

スポーツジャンルに応ずる ネタバレ特性分析と判定手法の提案

白鳥 裕士[†] 中村 聡史[‡]

[†] [‡] 明治大学総合数理学部 〒164-8525 東京都中野区中野 4-21-1

E-mail: [†] ev30621@meiji.ac.jp, [‡] satoshi@snakamura.org

あらまし スポーツの試合を録画して視聴することを楽しみにしている視聴者にとって、SNS 上で遭遇してしまうスポーツの試合結果を示唆するネタバレ情報は楽しみを減退させる忌むべきものである。我々はこれまでの研究において、Twitter におけるサッカーのネタバレ情報を直接的なものと間接的なものに分類したデータセットを構築し、特徴的な単語やその時間的特徴について明らかにした。本稿では、スポーツの種類を拡張してネタバレに関するデータセットを新たに構築し、スポーツのジャンル別にその特性を分析することによって、試合展開とネタバレの数や内容が大きく関係していることを明らかにした。また、試合展開を考慮した手法を提案し、試合展開を考慮することで適合率を向上させることができることを明らかにした。

キーワード ネットバレ防止, 機械学習, スポーツ, Twitter

1. はじめに

スポーツは筋書きのないドラマであるため、勝つか負けるか分からないというハラハラ感や予想もしない展開に対する驚きを味わうために、リアルタイムでの視聴をしたいと考えている視聴者は少なくない。しかし、仕事や学業、その他の用事などでリアルタイムでの視聴ができないため、仕方なく録画予約をして時間の許すときに視聴しようとすることもある。こうした状況において、録画視聴を楽しみにしている視聴者が、視聴前にそのスポーツの試合結果を知ってしまうと、ハラハラ感や驚きが失われてしまう。こうしたハラハラ感や驚きを大事にしている視聴者にとって、試合結果はネタバレ情報と呼ばれており、多くの視聴者はこれを避けるため情報遮断を積極的に行っている。ここで、ウェブ上で視聴者がネタバレ情報に出会うのはニュースサイトやウェブログ、検索サイトなど様々であるが、特にネタバレとの遭遇機会が多いのが Twitter に代表される SNS である。Twitter では、アクセスするだけで友人の現在の状況を知り、気軽にコミュニケーションを取ることができる。そのため、何気なくアクセスする視聴者が多く、その際にネタバレ情報も受け取ってしまう。

SNS 上にあるネタバレは、SNS サイトや SNS のためのアプリケーションを使用しないことで完全に断つことが可能である。しかし、SNS はコミュニケーションに利用されているものであり、完全に遮断することはコミュニケーションまで遮断してしまうことにつながるため現実的ではない。

我々はこれまで、自身の都合によりリアルタイムでの視聴ができず、録画した試合をハラハラ感や驚きを持って楽しむために、ネタバレを防止したいと考える視聴者を対象に、スポーツの試合開始から視聴者の視

聴開始までの間ネタバレ情報を遮断するための提示手法を提案[1]し、Twitter クライアントなどの形で実装[2]してきた。また、ネタバレがどのようなものなのかを深く分析するために、ネタバレについて文章内の単語などから簡単にネタバレと識別できる直接的なネタバレと、文脈からネタバレが暗に指し示される間接的なネタバレがあると考え、直接的ネタバレと間接的なネタバレがどのように出現しどのような特性を持っているのかを、サッカーの試合に関するネタバレデータセットを構築[3]することで明らかにしてきた。しかし、以前の研究では対象とするスポーツがサッカーのみと少なかった。

そこで本稿では、まず対象とするスポーツの種類を増やし、スポーツのジャンルに応じたネタバレの特徴を分析する。次に、ネタバレはその性質から視聴者が自らチェックすることはできないが、コンピュータはネタバレを見てしまってもショックを受けないことに注目し、コンピュータが試合展開を常時把握し、その展開にあわせたネタバレ防止を行うことで視聴者がネタバレに出会わないですむ手法を提案する。

2. 関連研究

ネタバレ防止に関連した研究として、情報フィルタリングがある。

情報フィルタリングに関する研究において以前から多くされているのは、SPAM メールに代表されるコンテンツに対するフィルタリングである。しかし、これらのフィルタリング手法はほとんどの視聴者にとって興味のない有害なものを遮断するものであり、ネタバレのような視聴者の興味の対象を動的に遮断するものとは異なる。

岩井ら[4]は機械学習アルゴリズムを用いてレビュ

一文のあらすじを分類・発見し、非表示にして提示することで、ストーリーに関する記述を読みたくない視聴者でもレビューを閲覧できる手法を提案している。あらすじが直接ネタバレとなるわけではないため、本稿で取り組んでいる内容とは異なる。また、我々と同じくコンテンツの時間的な流れに注目した前田ら[5]の研究では、ストーリーコンテンツに対するレビュー文を視聴者が参考にする際に、ネタバレを発見してしまうことを問題とし、ストーリーコンテンツに対する短文形式のネタバレデータセットを構築し、ネタバレに関する単語がストーリーコンテンツ内にどのように分布しているかを調査しており、コンテンツ自体からネタバレを判断する方法について検討している。田島ら[6]はテレビアニメなどのストーリーコンテンツにおいて、放送時間差によって、SNS上でネタバレをされてしまうことを問題としており、致命的なネタバレとなる「生死」「勝敗」などのトピックに対し、機械学習と独自の手法を組み合わせることでネタバレを判定する可能性について明らかにしている。対して我々は、ストーリーコンテンツではなく、決まったストーリー展開のないスポーツを対象コンテンツとしている。また、コンテンツではなくコンテンツに対するツイートからネタバレデータセットを構築し、ネタバレ判定を行うという点でアプローチが異なる。

一方、Golbeck[7]はテレビ番組に対して、放送時差による Twitter 上のネタバレコメントを判定する手法を提案している。Golbeck は対象コンテンツに対するあらゆるコメントを遮断することによって、再現率 100%を目指しているが、我々は SNS 上のコミュニケーションを可能な限り遮断せずに、ネタバレ情報のみを遮断することを目的としている。

田中ら[8]はニコニコ動画などのコメント機能付動画共有サービスにおいてネタバレとなるコメントを検知する手法を提案している。ニコニコ動画の性質上、そもそも視聴のタイミングは視聴者によって異なるものであり、SNS のようなリアルタイムに起きている事柄へのコメントではないため、対象が異なる。

3. ネットバレーデータセット構築

ここでは、スポーツの試合に対する Twitter 上の投稿について、ネタバレの正解データを構築する。以降この Twitter 上の投稿をツイートと呼ぶ。

3.1 収集対象と方法

まず、スポーツの試合に対応するツイートを収集する。以前我々はサッカー 3 試合についてツイートを収集し、ネタバレのデータセットを構築、分析までを行ったが、収集試合数が少なかったために、判定手法を吟味し精度を検証するまでに至らなかった。そこで、

本稿では収集するサッカーの試合数を 3 試合から拡張するとともに、興味度合いが高い人が多く（事前にアンケートを実施）、かつ 2015 年に比較的話題となった、バレーボールと野球についても新たに収集した。ここでは特にツイートが多く集まる日本代表の試合に注目し、表 1 に示す 18 試合を対象とした。

表 1 収集した試合

試合名	対戦チーム	勝敗
2018FIFA ワールドカップロシア アジア 2 次予選	アフガニスタン	勝利
	カンボジア	勝利
	シンガポール	勝利
国際親善試合	イラン	引き分け
2015EAFF 東アジアカップ 決勝大会	中国	引き分け
	韓国	引き分け
2015EAFF 女子 東アジアカップ 決勝大会	中国	勝利
2015FIFA 女子 ワールドカップカナダ 決勝トーナメント	イングランド	勝利
	アメリカ	敗北
2015FIVB ワールドカップ 男子大会	オーストラリア	勝利
	ロシア	敗北
	アメリカ	敗北
2015FIVB ワールドカップ 女子大会	中国	敗北
	セルビア	敗北
	アメリカ	敗北
世界野球 WBSC プレミア 12	アメリカ	勝利
	ベネズエラ	勝利
	韓国	敗北

ここで、リアルタイムなコンテンツに対してツイートする場合、ハッシュタグと呼ばれる検索・分類のためのテキストが付与される事がある。例えば、サッカーの日本代表の試合においては、「#daihyo」や「#JPN」などのハッシュタグが用いられている。ネタバレ対象のスポーツの試合に関するツイートに、対象スポーツに関するハッシュタグが付与されている場合は、そのハッシュタグを含むツイートをすべて遮断するだけでよい。しかし、実際にその試合に関連しているのにハッシュタグ無しでツイートされているものは多い。

あるスポーツの試合に対するツイートを全て集める場合、その時間にツイートされている全てのツイートを集め、そこからスポーツの試合に関するものを選別する必要がある。しかし、選別における精度の問題が生じるうえ、ツイートの内容を友人関係にしか開示していない視聴者のツイートを集めることは出来ない。また、Twitter 社が提供している Stream API では全てのツイートを収集することは出来ない。

そこで本稿では、あるスポーツの試合に対して投稿されているハッシュタグ付きのツイートは、その試合

に対して投稿されているすべてのツイートを代表していると仮定し、リアルタイムのツイートを見ることが出来る Yahoo! JAPAN の「Yahoo!リアルタイム検索」を用い、サッカーの日本代表の試合であれば、「#daihyo」や「#JPN」、バレーボールの日本代表の試合であれば、「#火の鳥 NIPPON」や「#龍神 NIPPON」といったような、以前に同様の試合で用いられているハッシュタグを試合開始前に抽出し選定、そのハッシュタグを含むツイートを Twitter Search API を用いて収集した。試合を楽しみにしている視聴者は、試合開始前や以前行われた同様の試合にも特定のハッシュタグを用いて気持ちを発信することが多く、この方法で十分ハッシュタグを特定できる。なお、収集においては、試合開始から試合終了後 2 時間までの間ツイートを収集した。

3.2 データの整形

収集したデータの中には分類および分析において適切でないツイートも多い。そのため、下記の手順で不適切なツイートの除去およびツイートの整形を行う。

1. ワールドカップの試合などでは対戦相手国からのツイートも多数登場し、多言語となる。ここではデータセット構築者が日本人であることを考慮し、日本語以外のツイートを除去する。なお、日本語以外のツイートの除去については、ツイート取得時にあらかじめ言語コードを取得し、言語コードが「ja」かそうでないかによって判断する。
2. 先頭に「RT」を含むツイートは、Twitter のリツイート機能と呼ばれる他の視聴者のツイートをそのままの形でツイートできる機能で、他の視聴者のツイートを自分のツイートを見ている人に対して発信できるものである。これは元々のツイートと内容が重複するものであるため、正規表現により除去する。
3. 収集したツイートからハッシュタグの除去を行う。ハッシュタグの除去については、「#」という文字から連続した空白・改行以外の文字までを正規表現により判定する。中身の無いツイートについても正規表現により削除する（ハッシュタグのみのツイートがあるため）。ここで先頭から最後までの間、空白および改行のみであった場合、そのツイートを除去した。
4. 無関係のスパムツイートを削除する。この作業については、スポーツのハッシュタグに対して URL を投稿しているツイートにスパムツイートが多かったため、「http://t.co/」もしくは「https://t.co/」を含むツイートを正規表現により判定し除去した。

3.3 データの分類

収集したツイートがネタバレなのか、それともネタバレではないのかということを知るため、人手による

分類を行った。

ここでは、先述の 18 試合のツイートからそれぞれ 1000 件ずつ無作為に抽出したツイートを対象とした。1000 件に限定したのは、各データセットのデータ量を揃えるためである。なお、分類作業を行ってもらうために、図 1 に示すウェブシステムを開発した。このシステムでは、視聴者は最初にアカウント名を入力してログインし、ページ上に提示されているツイートがネタバレだと感じたらクリックしチェックするというものである。ここで、抽出したものを 1 試合ずつ提示してしまうと、試合の詳細情報を知っている者が分類をする場合に、1 つのツイートから試合の前後関係や内容が鮮明に伝わってしまう恐れがある。例えば、香川選手が得点を決めた試合を分類してもらうと仮定して、「2018W 杯予選日本 vs シンガポール戦開始!」のようなツイートをその試合の結果を知っている者が見てしまうと、「かがわー」といった、試合の詳細を知らない者はネタバレとは言いがたいツイートまでネタバレと判断してしまう可能性がある。これは、実際に録画視聴を楽しみにしている視聴者を想定した場合、この視聴者は試合の事前情報を知ることであっても、試合内容までは知ることにはないはずであるので好ましくない。そこで、18 試合を 3 試合ずつまとめて、ランダムに提示することとした。つまり、合計 3000 件のツイートが 6 グループできることになる。さらに、視聴者は、試合開始からのある程度の経過時間については、視聴していなくとも把握することができる。経過時間が影響してしまう例として、「同点」というツイートがあったとすると、これが試合前半でのものなのか、試合後半でのものなのかにより、試合結果への影響の大きさが異なり、ネタバレの判断に必要であると考えた。そこで、ツイートは時間順ではなくランダムに提示され、各ツイートの下には、そのツイートがされた試合開始からのおおよその経過時間を表示した。例えば、ツイートの下に「0」と表示されていれば、試合開始 0 分から 10 分までのツイートであり、「60」と表示されていたら試合開始 51 分から 70 分までのツイートとなっている。なお、おおよその時間にした理由については、実際に録画視聴をしている視聴者を想定した場合に、視聴者は細かな時間までは知り得ないという点にある。

ウェブシステムでは 1 ページごとに 50 件ずつツイートが表示され、1 グループあたり 60 ページの分量となっている。図 2 に示す通りクリックすると背景色が赤に変わり、選択済みであることを示す。



図 1 開発したウェブシステム

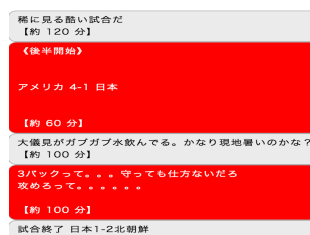


図 2 開発したウェブシステム（拡大）

データセット構築者には，作成したウェブサイトにもアクセスしてもらい，ツイートの分類を行ってもらった．なお，ツイートに対する分類結果は 1 ページごとにデータベースに記録されるため，途中で中断して再開することも可能となっている．データセット構築者は，1 グループにつき 5 人で，当該スポーツについてよく視聴していて，SNS でネタバレを受けたことがあり，Twitter を普段から用いている 19 歳から 22 歳の大学生とした．

4. ネタバレツイートの分析

表 2 データセット構築結果

グループ	件数	例	経過時間
男子サッカー	1085	おおお！ 香川が ゴール！！！！	20
女子サッカー	566	既に 2 失点 (´Д｀)	0
男子バレー	651	勝って 欲しかったな。	160
女子バレー	637	おお 1 セット 取ったー！	40
野球	633	日本 1 点追加、 筒香の タイムリー	140

3 章で構築したデータセットから，半数以上，つまり 3 人以上がネタバレと判定したものをネタバレツイートとし，正解データとした．半数以上とした理由は，以前の我々の研究[3]で明らかにしている．データセッ

ト構築の結果得られた，各グループのネタバレツイート件数とネタバレツイートの例を表 2 に示す．経過時間は，例に挙げたネタバレツイートがされた，およそその試合開始からの経過時間を表す．

分類した結果，どのスポーツも共通して，「○-○」「勝つ」「○点」など，特定のパターン記述や単語が含まれているものが多かった．サッカーでは「ゴール」，バレーボールでは「セット」，野球では，「ホームラン」「ヒット」「タイムリー」というように，スポーツごとに異なるパターン記述や単語による表現がされていたが，本質的には，試合展開や結果など同じような内容であった．ネタバレ件数は全体の 5 分の 1 ほどで，試合に対するツイートのうち致命的なネタバレがこれに当たることになる．

表 3 グループ別 TF 値上位 3 件

	単語	TF 値	単語	TF 値	単語	TF 値
	名詞		形容詞		動詞	
男子サッカー	点	0.050	良い	0.443	勝つ	0.051
	香川	0.030	無い	0.131	決める	0.049
	前半	0.020	欲しい	0.095	する	0.043
女子サッカー	日本	0.054	良い	0.181	見る	0.048
	なでしこ	0.054	強い	0.109	勝つ	0.046
	点	0.052	悔しい	0.086	返す	0.025
男子バレー	ロシア	0.026	凄い	0.164	取る	0.112
	選手	0.021	良い	0.133	頑張る	0.095
	男子	0.015	欲しい	0.090	勝つ	0.055
女子バレー	セット	0.073	良い	0.192	頑張る	0.094
	日本	0.061	悔しい	0.095	取る	0.076
	目	0.025	凄い	0.092	負ける	0.062
野球	日本	0.032	良い	0.283	勝つ	0.061
	回	0.024	凄い	0.141	打つ	0.041
	逆転	0.018	欲しい	0.056	取る	0.021

具体的に，ネタバレと判定されたツイートにおいて，どのような単語が多く含まれているのかを調べるため，競技ごとに単語の出現頻度を計算した．なお，単語には形態素解析エンジンである MeCab を用い，名詞，形容詞，動詞に絞って算出した．TF 値の上位 3 件ずつを

表 3 に示す。

まず名詞においては、どのスポーツでも国名や選手名、「点」「セット」などの試合の状況を直接表すような単語の TF 値が高かった。スポーツごとに特徴的な点としては、サッカーと野球で「前半」「後半」、「回」といった試合の経過を表す語が高くなっていた点と、サッカー、バレーボール、野球でそれぞれ「ゴール」、「セット」、「タイムリー」と試合状況の表現の仕方が異なる点があった。サッカーの「ゴール」という単語は、目で直接確認した時は多いと感じたが、TF 値を調べると思っていたよりも低い結果となった。これは、「ゴーーーーーール」のように母音を伸ばして感情の高ぶりを表現するようなツイートが多く存在しており、単語分けをする時に正しく判定されなかったからだと考えられる。

次に形容詞においては、どのスポーツも共通して、「良い」「悪い」といった評価を表す語や、「凄い」「悔しい」といった感動や感情を表現する語の TF 値が高かった。形容詞の上位単語の TF 値が名詞、動詞と比較して高い結果となっているが、ツイート全体を通して形容詞の種類と数が少ないのが原因となっている。

動詞においては、「頑張れ」のような応援を表す語や、「勝つ」「負ける」といった試合の結果を表す語が高くなっていた。動詞も名詞と同じく、サッカーでは「決める」、バレーボールでは「取る」、野球では「打つ」「取る」など試合中の選手の行動の表現の仕方にスポーツごとの特徴が出ていた。さらにツイートされた時点での勝敗ごとにみると、勝っているときは「勝つ」「逆転」、負けているときは「負ける」「悔しい」などが多く、試合展開でツイート内容が大きく異なっていることがわかった。

スポーツごとに試合経過に対するネタバレ数に違いがあるのかを調べるため、試合開始から 1 分ごとのネタバレ数の変化の様子をグラフ化した。スポーツごとに図 3～8 に示す。男子サッカーは数が多いため、2 つに分けた。グラフの横軸は試合開始からの経過時間を表し、縦軸はその時間でのネタバレ件数を表す。

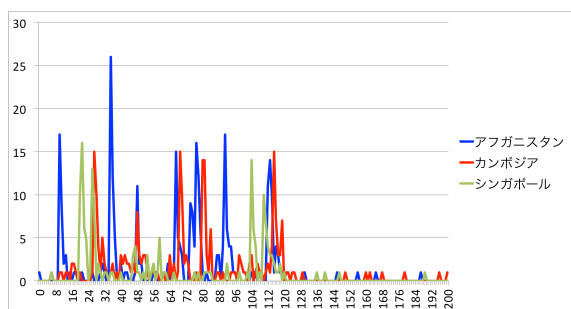


図 3 スポーツごとのネタバレ数の時間変化
男子サッカー①

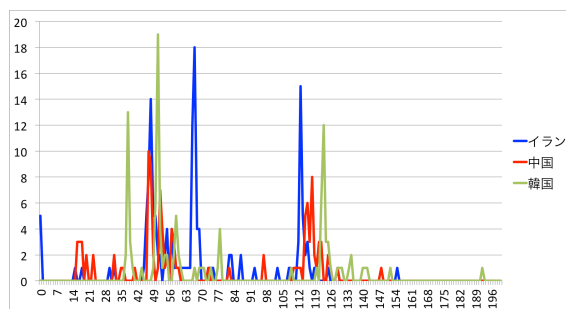


図 4 スポーツごとのネタバレ数の時間変化
男子サッカー②

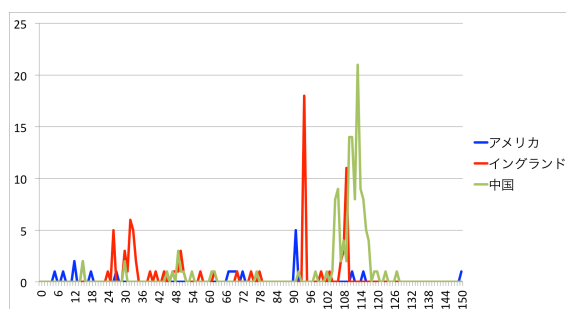


図 5 スポーツごとのネタバレ数の時間変化
女子サッカー

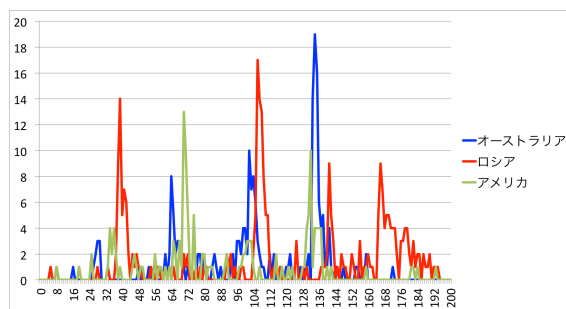


図 6 スポーツごとのネタバレ数の時間変化
男子バレー

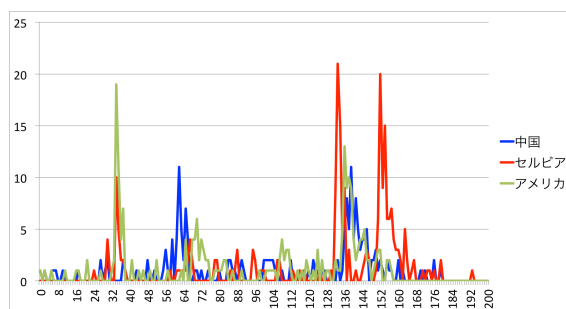


図 7 スポーツごとのネタバレ数の時間変化
女子バレー

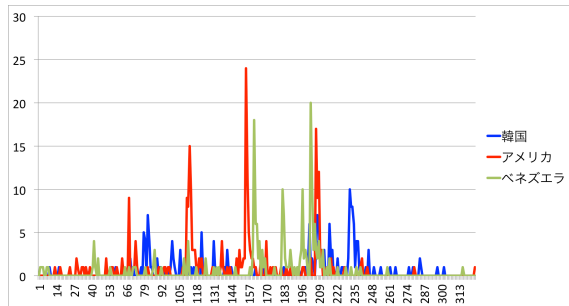


図 8 スポーツごとのネタバレ数の時間変化
野球

図 3～8 から、すべてのスポーツで動きが激しく特徴が出ている。まず図 8 の野球のグラフは、他のスポーツに比べ上下動が激しく、山の数が多くなっている。これはネタバレを引き起こすような得点シーンが野球では起こりやすいためだと考えられる。図 6, 7 のバレーボールのグラフは、定期的に上下しており、試合ごとに山の数異なっている。これは、野球よりも得点シーンは多いものの、得点ごとのネタバレツイートは少なく、セットを取ったタイミングでネタバレが多く起こることを表している。また、試合ごとに山の数異なるのは、取得したセット数が試合ごとに異なるからだと考えられる。さらに、ネタバレ件数が増えるタイミングは試合間で同期しており、セットを取りネタバレが多く起きやすいタイミングが比較的同じ時間帯であることを表している。

一方、図 1～3 のサッカーのグラフは、形として各試合間で共通している部分は少ない。これは、サッカーの点の動きが比較的少なく、かつ不定期に起こる性質にあると考えられる。サッカーでは前後半の開始から得点が決まるまでの間、野球では試合開始から得点が決まるまでの間はネタバレがほとんどなく、ネタバレが得点に対してされているものであることがわかる。それに対しバレーボールでは、比較的セットの取得以外の時間帯でのネタバレが多く、これはセット内の得点に対してもネタバレを感じる人が存在することを示している。試合展開と重ねると、それぞれで急激にネタバレ件数が増加した時間帯は、サッカーは点が決まった時と試合が終了した時、バレーボールはセットを取った時と試合が終了した時、野球は点が動いた時と試合が終了した時であった。これは、点やセットといったように競技ごとにツイートの表現に違いはあるが、すべての試合に共通しており、試合の動きがあった時（バレーボールの場合は日本代表にプラスの動きがあった時）と試合が終了した時にネタバレ件数が伸びている結果となり、ネタバレ件数と試合展開には関係性があることがわかった。なお、サッカーと野球では応援しているチームにマイナスの試合の動きがあった時

もネタバレ件数が伸びていたが、プラスの試合の動きがあった時よりも早くネタバレ件数が減少しており、これは気持ちの萎えが原因だと考えられる。

5. 提案手法

4 章で述べたように、スポーツの試合に対応するツイートは試合の展開によって内容が大きく異なる。そこで、過去のツイートを学習する際に、リードしている時間帯では「勝ちモデル」、リードされている時間帯は「負けモデル」など試合展開ごとに学習モデルを作成しておき、試合展開に応じて利用するモデルを切り替えながらネタバレを判定する。この手法により精度を上げることができると期待される。

提案手法をシステムとして実装する際には、試合展開をリアルタイム速報のウェブページを常時監視することで取得し、それに応じて使用するモデルを決定したうえで、判定対象の試合に適用する。例えば、あるサッカーの試合に対してネタバレ判定を行うときには、システムがリアルタイム速報をリアルタイムで監視し、得点が動く度に、勝っている状況なのか、負けている状況なのか、同点の状況なのかを判断し、適切なモデルを用いる。つまり、視聴者に届くツイートに多少の遅延をもたせ、システムがネタバレのチェックを行い遮断することでネタバレ防止を実現する。

6. 評価実験

提案手法の有効性を確かめるために、提案手法と比較する手法を 2 つ用意する。

- **パターンマッチ手法**: 4 章でネタバレとして出現頻度の高かった単語をキーワードとし、キーワードにマッチする単語を含むツイートをネタバレと判定する。
- **ベースライン手法**: SVM による機械学習において、試合展開などは考慮せず、試合全体で 1 つの学習モデルを作成する。

ベースライン手法ないし提案手法では、構築したデータセットのうち判定対象となる試合以外の試合のツイートをを用いて SVM のモデルを作成し、そのモデルを用いて判定対象となる試合のツイートを判定する。なお、ベースライン手法では、判定対象となる試合以外の試合のツイート全体で 1 つの学習モデル、提案手法では、判定対象となる試合以外の試合のツイートのうち、リードしている時間帯のツイートでは「勝ちモデル」、リードされている時間帯では「負けモデル」、同点の時間帯では「引き分けモデル」と 3 つの学習モデルを作成する。例えば、提案手法で男子サッカーの試合「日本 vs アフガニスタン」を判定するときは、「日

本 vs カンボジア」「日本 vs シンガポール」「日本 vs イラン」「日本 vs 中国」「日本 vs 韓国」5つの試合のツイートを用いて「勝ちモデル」「負けモデル」「引き分けモデル」の3つのモデルを作成し、「日本 vs アフガニスタン」の1000件のツイートに対し、勝っていれば「勝ちモデル」を、負けていれば「負けモデル」を、同点であれば「引き分けモデル」を適用し、判定する。勝敗についてはサッカー、野球では得点によるスコア、バレーボールは取得セットによるスコアの進行状況とする。モデルの作成にあたっては、ネタバレツイートよりもネタバレでないツイートの数が多くなってしまったため、アンダーサンプリングを行う。以下では、SVMのモデルはアンダーサンプリングして作成したものとする。SVMのための単語ベクトルの生成については、MeCabを用いた形態素解析を行う。なお、単語ベクトルは名詞、動詞、形容詞、連体詞、副詞の5つの品詞を利用して生成する。

パターンマッチ手法、ベースライン手法、提案手法を適合率、再現率で比較する。なお、「1-0」などのスコア情報は確実にネタバレであることがわかるため、スコア情報を含むツイートは事前に正規表現により取り除いた。

7. 実験結果

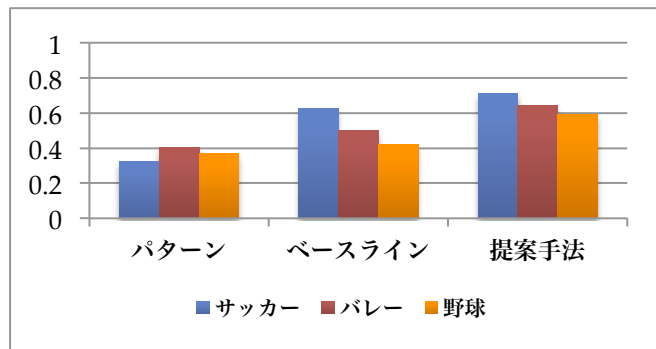


図9 手法ごとのネタバレ判定適合率

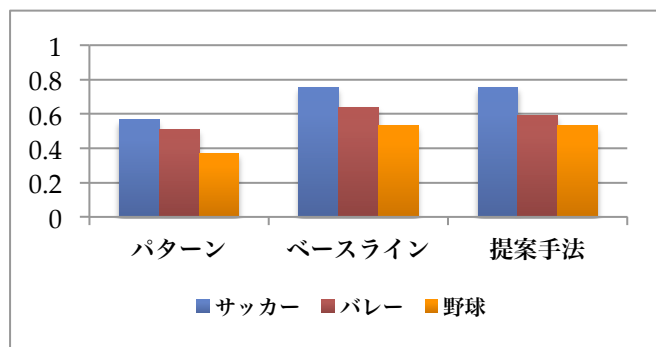


図10 手法ごとのネタバレ判定再現率

パターンマッチ手法、ベースライン手法、提案手法の適合率、再現率を比較した結果を図9、図10に示す。「サッカー」は男子サッカーと女子サッカーの平均、「バレー」は男子バレーと女子バレーの平均を表す。

実験の結果、いずれのスポーツにおいても適合率、再現率ともにパターンマッチ手法に比べ、ベースライン手法および提案手法の方が有効であった。適合率は、他の手法と比較し、提案手法がすべてのスポーツで高くなっており、試合展開ごとで分けたことによって、試合展開とは関係のないツイートを無視することができ、学習モデルのノイズが軽減されたと言える。「バレーボール」「野球」が「サッカー」よりも適合率が低かったのは、セットによる勝敗が「負け」であってもセット内の途中の得点では勝っていることもあり、勝敗ごとの特徴があまり出ないこと、同じヒットでも得点につながるヒットのみを人はネタバレだと感じるが、システムは得点につながらないヒットもネタバレだと判断してしまうことなどが原因だと考えられる。

提案手法でも適合率がそこまで大きくならなかったのは、対象の試合に対する「勝つぞ」といったツイートがネタバレと判定されてしまっていることが原因であると考えられる。この問題については、「勝つぞ」など現在形のときはネタバレでないとし、「勝った」など過去形のときはネタバレとすることが考えられるが、「次は勝つぞ」などは現在形で書かれているが、未来へ向けた発言であり、間接的に今回の試合結果が伝わってしまう。従って、勝敗の試合展開ごとのモデル作成に加え、試合の前半と後半などの試合経過によるモデル作成（試合前半の「勝つぞ」は対象試合に対してなのでネタバレではなく、試合終了後の「勝つぞ」は未来の試合に対してなのでネタバレ）を行うことや、CaboChaを用いて係り受け解析を行い、「次」が「勝つ」に掛っていたらネタバレとすることなどを考えている。

今回の実験では、再現率はベースライン手法と提案手法で変化はなかった。これは実験に用いたデータ量が少なく、試合展開特有のネタバレがあまり存在しなかったことが原因であると考えられる。

また、試合展開ごとのネタバレ判定精度も調査したところ、勝っている展開の適合率が高く、負けている展開では全体的にやや適合率が低かった。これは、勝っている展開では提案手法が有効であることを示している。また、負けている展開についても、負けている展開の多かったバレーボールでは適合率が高くなっており、学習するデータの数を増やせば、他のスポーツも高くなると考えられる。さらに、引き分けている展開の適合率が突出して低かった。これは、同じ引き分けでも「0-0」なのか、追いついたのか、追いつかれたのかによってツイート内容が大きく変化するからだ

8. 考察

9. おわりに

- [1] 中村聡史，小松孝徳：スポーツの勝敗にまつわるネタバレ防止手法の検討，情報学会論文誌，Vol.54, No.4, pp.1402-1412（2013）．
- [2] 中村聡史，川連一将：スポーツのネタバレを防止する Twitter クライアントの開発と諸検討，ARG WI2 No.4（2014）．
- [3] 白鳥裕士，中村聡史：SNS 上でのサッカーの試合に対する直接的・間接的ネタバレの分析，研究報告グループウェアとネットワークサービス（GN），2015-GN-96, Vol.8, pp.1-8（2015）．
- [4] 岩井秀成，池田郁，土方嘉徳，西田正吾：レビュー文を対象としたあらすじ分類手法の提案，電子情報通信学会論文誌 D, J96-D, No.5, pp.1222-1234（2013）．
- [5] 前田恭佑，土方嘉徳，中村聡史：ストーリー文書内のネタバレの記述に関する調査とレビュー文書でのネタバレ検出の試み，第 8 回 Web とデータベースに関するフォーラム論文集，pp.32-39（2015）．
- [6] 田島一樹，中村聡史：ストーリーコンテンツに対するネタバレの基礎調査とその判定手法の検討，研究報告グループウェアとネットワークサービス（GN），2015-GN-96, Vol.7, pp.1-6（2015）．
- [7] Golbeck, J.: The Twitter Mute Button: A web Filtering Challenge, Proc.2012 ACM Annual Conference on Human Factors in Computing Systems, pp.2755-2758（2012）．
- [8] 田中駿，廣田壮一郎，高村大也：コメント機能付動画共有サービスにおけるネタバレ検知，第 29 回人工知能学会全国大会，査読なし（2015）．
- [9] Brody, S. and Diakopoulos, N.: Cooooooooooooooooo!!!!!!!!!!!!!!!!!!!!!! Using word lengthening to detect sentiment in microblogs, Proc.Conference on Empirical Methods in Natural Language Processing, pp.562-570（2011）．