

スポーツを「いいとこどり」で観戦するための実況ツイート解析手法

藤本 和[†] 牛尼 剛聡^{††}

[†]九州大学芸術工学部 〒815-8540 福岡県福岡市南区塩原 4-9-1

^{††}九州大学芸術工学研究院 〒815-8540 福岡県福岡市南区塩原 4-9-1

E-mail: [†]fujimoto.nodoka.017@s.kyushu-u.ac.jp, ^{††}ushiana@design.kyushu-u.ac.jp

あらまし 近年、SNSの利用用途としてソーシャルビューイングが注目されている。ソーシャルビューイングは、TV番組やスポーツ中継などを視聴しながらコメントや応援などをSNSに投稿する視聴形態であり、代表的なSNSの一つであるTwitterで投稿されるツイートは「実況ツイート」と呼ばれる。また、スポーツ中継に関するサービスも様々なものが存在する。例えば、ユーザが忙しいときには、中継の全てを視聴することができないため、得点が入ったことを通知するサービス等が活用されている。しかし、そのようなサービスで行っていることは、既に起こってしまった出来事の通知である。もし、その出来事をリアルタイムで体験できればユーザの満足度が向上すると考えられる。そこで、スポーツの試合中に、「いいところ」が期待できるときに、ユーザに通知を行うことで、この問題を解消しようと考えた。本研究では、野球の試合が行われている時に、「いいところ」となる兆候をリアルタイム検出することを目的とする。具体的には、実況ツイートを取得してLDAによるトピック抽出と重回帰分析を用いてファンの喜び度合い（Pleasure値）を推定し、LSTMを利用してPleasure値を予測することで試合の「いいとこどり」を行う。

キーワード Twitter, 実況ツイート, LDA, トピック抽出

1 はじめに

近年のソーシャルネットワークサービス（SNS）の普及により、誰もが気軽に情報を発信できるようになった。SNSは情報収集、宣伝、コミュニケーションなど様々な用途のために利用されており、SNSの利用方法のひとつとして、ソーシャルビューイングがある。ソーシャルビューイングは、ドラマ、アニメ、スポーツ中継などを視聴しながらコメントや応援などを投稿する視聴形態であり、代表的なSNSであるTwitterで投稿されるツイートは「実況ツイート」と呼ばれている。ソーシャルビューイングでは、TV番組を視聴しながら実況ツイートを見ることで、同じ番組の視聴者とその感想や意見を共有して盛り上がるができる。

また、近年、スポーツ中継に関する様々なサービスが登場している。例えば、ユーザが忙しいために中継の全てを視聴することができないときは、スポナビ[1]等のサービスが役に立つ。スポナビはお気に入りのチームを登録することで、設定によって追加点を入れた時に通知することができる。このようなサービスを利用することで、気になっている試合の状況がある程度リアルタイムに知ることができる。しかし、これらのWebサービスや通知サービスは、過去の出来事を伝えるものであり、その出来事をリアルタイムに体験することはできない。そのため、中継をリアルタイムに視聴をしている時とは大きく満足度が異なる。得点の通知や試合結果がわかってから編成されるダイジェストだけでは試合の「流れ」や「ファンの盛り上がり」などの臨場感を伝えることはできない。また、そのような臨場感を共有するためには常に中継を見ておく必要があり、ユーザに対

する時間的な負担が大きい。結果から振り返って「いいところ」を知らせるのではなく、リアルタイムで「いいところ」になりそうな時を通知することができれば、リアルタイムに「いいところ」だけを見たいというユーザの満足度の向上につながる。

従来の野球中継の特徴的シーンを抽出する研究では、試合が終わった後に試合全体からシーンを選択している[2]~[5]。この方法では、ユーザはリアルタイムに「いいところ」を体験することはできない。そこで、リアルタイムに観戦者の主観的データを分析することによって、「いいところ」になる予測を行う。本研究では、観戦者の主観的データとして試合の「流れ」や「ファンの盛り上がり」などを反映していると考えられる、実況ツイートをを用いる。「いいところ」には様々な基準が存在するが、本研究では「ファンの喜び度合い」が高いところを「いいところ」と考える。「ファンの喜び度合い」とは、ファンが観戦している試合に対してポジティブな印象を持つ度合いである。実況ツイートから「ファンの喜び度合い」を推定し、将来における「ファンの喜び度合い」の高い「いいところ」となる可能性が高い部分予測することで、ユーザが試合の「いいとこどり」をした観戦を可能とする。

2 アプローチ

本研究では、野球中継に対するの実況ツイートを分析し、リアルタイムに「ファンの喜びの度合い」に基づいて「いいところ」を予測することを目的とする。

実況ツイートは、チームに対する応援や感想などが投稿されており、試合の状況が変化すると視聴者が投稿する内容が異なる。試合の状況は、客観的なデータからはわからない要素が多

く存在し、同じ結果であっても視聴者の反応は異なる。そのため、観戦者の主観を表すデータとして実況ツイートは有用だと考えた。

そこで、視聴者が投稿した実況ツイートのテキストから喜怒哀楽等の感情を分類することが考えられる。従来、感情辞書から動画コンテンツにおける特徴的なシーン抽出をする研究は多く行われている。しかし、実況ツイートには、一般的に感情を表してはいない「ホームラン」、「ヒット」、「アウト」など出来事の名前のみのツイートなども含まれる。これらは、どちらのチームについての言及なのか、状況によってユーザーの持つ感情は異なる。また、一般的に感情が推定できないようなファン独特の表現である場合や、複雑な状況によりファンの中で必ずしも感情が一致しないという場合がある。そのため、ツイートの単純な感情分類は適切でないと考えた。

上記の問題を解決するために、本研究ではファンがどの話題でどのくらい盛り上がっているかに注目する。本研究ではユーザーの盛り上がりを出すために、特徴量としてツイート数とトピックを利用する。経験的に、スポーツの実況ツイートのツイート数は、ユーザーの興味を引く状況において増加する傾向にある。また、トピックは、感情だけではなく投稿されている話題の分布から、状況の種類を抽出できる。

本研究の提案手法の概要と概要図を以下に示す。①～③で、試合中の短い時間区間での特徴から「ファンの喜びの度合い」を推定し(図1)、④で、その値に基づいて未来の「ファンの喜びの度合い」を予測する(図2)。

- ① 実況ツイートの収集
- ② 状況の特徴分析
 - 窓関数を利用して試合中の特定の時間窓幅 w を設定する
 - 窓区間内のツイートを取り出し1つの文書とする
 - 時間窓ごとに品詞ごとのBow(Bag of Words)を生成する
 - 単語の正規化として、汎化とフィルタリングを行う
 - 品詞ごとにLDA [6]を用いたトピックの抽出を行う
 - 時間窓ごとにトピック分布の特徴を求める
 - 時間窓ごとのツイート数、差、速度の特徴を求める
- ③ ファンの喜び度合いの推定する式の導出
- ④ 「いいところ」の予測

3 ファンの喜び度合いの推定手法

3.1 試合状況の分析

野球中継を対象に、応援するチームのハッシュタグを利用して実況ツイートを収集する。収集した実況ツイートを投稿時刻にもとづいて時系列化し、そこから試合の状況を抽出するために、時間窓を適用してツイート数とトピックの特徴を取り出す。

3.2 時間窓の適用

窓関数を利用して時間窓幅 w におけるツイート数とトピックを利用して分析する。 i 番目の時間窓の時間区間を $(\frac{w}{2}i, \frac{w}{2}(i+2))$ とする。

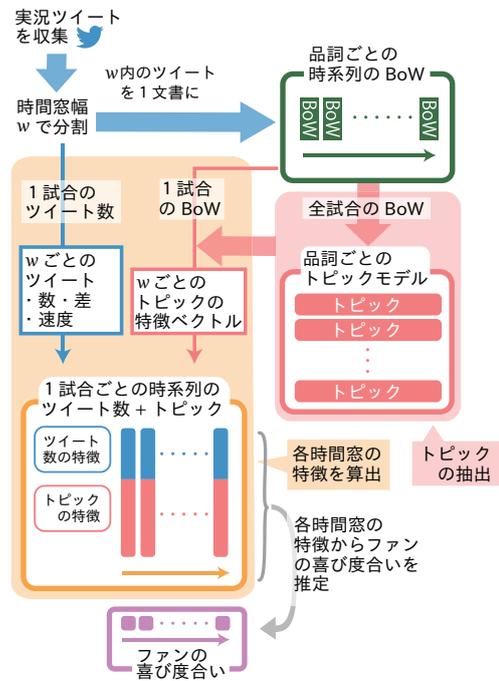


図1 本研究の概要図1

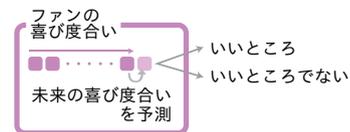


図2 本研究の概要図2

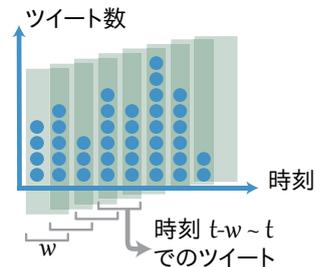


図3 時間窓幅 w の設定

3.2.1 品詞ごとの時系列の BoW を生成

実況ツイートに出現する単語は、品詞によって個数が異なる。また、LDA を用いて時間窓ごとのトピックの特徴を抽出する際に、品詞の組み合わせを変更してトピックを抽出する。そこで、時間窓ごとに取り出されたツイート文書を MeCab によって形態素解析を行い、時間窓ごとに品詞ごとの BoW(Bag of Words) として特徴を表現する。BoW とは、並び順を考えない単語の集まりである。品詞は、一般的にストップワードとならない名詞、動詞、形容詞、形容動詞、副詞を扱う。

形態素解析は、新語や固有表現に対応できる mecab-ipadic-NEologd 辞書 [7] を用いる。この辞書は Web 上の言語資源から得た新語を活用しており、実況ツイートのような、ネット用語や SNS 上に現れやすい崩れ表記語等にも対応しているため、本研究の用途に適している。そして、形態素解析をした単語の原形を BoW の要素とする。

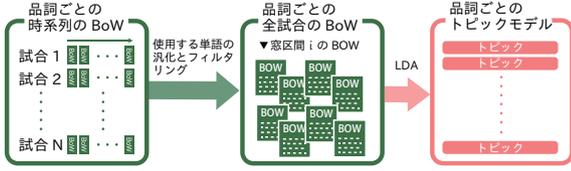


図4 単語の汎化とフィルタリングとトピックの抽出

3.2.2 単語の正規化

単語の文字種や表記揺れの統一を行うことで、計算量を少なくし、まとまりのあるトピックの抽出を行えるように前処理を行う。本研究の目的に沿って単語を汎化し、またストップワードを取り除く。

実況ツイートでは、多くの選手名やニックネームが出現しており、このままトピックを抽出すると選手名によってトピックの特徴が決まってしまう恐れがある。そこで、味方チームの選手であれば「味方」、敵チームの選手であれば「敵」という単語に置き換え、選手名を汎化する。また、打率等の成績などの数字が頻繁に現れる。「1点」「3回」などの単語は、単位ごとにひとつにまとめ、「(数)点」「(数)回」となるよう汎化する。多くのツイートの中で数回しか出現しない単語は、試合に関係のないツイート文章の一部であったり、誤字や表記揺れである可能性が高い。それらの単語を BoW から削除するために、単語の出現頻度によって、使用する単語を制限する。また、本研究の目的に沿わない単語をストップワードとして削除する。

これらの処理を行った BoW を、トピックの抽出やトピックの分布を求める際に使用する。

3.2.3 トピックの抽出

LDA (潜在的ディリクレ配分法/Latent Dirichlet Allocation) [6] を用い、実況ツイートに対するトピックの抽出を行う。LDA とは、並び順を考えない単語の集まり BoW で表現された文書の生成過程をモデル化したものであり、潜在的なトピックを複数の単語の分布で表現することができる。

実況ツイート収集した全試合の全時間窓から、トピックの抽出を行う。品詞ごとに出現頻度が大幅に異なるため、品詞の種類とトピック数を検討し、トピック抽出を行う。トピック抽出の概要を図4に示す。

3.2.4 トピックの特徴の算出

LDA によって抽出した複数のトピックと、窓区間ごとの BoW から、それぞれの窓区間におけるトピックの確率分布を算出する。この分布をトピックの特徴のとして扱い、時間窓ごとのトピックの確率ベクトルを求める。品詞 p のトピック数を N_p とすると、時間窓内の BoW から N_p 次元のトピックの確率ベクトルが算出できる。窓区間 i と品詞の種類 p ごとのトピックの特徴ベクトルを $topic_i(p)$ とし、以下の (1) 式で表す。

$$topic_i(p) = \begin{bmatrix} topic_i(p, 0) \\ \vdots \\ topic_i(p, N_p) \end{bmatrix} \quad (1)$$

3.2.5 ツイート数の特徴の算出

ツイート数の特徴として、ツイート数、ツイート数の前後の差、ツイート数の増減の速度を考える。

幅 w の窓区間ごとのツイートの数を $Count$ 、1 つ前の窓区間内のツイートとの差を $Diff$ 、ツイートの増減の速度を $MACD$ とする。窓区間 i でのツイート数を f_i としたとき、窓区間 i の $Count$ ($Count_i$) と $Diff$ ($Diff_i$) は、それぞれ (2) 式と (3) 式で表される。Diff は一瞬のツイート数の増加を表すことができる。MACD (移動平均線収束拡散法/Moving Average Convergence Divergence) [8] とは、主に株価や為替市場のトレンド分析などで用いられる指標である。値の変動の速さを表し、Diff と違い、ツイート数のゆるやかな増減を表すことができる。これは、長さの異なる EMA (指数移動平均/exponential moving average) から計算され、(4)、(5) 式で表される。 n_1 と n_2 は、経験的に $n_1 = 2$ 、 $n_2 = 4$ とする。

$$Count_i = f_i \quad (2)$$

$$Diff_i = f_i - f_{i-1} \quad (3)$$

$$EMA_i^{(n)} = \sum_{k=0}^{n-1} \alpha(1-\alpha)^k f_{i-k} \quad \left(\alpha = \frac{2}{n+1} \right) \quad (4)$$

$$MACD_i = EMA_i^{(n_1)} - EMA_i^{(n_2)} \quad (5)$$

これらの値から、窓区間 i でのツイート数の特徴ベクトル $twNum_{(i)}$ を以下の (6) 式で表す。

$$twNum_{(i)} = \begin{bmatrix} Count_i \\ Diff_i \\ MACD_i \end{bmatrix} \quad (6)$$

3.3 窓区間ごとの Pleasure 値の推定

窓区間の試合状況の特徴から、その窓区間の「ファンの喜び度合い」を推定する。窓区間 i のツイート数の特徴ベクトル $twNum_i$ と、品詞の種類 (p_1, p_2, \dots) ごとのトピックの特徴ベクトル $topic_i(p)$ を組み合わせ、窓区間 i の「試合状況ベクトル」 $\mathbf{x}_{(i)}$ を構成する。このベクトルから、 $\mathbf{x}_{(i)}$ と同じ次元の係数 \mathbf{c} のベクトル及び切片 $intercept$ を用い、 $\mathbf{x}_{(i)}$ を変数とした、窓区間 i における Pleasure 値の推定値を求める式を以下の (7)、(8) 式のように導出する。

$$Pleasure_i \text{の推定値} = \mathbf{c}^T \mathbf{x}_{(i)} + intercept \quad (7)$$

$$\mathbf{x}_{(i)} = \begin{bmatrix} Count_i \\ Diff_i \\ MACD_i \\ topic_i(p_1, 0) \\ \vdots \\ topic_i(p_2, 0) \\ \vdots \end{bmatrix} \quad (8)$$

ここで、係数のベクトル \mathbf{c} と切片 $intercept$ を決定する必要がある。そのため、時間区間内のツイート数とトピックの特徴ベクトルと「ファンの喜び度合い」の関係について、重回帰分

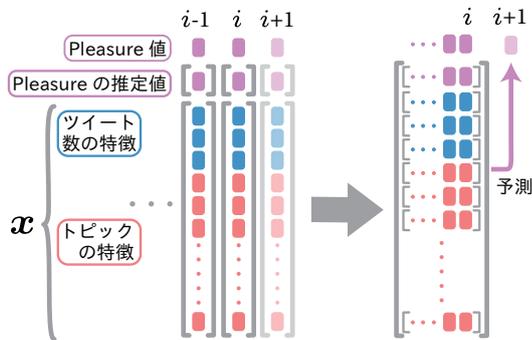


図 5 予測モデルの概要図

析を行う。目的変数となる「ファンの喜び度合い: *Pleasure*」の値は、被験者実験における主観評価によって正解データを生成する。被験者実験では、被験者は窓区間 i の実況ツイートを見て、「ファンの喜び度合い」を 5 段階 (1~5) で評価する。重回帰分析の結果から求められる偏回帰係数を c 、誤差 (切片) を *intercept* とし、関数のパラメータを求める。

4 「いいところ」の予測

4.1 未来のファンの喜び度合いの予測

「いいところ」の予測を行うために、未来のファンの喜び度合いである *Pleasure* 値を LSTM (Long short-term memory) [9] モデルを用いて予測する。LSTM は代表的な再帰的ニューラルネットワーク、RNN (Recurrent Neural Network) の一種である。RNN は系列データを扱うためのモデルで、前時刻の中間層を現時刻の入力としても用いて学習を行う。RNN には学習時に起こる勾配が減衰するという問題が存在するが、それを解決したものが LSTM である。

LSTM モデルを使って、図 5 のように、窓区間 i の試合状況ベクトル $x_{(i)}$ と、3.3 の *Pleasure* 値の推定手法から求めた $Pleasure_i$ の推定値から、窓区間 $i+1$ の $Pleasure_{i+1}$ の値を予測するモデルを生成する。LSTM で学習する時間窓の個数 (*look.back*) を定め、例えば *look.back*=2 の場合は、 i と $i-1$ のデータを用いて $i+1$ の予測を行う。

4.2 「いいところ」の予測

予測されたファンの喜び度合いに対して閾値を定め、値が閾値以上となった時に「いいところ」としてユーザに通知する。

5 パラメータの決定

1 試合のうちの時間区間内のトピックの特徴のパラメータを決定するため、まず収集した試合の実況ツイート全体からトピックを抽出する。トピックの抽出には、BoW を作成するための時間窓幅 w と、まとまりの適切なトピックを抽出するために、使用する単語を汎化及びフィルタリングして種類を決定する必要がある。

本研究では、阪神タイガースのファンの実況ツイート (#han-shin を含むツイート) を収集し、トピック抽出を行った。

表 1 トピックを抽出した単語の種類

パターン	使用する品詞	出現頻度の閾値	単語の種類
①	名詞	80	569
②		100	458
③			
④	動詞	10	554
⑤		20	366
⑥		30	267
⑦	形容詞形容動詞	1	312
⑧		5	180
⑨		10	133
⑩	副詞	10	219
⑪		20	129
⑫		30	96
⑬	(1) 形容詞形容動詞	(1)5, (2)5	411
⑭	, (2) 副詞	(1)5, (2)10	399
⑮		(1)5, (2)20	309
⑯	(1) 動詞, (2) 副詞	(1)30, (2)10	486
⑰		(1)30, (2)20	396
⑱	(1) 名詞, (2) 形容詞形容動詞	(1)100, (2)5	638

5.1 時間窓幅の決定

時間窓幅 w を決定するため、時間窓幅を 1 分、2 分、4 分、6 分と変更して検討を行った。

2018 年に行われた 27 試合を対象に、実況ツイート及び「スポーツナビ」から取得した試合の点数差の関係を調査したところ、試合によって実況ツイートの数には、ばらつきがあった。時間窓幅を 1 分、2 分とすると、十分な文書の長さを確保することができず、ツイート数が少ない時間帯の場合は適切にトピックを抽出することができない。6 分の場合は試合の展開が早い場合には不適切であった。よって、本研究では、時間窓幅を 4 分とする。

5.2 使用する品詞の決定

提案手法による単語の汎化を行った上で、名詞を対象に手動で球団名、球場名、明らかに不要な語 (bot で使われる単語や顔文字等に使われる語など) を除外した。

トピック抽出に使用する単語の種類を決定するため、フィルタリングされた BoW を以下の表 1 のように用意し、トピックを抽出する実験を行った。単語のフィルタリングでは、品詞に「名詞」「動詞」「形容詞形容動詞」「副詞」の 4 種類とそれら組み合わせたものを用意し、それぞれ単語の出現頻度の閾値 (単語の最低頻度) を変化させた。

名詞は「ヒット」「ホームラン」「アウト」など、試合の状況を表す重要な単語が含まれており、試合の状況を表すためには必要な要素だと考えた。一方、動詞は、使用する単語の出現頻度の閾値を調整しても、意味的にまとまりのあるトピックとはならなかった。その理由として、野球の実況ツイートは「ホームラン (を打った) !」など、状況を名詞のみで表すものが多いという特徴があり、動詞が隠れている場合があるからだと考えられる。このことから、動詞はトピック抽出には使用しない。形容詞形容動詞と副詞は、実況ツイートを行っているファンの感情をよく表している。ファンの喜び度合いを求めることを目

表 2 名詞のトピック

リクエスト:0.145	セーフ:0.113	バント:0.078	...
プロ:0.335	打点:0.214	ヒット:0.094	...
ゲッター:0.355	併殺:0.211	送球:0.040	...
とら:0.398	カッター:0.050	初勝利:0.047	...
初回:0.108	失点:0.071	先制:0.036	...

表 3 形容詞形容動詞のトピック

うまい:0.118	多い:0.094	やすい:0.091	...
怖い:0.150	すごい:0.061	軽い:0.057	...
浅い:0.150	珍しい:0.135	不甲斐ない:0.063	...
低い:0.140	おもしろい:0.114	美しい:0.108	...
楽しい:0.081	でかい:0.065	凄い:0.062	...

表 4 副詞のトピック

ちゃんと:0.127	どうぞ:0.101	いよいよ:0.083	...
そんなに:0.197	サクサク:0.114	もっと:0.101	...
きっちり:0.214	しっかり:0.096	さすが:0.086	...
なかなか:0.248	まったく:0.185	もっと:0.041	...
あっさり:0.213	あっという間に:0.090	何とか:0.050	...

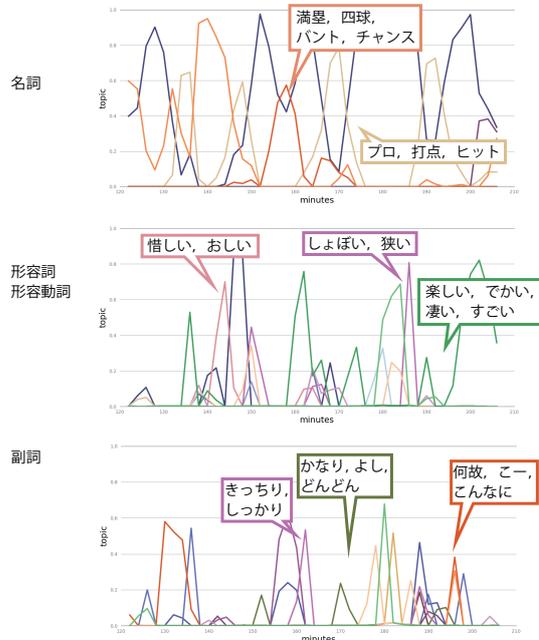


図 6 トピック分布の推移 (例)

的としているため、必要な要素とした。

なお、⑬～⑱のように、種類の異なる品詞を同時に使用するトピック抽出では、ツイート内で出現頻度が高い方の品詞の単語に偏って抽出されてしまった。そのため、種類の異なる品詞を同時に使用する利点はないと考えられる。

すべての品詞において、使用する単語の最低頻度の調整を行った。最低頻度の閾値が低すぎると単語の種類が多くなり、まとまりのないトピックとなってしまう、逆に閾値が高すぎると重要な単語が抜け落ち様々な状況を抽出できなくなってしまった。それらを考慮し、②, ⑧, ⑬の3種類の品詞を利用したトピックモデルが適切だと考えた。

5.3 トピック数の決定

トピック数を変化させてトピックを抽出し、トピックの内容を比較して適切なトピック数を検討し、決定した。

検討したトピック数は20, 30, 40, 50の3種類であった。どの品詞であっても、トピック数15では、トピック数が少なすぎるために異なる種類の単語が同じトピックに出現してしまった。それに対して、トピック数45では、類似した内容のトピックが複数存在する、もしくは同じトピックに似た単語が含まれていた。この結果より、最もトピック数30のときが適切に分類がされていると判断し、トピック数は品詞の種類に関わらず30と定めた。

5.4 トピックの抽出

品詞(名詞, 形容詞形容動詞, 副詞)とトピック数(全て30)を設定し、トピックの抽出を行った。それぞれ30のトピックのうち、主なトピック及びトピックの内容は以下の表2, 表3, 表4の通りである。結果、名詞, 形容詞形容動詞, 副詞の全90個のトピックが抽出された。そのため、窓区間でのトピックの特徴は最大90次元のベクトルとなった。

5.5 トピック分布の算出

各試合の各窓区間に対して試合状況を表すトピック分布を求

めた。図6はその推移の一部である。

6 実験と評価

2018年の阪神タイガースの試合のうち20試合を、勝敗が均等になるよう、勝った10試合、負けた10試合のデータを扱い、16試合(勝:8試合, 負:8試合)を学習データ, 4試合(勝:2試合, 負:2試合)をテストデータとして実験を行った。

6.1 実験手法

まず正解データであるPleasure値を決定したのち、提案手法の中で、Pleasure値の推定式と未来のPleasure値の予測モデルの精度を検証し、「いいとこどり」できるかの性能評価を行う。

6.1.1 正解データの準備

20試合のデータにおいて、窓区間ごとのPleasure値の正解データを決定した。

窓区間*i*でのPleasureの値($Pleasure_i$)を設定するために、ファンの喜び度合いに関する主観評価を行った。クラウドソーシングによって、阪神タイガースの20試合について、1試合に対して5人が窓区間内のツイート群からファンの喜び度合いの5段階評価を行った。評価の外れ値を断定するために、5人中4人以上が同じ評価をしている場合は、その値をPleasure値とした。5人中で評価の数値(1~5)がわれている場合は、トリム平均(Trimmed Mean)を用いて評価の最大値と最小値を取り除いた3つの評価値の平均をとり、Pleasure値とした。

6.1.2 Pleasure値の推定式の定式と評価

試合状況ベクトル \mathbf{x} を説明変数、主観評価によって得られたPleasure値を目的変数として学習データの重回帰分析を行った。偏回帰係数 \mathbf{c} と切片 $intercept$ からPleasure値の推定式を定式し、その推定の精度を評価した。試合状況ベクトル \mathbf{x} は、ツイート数の特徴ベクトル(3次元)とトピックの特徴ベクトル

ル (90 次元) を合わせた 93 次元の \mathbf{x}_{93} である。この比較対象として、トピックの特徴のみの \mathbf{x}_{90} (90 次元) とツイート数の特徴のみの \mathbf{x}_3 (3 次元) も同様に評価した。それぞれの推定式を「Estimate」, 「EstimateNoTw」, 「EstimateNoTp」と呼ぶ。

重回帰分析での推定の精度は、交差検定のうちの k-Fold 法を用い、平均平方二乗誤差 (RMSE : Root Mean Square Error) を求めることを行った。データを分割する個数 k は、5, 10, 20 と変化させて実験し、偏回帰係数および切片は、 k 通りの検定のうち、RMSE が最も小さいときのものを扱った。

6.1.3 Pleasure 値の予測モデルの構築

Pleasure 値の推定値及び試合状況ベクトル : \mathbf{x} から、LSTM によって対象とする窓区間のひとつ次の窓区間の Pleasure 値を予測するモデルを構築した。Pleasure 値の推定値は、3 種類の推定式 (Estimate, EstimateNoTw, EstimateNoTp) から算出された値を用いる。 \mathbf{x} の次元数は、Pleasure 値の推定式を学習した時と同様の 3 種類を扱う。LSTM の学習は、表 5 に示すにパラメータと変数を設定し、 $(A_1) \sim (J_2')$ のモデルを生成する。モデルの評価は、正解となる Pleasure 値との RMSE で行って比較する。

$(A_1) \sim (B_2)$, $(E_1) \sim (F_2)$, $(I_1) \sim (J_2)$ は Pleasure 値の推定値及び試合状況ベクトル \mathbf{x} を用いた予測である。その比較として、試合状況ベクトル \mathbf{x} のみから未来の Pleasure 値を予測する $(C_1) \sim (D_2)$, $(G_1) \sim (H_2)$, $(K_1) \sim (L_2)$ のモデルを用意した。また $(I_1') \sim (J_2')$ は、 $(I_1) \sim (J_2)$ での用いた推定式を Estimate に変更したモデルである。

6.1.4 「いいところ」予測の性能評価

窓区間 i の分析によって予測された $Pleasure_{i+1}$ に対して閾値 $Threshold$ を設定し、 $Pleasure_{i+1}$ が $Threshold$ 以上となったときに、「いいところと予測した」とする。正解となる「いいところ」は、主観評価実験によって定められた Pleasure 値が 4 以上のときとする。表 6 と (9) 式に従い、ROC 曲線 [11] の AUC によって、生成したモデルを評価する。

$$\text{偽陽性率} = \frac{FP}{FP + TN} \quad \text{真陽性率} = \frac{TP}{TP + FN} \quad (9)$$

この評価方法は、偽陽性率を小さく、なおかつ真陽性率を高くするモデルの識別能力を評価できる。ROC 曲線は、予測値の閾値を変化させた時に、偽陽性率を横軸、真陽性率を縦軸としたグラフである。この曲線の AUC (Under the Curve) とは、図 7 のように、ROC 曲線と x 軸 y 軸で囲まれた部分を指す。また、ROC 曲線は、右に行くほど値が下がることがないという特徴があり、偽陽性率の値が小さい時点で、高い真陽性率であるモデルほど良いということから、AUC の面積が大きくなるモデルの方が性能が良いと言える。

$(A_1) \sim (J_2')$ のモデルにおいて、 $Threshold$ の値を $(1 < Threshold < 5)$ の範囲で 0.1 ずつ変化させ、AUC の面積の値によって各モデルの性能を比較する。

6.2 実験結果

6.2.1 Pleasure 値の推定式の定式と評価

重回帰分析による窓区間ごとの Pleasure 値を推定した精度

表 5 LSTM でのパラメータ, 変数

損失関数	mean_squared_error
最適化手法	Adam [10]
バッチサイズ	1

	(A ₁)	(A ₂)	(B ₁)	(B ₂)	(C ₁)	(C ₂)	(D ₁)	(D ₂)
\mathbf{x} の次元	ツイート数+トピック 93 次元							
推定式	Estimate				なし			
入力次元	94				93			
look_back	1		2		1		2	
epoch	100	500	100	500	100	500	100	500

	(E ₁)	(E ₂)	(F ₁)	(F ₂)	(G ₁)	(G ₂)	(H ₁)	(H ₂)
\mathbf{x} の次元	トピックのみ 90 次元							
推定式	EstimateNoTw				なし			
入力次元	91				90			
look_back	1		2		1		2	
epoch	100	500	100	500	100	500	100	500

	(I ₁)	(I ₂)	(J ₁)	(J ₂)	(K ₁)	(K ₂)	(L ₁)	(L ₂)
\mathbf{x} の次元	ツイート数のみ 3 次元							
推定式	EstimateNoTp				なし			
入力次元	4				3			
look_back	1		2		1		2	
epoch	100	500	100	500	100	500	100	500

	(I ₁ ')	(I ₂ ')	(J ₁ ')	(J ₂ ')
\mathbf{x} の次元	ツイート数のみ 3 次元			
推定式	Estimate			
入力次元	4			
look_back	1		2	
epoch	100	500	100	500

表 6 「いいところ」を予測する分割表

	いいところである	いいところでない
いいところと予測	TP (真陽性)	FP (偽陽性)
いいところでないと予測	FN (偽陰性)	TN (真陰性)

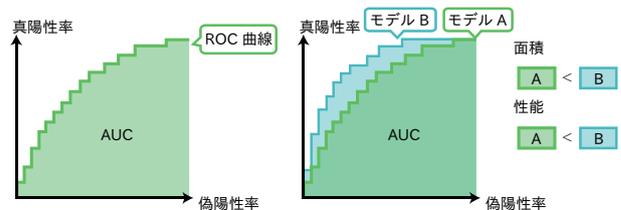


図 7 ROC 曲線と AUC の関係

表 7 学習データ (交差検定), テストデータでの RMSE

推定式	Estimate	EstimateNoTw	EstimateNoTp
\mathbf{x} の次元数	ツイート数+トピック	トピックのみ	ツイート数のみ
	93	90	3
学習データ	0.6901	0.6924	0.8678
テストデータ	0.6407	0.6419	0.7487

は、k-Fold 法を用いた学習データにおいて、 $k=20$ の時の精度が一番高かった。その時の RMSE と、テストデータにおける RMSE は、表 7 の通りである。

結果として、説明変数にトピックの特徴が含まれている Estimate と EstimateNoTw の方が、ツイート数のみの特徴で推定された EstimateNoTp よりも精度が高かった。

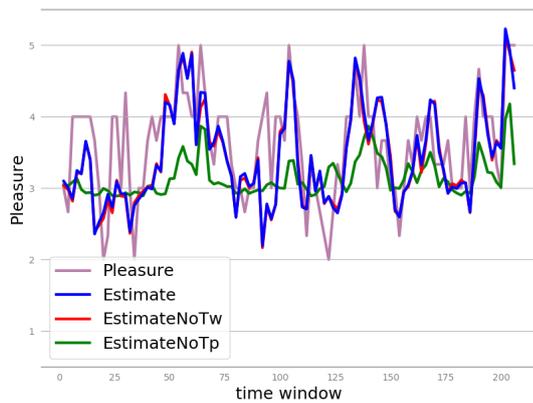


図 8 推定式

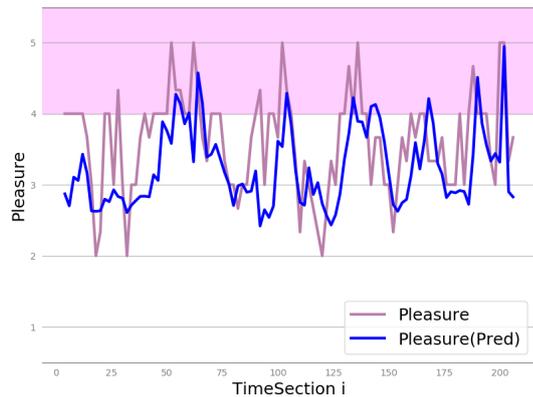


図 10 モデル (J₂') の予測

テストデータにおいて、それぞれの推定式 Estimate, EstimateNoTw, EstimateNoTp と正解データである x にツイート数とトピックの両方の特徴を用いたときの推定式 (Estimate) で、正解である Pleasure とその推定値 (Estimation) を比較したグラフの一部を図 8 に示す。

6.2.2 Pleasure 値の予測モデルの評価

予測された Pleasure 値と正解データである Pleasure との RMSE は、学習データ (train) とテストデータ (test) で、図 9 のようになった。

(I₁')~(J₂') のモデルが、他のモデルと比較して学習データ及びテストデータで高い精度であった。例として、(J₂') の予測モデルでテストデータを予測した値のグラフを図 10 に示す。

6.2.3 「いいところ」予測の性能評価

図 11 は、(A₁')~(J₂') のモデルで、ROC 曲線の AUC の面

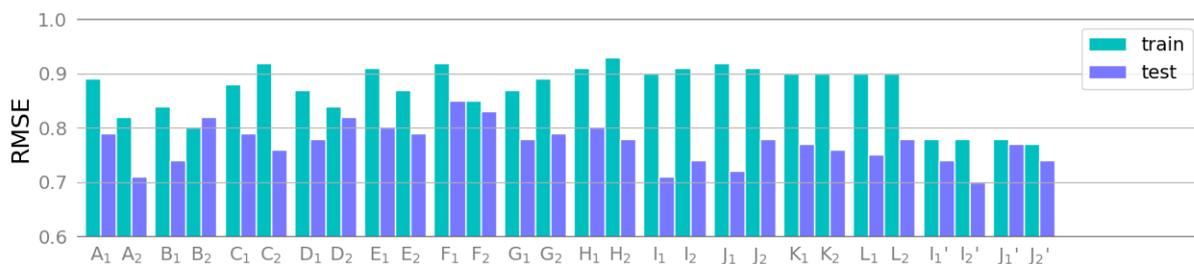


図 9 各モデルの評価 (RMSE)

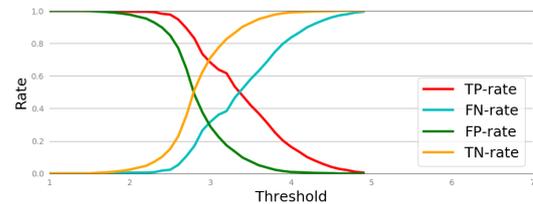


図 12 (J₂') の予測精度 (学習データ)

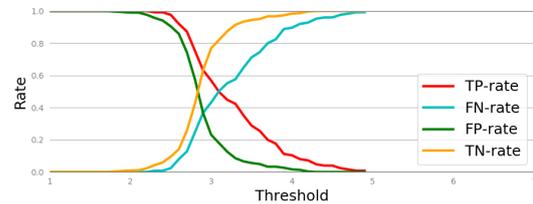


図 13 (J₂') の予測精度 (テストデータ)

積をそれぞれ表したグラフである。モデル (J₂') の性能が一番高い結果となった。

また、(J₂') のモデルにおいて、Threshold の値を (1 < Threshold < 5) の範囲で 0.1 ずつ変化させたときの真陽性率 (TP-rate)、偽陰性率 (FN-rate)、偽陽性率 (FP-rate)、真陰性率 (TN-rate) を、図 12, 13 のグラフに示す。対象の時間窓の数は、学習データで 1631 個、テストデータで 428 個である。

6.3 考察

Pleasure 値の推定式に関しては、トピックの特徴の方がツイート数の特徴に比べて、Pleasure 値への寄与が大きいという結果となった。そして、推定式の変数の要素で、「トピック+ツイート数」と「トピックのみ」では大きな違いは見られなかった。推定式の誤差に関しては、各試合のトピックの特徴抽出が安定化できれば精度は向上すると考えられる。

未来の Pleasure 値の予測モデルの評価では、look_back の値の違いによる精度の相関は見られなかった。(A₁')~(B₂') と (E₁')~(F₂') を比較すると、推定式 Estimate と EstimateNoTw の精度が大きく変わらないことから、試合の状況を表す x のベクトルにツイート数の特徴が含まれていることは重要だと判断できる。また、(I₁')~(L₂') と (I₁')~(J₂') を比較すると、(I₁')~(J₂') で使用した推定式 Estimate の有効性がうかがえる。(A₁')~(B₂') 及び (E₁')~(F₂') は、学習の入力ベクトルの次元が多ことから、学習がうまく進まなかったと考えられる。これらのことから、(J₂') のように、Pleasure 値の値をツイート数とト

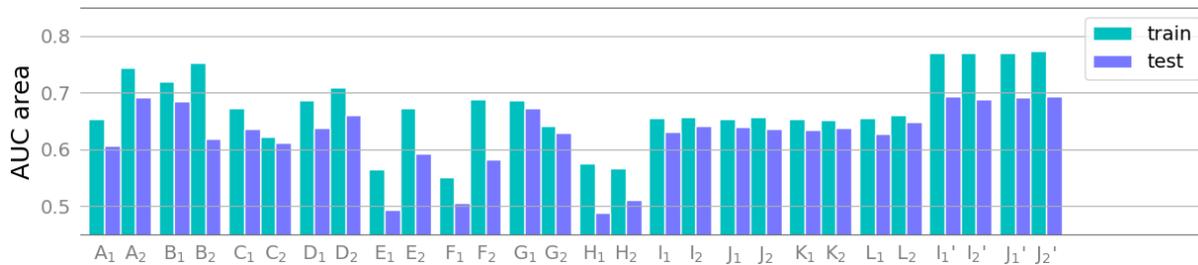


図 11 各モデルの評価 (AUC の面積)

ピックの特徴から推定し、その推定値及びツイート数の特徴から未来の Pleasure 値を予測することが、提案手法よりも有効であることが示された。

予測された Pleasure 値に対する *Threshold* の値は、実験結果の図 12, 13 での、「FP-rate と TN-rate の交点」から「TP-rate と FN-rate の交点」の間の 3.0 程度に設けることが適切だと考えられる。例として、テストデータにて *Threshold* = 2.9 とすると、全体の 29.2%にあたる「いいところ」のうち、63.2%の割合で「いいところ」だと予測できた。

7 関連研究

本研究では、野球の試合の「いいところ」を検出するため、実況ツイートを用いた手法を提案した。これまでも、実況コメントを用いて動画内の特徴的シーンを抽出する研究は様々な手法で行われている。

まず、コメント数の推移は、ユーザの興味や盛り上がりを表す指標として多くの研究で使われている。本研究でも、「ファンの喜び度合い」を表す要素の一部としてツイート数を扱った。中澤ら [2] は、テレビ番組に対しての実況ツイートのツイート数の変動から、重要シーンを自動的に検出している。この研究では、ツイート数のピークのみを重要シーンとしている点が、本研究とは異なる。コメント数以外からシーンの特徴を検出する研究として、山内ら [3] は、ニコニコ動画コメントから、感情極性 (PN) を用いて視聴者をクラスタリングし、「観点入れ替わり」している部分の特徴的シーンとみなし、ダイジェストを作成する手法を提案している。また、佃ら [4] は、映像検索や推薦することを目的に、ニコニコ動画コメントを用い、喜び、悲しみ、驚き、肯定、否定の感情辞書から登場人物の活躍パターンを分析し、動画の類似度を算出する手法を提案している。これら手法では、コメントに出現する一つの単語のごとの感情を辞書から割り当てているが、本研究ではトピック分布のパターンを扱っている点で異なる。牛島ら [5] は、LDA モデルを用いて、TV ドラマの実況ツイートから時間区間ごとのトピックを抽出し、その分布から、別のドラマとの類似度を計算する。本研究は、視聴者の反応のパターンを検出している点と同じだが、そこから全く別の指標である「ファンの喜びの度合い」を算出している点で異なる。

8 まとめ

本研究では、野球の試合の「いいところ」を目的とし、試合に対する実況ツイートから、ファンが喜ぶ「いいところ」を予測する分析を行った。結果として、Pleasure 値の推定式を利用した「いいところ」の予測の有効性は示せた。推定式の精度を上げることによって、Pleasure 値の予測モデルの精度も向上させる必要がある。

今後は、「いいところ」を予測したときにユーザに話しかける文章の生成等を検討している。

文 献

- [1] Yahoo Japan, “【無料アプリ】スポナビ プロ野球速報 - スポーツナビ”, <https://sports.yahoo.co.jp/promo/app/bblive>, 参照 2019-1-31.
- [2] 中澤昌美, 帆足啓一郎, 小野智弘, “Twitter によるテレビ番組重要シーン検出及びラベル付与手法”, 第 73 回全国大会講演論文集, 2011, 1, pp.517-519, 2011.
- [3] 山内嶺, 北山大輔, “ダイジェスト映像自動生成のための観点入れ替わりに基づいた特徴的シーン抽出”, 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2014), F4-2, 2014.
- [4] 佃洗撰, 山本岳洋, 中村聡史, 田中克己, “映像に付与されたコメントを用いた 登場人物が注目されるシーンの推定”, 情報処理学会論文誌 Vol. 52 No. 12 3471–3482 (Dec. 2011), 2011.
- [5] 牛島実桜, 南大智, 牛尼剛聡, “視聴者はドラマの「どこ」に「どう」反応しているのか? -実況ツイートを利用したドラマのシーン特徴の抽出-”, 第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2017), D6-1, 2017.
- [6] David M Blei, Andrew Y Ng, and Michael I Jordan, “Latent dirichlet allocation”, the Journal of machine, Learning research 3 pp.993-1022 2003.
- [7] mecab-ipadic-NEologd : Neologism dictionary for MeCab, <https://github.com/neologd/mecab-ipadic-neologd/>, 参照 2019-1-31.
- [8] G.Appel, “Technical Analysis: Power Tools for Active Investors”, FT Press, 2005.
- [9] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory”, Neural computation, Vol.9 No.8 pp.1735–1780, 1997.
- [10] Kingma, Diederik, and Jimmy Ba. “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980, 2014.
- [11] Akobeng AK, Understanding diagnostic tests 3: Receiver operating characteristic curves, Acta Paediatrica 2007 96 pp.644–647, 2007.