの Tweet に含まれていないためであったと推察される。

また、表 7 において、掘り出し物スポットよりがっかりスポットの実験結果がアンケート結果に相関があった。原因とし

表 5 期待仮説に基づいたがっかりスポットのランキング

閾値 0.2		閾値 0.3	
PLACE	$SC_a^{neg}(g_n)$	PLACE	$SC_a^{neg}(g_n)$
両国駅から南西に約 100m 薬局, ファミマ前	0.93	五反田駅前付近	0.71
溝の口駅前ペDESTロリアンデッキ 溝の口駅付近	0.71	自由が丘駅北口付近 飲食店街	0.61
五反田駅前付近	0.68	下北沢駅 (北口方面)	0.61
駒沢大学駅西口前 ラーメン駒大商店 住宅地	0.66	保谷駅から北西に約 700m の住宅街	0.59
西武新宿駅 西武新宿ペペ 新宿プリンスホテル付近	0.65	明治通り 宮益坂下付近 周囲はカフェ, サロン, レストラン, ゲームセンターなど	0.59
拝島駅	0.65	新柴又駅から東南東に約 500m の住宅街 江戸川河川敷付近	0.59
自由が丘駅北口付近 飲食店街	0.64	新宿駅西口地下前タクシー乗り場付近 マクドナルド付近	0.58
品川駅 (北改札内付近)	0.63	LUMINE 立川	0.57
渋谷区役所前交差点 NHK スタジオオパーク付近 渋谷税務署付近	0.63	池袋東口付近 ビックカメラ周辺	0.57
新百合ヶ丘駅 (北口付近)	0.62	国道 2 0 号線 新宿駅東南口付近 周囲に飲食店, ゲームセンター	0.57

表 6 引きずり仮説に基づいたがっかりスポットのランキング

閾値 0.2		閾値 0.3	
PLACE	$SC_b^{neg}(g_n)$	PLACE	$SC_b^{neg}(g_n)$
両国二丁目 回向院付近	0.92	池袋駅から北東に約 400m のビル街	0.67
新横浜公園	0.86	道玄坂二丁目交差点 ドンキホーテ渋谷店前	0.66
大崎駅前周辺	0.71	五反田駅西口周辺 周囲はビル街	0.61
道玄坂上交差点付近 東急ホームズ前 道玄坂ヒミツキチラボ	0.69	渋谷駅周辺宮益交差点付近 タイトステーション前 カフェ, レストランなど	0.61
竹下口交差点付近 (明治神宮駅付近) 周辺はアパレルショップ	0.68	渋谷モディ付近 APPLE STORE 付近 ディズニーストア付近	0.61
川崎 LA CITTADELLA	0.66	東京駅構内 レストラン街付近	0.60
五反田駅西口周辺 周囲はビル街	0.66	上北沢駅から北西に約 200m の住宅街	0.60
明治通り明治神宮前駅周辺	0.66	中野駅	0.60
さいたまスーパーアリーナ けやきひろば近く	0.65	渋谷駅から北に約 100m レストラン多数	0.59
銀座コリドー街付近	0.64	渋谷駅から北西に約 300m の明治通り沿い ZARA 渋谷店付近	0.58

表 7 ベースライン・実験結果とアンケート結果との順位相関係数

対象ランキング	相関係数
ベースライン	-0.043
期待仮説 閾値 0.2 掘り出し物スポットランキング	-0.324
期待仮説 閾値 0.3 掘り出し物スポットランキング	-0.226
引きずり仮説 閾値 0.2 掘り出し物スポットランキング	0.060
引きずり仮説 閾値 0.3 掘り出し物スポットランキング	0.041
期待仮説 閾値 0.2 がっかりスポットランキング	0.017
期待仮説 閾値 0.3 がっかりスポットランキング	-0.124
引きずり仮説 閾値 0.2 がっかりスポットランキング	-0.367
引きずり仮説 閾値 0.3 がっかりスポットランキング	-0.333

て、ユーザがどこかへ訪問する時に期待していないスポットに訪れることは少なく、期待を上回る可能性がある場所にあまり訪れないことが考えられる。結果として、ユーザが訪問する時に、期待して訪れた場所が期待通りであったか、それを下回ることが多いと推察される。

引きずり仮説においても、期待仮説においても、スポットが直後の Tweet , 直前の Tweet に影響を及ぼしている Tweet のペアはほとんど見受けられなかった。そこで、今回は、表 8 に示す、引きずり仮説を用いて抽出された Tweet とその直後の Tweet の一部について考察する。表 8 において、Tweet に含まれる URL やハッシュタグは省略した。1 つ目、2 つ目の Tweet のペアは引きずり仮説を用い、閾値 0.2 の時のがっかりスポットランキングの上位 3 件と感情変化が大きいと予想されるさいたまスーパーアリーナ付近 (9 位) で投稿されたものである。1 つ目の Tweet のようにスポットに関する感情の変化ではないが、抽出された Tweet より直後の Tweet の方が感情スコアが下がっている Tweet ペアを抽出できた。一方で、2 つ目の Tweet のように感情に関係ないと思われる Tweet も抽出

されてしまっている。抽出された Tweet の多くがこのように実際は感情が変化していないような Tweet だった。また、3 つ目、4 つ目の Tweet のペアは引きずり仮説を用い、閾値 0.2 の時の掘り出し物スポットランキングの上位 3 件と感情変化が大きいと予想される横浜アリーナ (9 位) で投稿されたものである。3 つ目の Tweet のペアでは抽出された Tweet は個人的な出来事に対する感情であるが、直後の Tweet はスポットに関しての感情を含んでいるものが抽出できたと推察される。4 つ目の Tweet ペアのように、抽出された Tweet がネガティブな感情を含み、相対的に直後の Tweet の感情スコアが大きくなった例も存在した。このように、弱い相関が認められたとはいえ、その内容には多くのノイズを含んでいることがわかった。また、除去しきれていない BOT などの固定された表現を含む Tweet も残っていた。データ整形の段階でよりノイズを除去することでランキングの相関はより大きくなると推察される。

ランキング外で、著名な観光スポット周辺で投稿された Tweet についても確認した。5 つ目、6 つ目のペアのように、引きずり仮説を用いた中には、直前の Tweet はスポットと関係のないが、直後の Tweet はスポットに関する情報を含む場合が存在した。また、7 つ目のペアでは、テキストの内容から感情の変化が読み取れないものも存在した。これらの原因として、Tweet の感情スコアの計算方法が不適切だったことが推察される。実際、7 つ目の抽出された Tweet の感情スコアは-0.47 で、その直後の感情スコアは-0.82 であった。直後の感情スコアが低くなったのは、「死ぬ」という単語が辞書のスコアでは強いマイナスの意味を持つためである。一方で、ユーザは「死ぬ」という言葉を強いプラスの意味で使用している。このように前後の文脈や日本語の言い回し、独特な表現を使用している場合、感情スコアと実際の Tweet の感情に大きなギャップが生じている

表 8 引きずり仮説により抽出された Tweet とその直後の Tweet の内容 (一部)

	抽出された Tweet	直後の Tweet
1	今回の旅、最大の試験、ついに終わりました、みなさんおつかれさまでした！	まだ悔しいことありますけど、ね
2	車内アナウンスは普通だな	びよんびよんしたいよー
3	胃が。苦しい	MM6 やばかったかわいかったどうしよう
4	足が疲れるううう	ノンスไตล์いいね！ 井上の聞き直り感 笑
5	雨が降りそう。。。	久しぶりにボンルパのバゲット！ シラーにおでん！ さいこーじゃない？ そして、すでに酔いました！
6	めぐチャンスじゃん！	雨降ってるなあ、博多はバリ晴れてたから傘なんて用意してねえぞ。荷物満載でバス停から歩くの正直言ってシンドイ！
7	くはは。なかなかの神席かもしれん。鼻血でる。	やばい。2 列目。花道。くそ近い。死ぬ自信ある。
8	パンフレットくばりはじめた	やばいおなかいたいぞ
9	今日はここから！ 絶対勝つぞ！	まさかの三凡

ことがわかる。より相関の大きいランキングを作成するためには、上記の問題を解決必要があると推察される。

8 つ目のペアにおいては、スポットによる感情変化ではなく、個人の状態が変化したことによる感情変化である。本手法ではこのような Tweet に対応することが難しいと推察される。そのため、このような Tweet をあらかじめ除外するなどの対処が必要であると考え、9 つ目のペアは野球場で投稿された Tweet である。これらの Tweet は、スポットに対する感情変化を示してはいないが、そこで行われた試合に対して感情変化があったと思われる。そのため、このスポットを含むセルは、感情変化を起こす事象が発生する可能性があるセルであると推察される。本手法で意図したものではないが、感情変化が起きやすいスポットのひとつとして Tweet の中から抽出することができた。

6. ま と め

本論文では、Twitter に投稿されたテキスト情報を利用して「多くの人が良くも悪くも期待を裏切られたと感じるスポット」を発見する手法を提案した。期待仮説と引きずり仮説という 2 つの仮説を元に実験を行い、掘り出し物スポットとがっかりスポットのランキングをそれぞれ作成し、被験者が作成したアンケート結果と比較した。その結果、引きずり仮説を用い作成したがっかりスポットランキングとアンケートから作成したランキングに弱い相関がみられた。しかし、それ以外のランキングについては、無相関、逆相関であった。

今後の課題について述べる。今回の実験では、抽出したそれぞれのセルに対して名前を手で付与した。しかし、本研究で得られたギャップスポットをユーザに提示する際に、ユーザにとってわかりやすいスポットを表す名前を自動的に作成することは課題のひとつである。また、投稿時の間隔の差を変更することによる相関や、ランキングの変化について検討する。実際に Tweet の内容を確認したところ、訪れたスポットに対してスポット名を直接に含む Tweet は多くない。そこで、スポット名を含まない Tweet と、実際に訪れたスポットを関連づけることを検討する。最後に、本論文で用いた Tweet の感情スコアの計算方法を検討することがあげられる。

謝 辞

本研究は、首都大学東京傾斜的研究 (全学分) 学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」及び JSPS 科研費 16K00157, 16K16158 による

文 献

- [1] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, Vol. 1, No. 12, 2009.
- [2] 伊川 洋平村上 明子. Twitter を用いた災害時の住民感情の分析. 第 7 回 データ工学と情報マネジメントに関するフォーラム, pp. C7-4, 2015.
- [3] Kristina Lerman, Megha Arora, Luciano Gallegos, Ponnurangam Kumaraguru, and David Garcia. Emotions, demographics and sociability in twitter interactions. In *Tenth International AAAI Conference on Web and Social Media (ICWSM2016)*, 2016.
- [4] 菅原久嗣, Alena Neviarouskaya, 石塚満. 日本語テキストからの感情抽出. 第 23 回 人工知能学会 全国大会 2009, pp. 1-2, 2009.
- [5] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsn*, Vol. 11, pp. 538-541, 2011.
- [6] 三好辰明, 太田学. ツイートに出現する顔文字等の文字と記号に着目した感情分類. 第 5 回 データ工学と情報マネジメントに関するフォーラム, pp. D9-2, 2013.
- [7] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, Vol. 8, No. 5, p. e64417, 2013.
- [8] 小原基季, 森田和宏, 泓田正雄, 青江順一. Twitter 本文を用いた観光情報抽出及び分析システムの構築. 人工知能学会第 29 回全国大会講演論文集, 2015.
- [9] 奥村秀人, 徳久雅人, 村上仁一, 村田真樹. 観光地に対する長所情報の収集と分類の試み. 電子情報通信学会技術研究報告. TL, 思考と言語, Vol. 110, No. 244, pp. 25-30, 2010.
- [10] 杉本祐介, 水野忠則, 菱田隆彰. 口コミに含まれる感情語を利用した観光地分類の検討. マルチメディア, 分散協調とモバイルシンポジウム 2014 論文集, Vol. 2014, pp. 1345-1350, 2014.
- [11] 村上嘉代子, 川村秀憲. 外国人から見た日本旅行: 英語ブログからの観光イメージ分析. 人工知能学会誌, Vol. 26, No. 3, pp. 286-293, 2011.
- [12] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, pp. 133-140. Association for Computational Linguistics, 2005.