ツイート発信タイプと感情表現語の日内変動

養老 毅暁　　久野 雅樹
電気通信大学大学院 情報理工学研究科 総合情報学専攻
yourou@uec.ac.jp, hisano@hc.uec.ac.jp

1 はじめに

近年、ソーシャルネットワーキングサービス（SNS）上でのコミュニケーションが盛んになってきた。中でもマイクロブログである Twitter[*1] は多くの人に利用されており、その投稿内容も様々である。Twitter では、投稿のことをツイートと呼び、利用者をユーザと呼ぶ。ユーザは自身のツイートの一般公開／非公開（プロテクト）を設定することができる。一般公開のツイートは、Twitter のアカウントを持たない人でもインターネット上で閲覧することができる。

現時点では、Twitter は投稿可能な文字数が 140 字に制限されている。比較的長い文章を時間をかけて投稿する傾向があるウェブログよりも、気軽に投稿することができると考えられる。このことから、Twitter はユーザの状況をよりリアルタイムに反映するといえる。

本研究は、先の研究をするための準備段階として、Twitter の利用状況を押さえておきたいという動機から始まった。感情表現に着目するなど、多くの発展的な関連研究はあるものの、分析対象が全体像の中でどのような分布にあるのかという、全体像に対する基礎研究は十分ではないと感じている。まずは Twitter の全体的な日内変動を調べ、単純な分類方法で特徴を調べることが本研究の第 1 の目的である。全体的な日内変動の影響が、先の研究として考えている感情表現の分析にどのような影響をもたらすかを知ることが、第 2 の目的である。

2 関連研究

Twitter を対象として時系列変化に注目した研究は 2 種類に分けることができる。特定の 1 ユーザに関しての時系列変化を調べる研究と、ユーザを限定せずに複数ユーザのコミュニティ全体に対しての時系列変化を調べる研究である。前者については、ユーザが抱いている感情の変化を抽出するもの [1] や、ツイートの履歴から性格・プロフィールなどの推定 [2] を行うものがあり、それらの精度の向上を目的とする研究が多い。後者については、トピックを限定して世の中のユーザが抱いているトピックに対する印象を調べるもの [3][4] が多い。Yahoo! JAPAN が提供する「リアルタイム検索」[*2] はその好例である。

本研究は後者にあたるが、トピックを限定せず、一般公開されている全てのツイートに対して、そのサンプリング調査を行ったものである。これにより、Twitter を一般公開で利用しているユーザ全体の傾向・偏りを大きく把握することができる。本論文で報告すること以外にも、ユーザ全体の大規模サンプリングデータを対象とした様々な属性情報についての調査も行っている。

3 分析対象コーパスの構築

3.1 ツイートの収集

Twitter 社が提供する Twitter Streaming API には、全公開ツイートの約 1%を収集できる sample という機能がある。sample は全世界の公開ツイートについて、ツイート本文を含む様々な属性情報を集めることができる。本研究では、それを利用するための Java ラッパである Twitter4J（Twitter 非公式）[*3] を利用して、一般公開されているツイートを収集した。ここでは、ユーザの情報は User、ツイートの情報は Status として提供されている。このうち、ユーザの言語設定（User の lang 属性）とツイートの言語設定（Status の lang 属性）がともに日本語であるものを収集することとした。収集は 2015 年 9 月 1 日に開始し、現在も続けている。1 日には約 70 万件以上が集められている。

今回の分析は、更にツイート本文に日本語の文字が含まれているかを判定するフィルターをかけ、設定情報は日本語であっても、外国語のみで書かれたものや顔文字のみのツイートは除外し、1 日平均 68 万件分を対象とした。その他の属性情報として、本研究ではツイート本文（Status の text 属性）以外に、公式リツイートであるか、リプライであるかどうか、投稿の時刻の情報を使用する。

[*1] http://www.twitter.com/
[*2] http://search.yahoo.co.jp/realtime
[*3] http://twitter4j.org/ja/index.html

収集開始の 2015 年 9 月 1 日から 2015 年 12 月 31 日までの 4ヶ月間で、何らかの不具合（Twitter Streaming API のエラーなど）で 1 日中のツイートが集められなかった日を除き、祝日を除く各曜日 2 日ずつ計 14 日をランダムに選出し、分析対象日とした。ツイート件数は計 9,583,979 件である。

3.2 発信タイプの分類とツイート頻度の日内変動

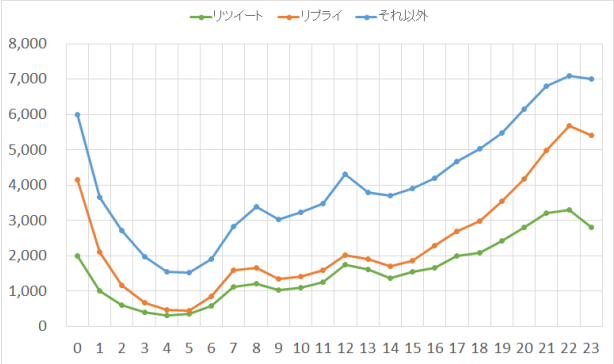


図 1: ツイート発信タイプ別 ツイート頻度の日内変動

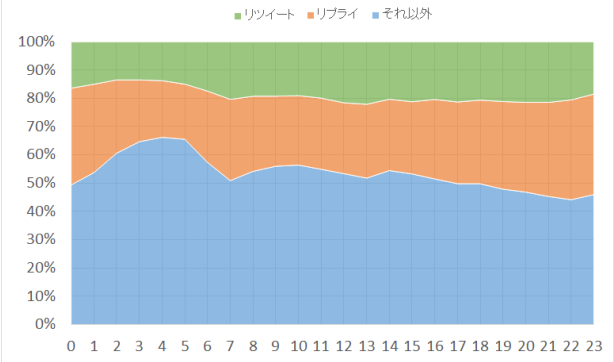


図 2: ツイート発信タイプ別 全体に対する割合の日内変動

図 1、図 2 は、全ツイートを公式リツイート・公式リプライ・それ以外の 3 つの発信タイプに分け、そのツイート総数の時系列変化と全体に占める割合を表したものである。計算の負荷を減らすため、50 分の 1 にあたる計 191,680 件のデータで作図をした（新たにツイートされた全公開日本語ツイートの約 0.02%）。各発信タイプのツイートの全体に対する割合の 1 日平均は、リツイートが約 20%、リプライが約 29%、それ以外が約 51%となった（図 3 も参照）。

図 1 によると、ツイート総数については、まず、全 3 タイプのツイートとも 22 時台が最大で、4 時・5 時台には最小となることが分かる。また、9 時台と 12 時台にピークがあることが伺える。図 2 によると、全体に占める割合については、まず、7 時台から 22 時台までは、ほぼ一定の割合を保ちながら、徐々にリプライが増えているとみることができる。一方で、深夜 0 時ごろから翌 4 時にかけてはリプライが減り、リプライでもリツイートでもないツイートの割合が増えている。5 時台・6 時台にかけて、再び昼間の一定割合に戻っていく。リツイートについては深夜帯には少し割合が減っているが、1 日を通してほとんど変わらないようである。

深夜帯は、ツイート総数自体が減っているとおり、実際に Twitter を使用中のユーザが減っているものと考えられる。それでも一定量のツイートが新たに投稿されているのは、実際に集めたツイートを確認したところ、自動ツイート機能である bot が影響しているようである。

4 感情表現語を含むツイート

4.1 感情表現語の抽出

ツイート本文に次ページの表 1 に示す 14 の感情表現語が含まれている個数を、スクリプト言語 Ruby の正規表現を用いてカウントした。

感情表現語はそれぞれの派生語とその活用形にマッチするものを数えることとした。例えば、形容詞「楽しい」については、名詞「楽しさ」「楽しみ」、動詞「楽しむ」「楽しぶ」「楽しがる」の全ての活用形、可能動詞「楽しめる」「楽しべる」「楽しがれる」の全ての活用形、「楽しな」「楽しそう」を対象とした。ただし、「楽しくない」「楽しめなかった」といった否定の表現もカウントしている点に注意しておきたい。同じ感情表現には「たのしい」「タノシイ」など、漢字以外の表現方法もあるが、「泣く」を対象としたため、平仮名・片仮名での表現は含まないこととした。

上記のルールで数え上げた感情表現語を 1 つ以上含むツイートと 1 つも含まないツイートとで、各発信タイプの全体に対する割合を表したものが図 3 である。感情表現語を 1 つ以上含むツイートは計 191,680 件中 13,621 件見つかった。これは、全サンプルの約 7%にあたる。

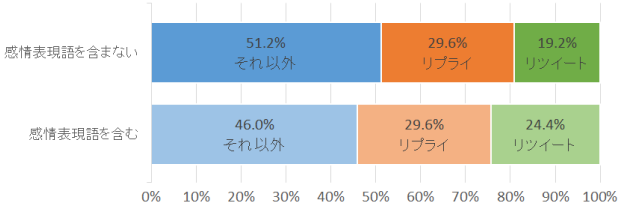


図 3: 発信タイプ別 感情表現語を含む／含まないツイートの全体に対する割合

図 3 から分かることは、感情表現語を含まないツイートに比べて、感情表現語はリツイートに使われることが多く、その分はリプライには影響せずに、リプ

表 1: 感情表現語の個数とその割合

語	全体			リプライ			リツイート			それ以外		
	頻度	%	順位	頻度	%	順位	頻度	%	順位	頻度	%	順位
楽しい	37,953	24.4	1	11,203	25.7	1 →	11,228	28.2	1 →	15,522	21.5	1 →
頑張る	20,172	13.0	2	8,610	19.8	2 →	3,939	9.9	4 ↓	7,623	10.6	3 ↓
嬉しい	16,201	10.4	3	6,785	15.6	3 →	3,964	9.9	3 →	5,452	7.5	6 ↓
笑う	15,467	9.9	4	3,759	8.6	4 →	4,767	12.0	2 ↑	6,941	9.6	4 ↓
死ぬ	14,397	9.2	5	2,633	6.0	5 →	3,020	7.6	6 ↓	8,744	12.1	2 ↑
泣く	10,742	6.9	6	2,118	4.9	6 →	3,052	7.7	5 ↑	5,572	7.7	5 ↑
面白い	9,486	6.1	7	1,944	4.5	8 ↓	2,308	5.8	7 →	5,234	7.2	8 ↓
痛い	7,739	5.0	8	1,343	3.1	9 ↓	1,035	2.6	11 ↓	5,361	7.4	7 ↑
怖い	7,366	4.7	9	2,034	4.7	7 ↑	1,532	3.8	8 ↑	3,800	5.3	9 →
怒る	4,402	2.8	10	954	2.2	10 →	1,314	3.3	9 ↑	2,134	3.0	10 →
悲しい	3,539	2.3	11	684	1.6	12 ↓	899	2.3	12 ↓	1,956	2.7	11 →
喜ぶ	3,532	2.3	12	785	1.8	11 ↑	1,187	3.0	10 ↑	1,560	2.2	12 →
驚く	2,415	1.6	13	428	1.0	13 →	886	2.2	13 →	1,101	1.5	14 ↓
苦しい	2,251	1.4	14	268	0.6	14 →	743	1.9	14 →	1,240	1.7	13 ↑
合計	155,662	100.0		43,548	100.0		39,874	100.0		72,240	100.0	

※ 計 14 日のサンプルから感情語を含むツイートを取り出し、その 5 分の 1 のデータから作成

ライ・リツイート以外の割合が減っているということである。このことについて、リツイートをする目的を考えると、その目的は、リツイートしようとするツイートに対して、共感していることを知らせたり、そのツイートを広めようとすることであるから、感情表現語が使われることが多いと考えられる。

4.2 発信タイプ別にみた感情表現語

図 4 は、発信タイプごとに各感情表現語がそれぞれの全体に対する割合を示したものである。語の順は、分析対象全体でその語を含むツイートの総数が多い順である（表 1 参照）。

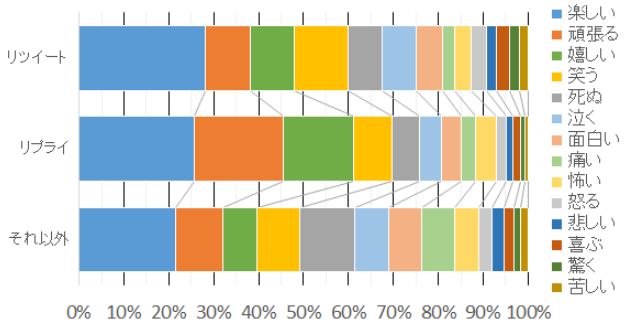


図 4: 発信タイプ別 各感情表現語が全体に占める割合

第 1 位の「楽しい」については、リツイート・リプライ・それ以外の順に多くの割合を示すことがわかった。第 2 位の「頑張る」は、リプライにおいて占める割合が多く、リツイートのそれを逆転している。第 3 位の「嬉しい」もリプライが最も多くを占める結果となった。リプライにおいて「頑張る」「嬉しい」が多いのは、一般に、友人であるユーザに「頑張って！」と応援したり、「嬉しかった」と感謝の気持ちを伝えるやり取りが多いためと考えられる。

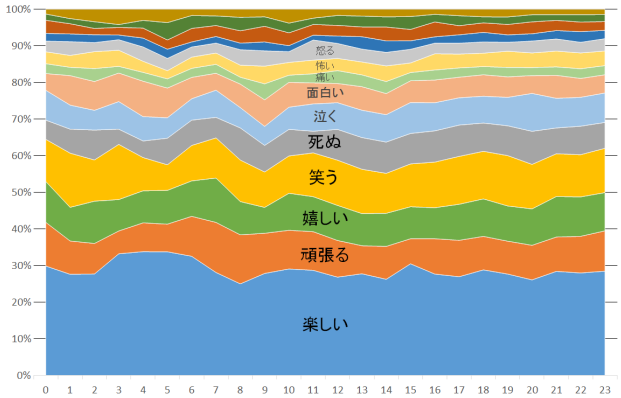
リプライの上位 3 位が占めている 60%に着目すると、リプライは「楽しい」「頑張る」「嬉しい」だけで 60%を示しているのに対し、リツイートは 4 位の「笑う」を加えて、それ以外のツイートは 5 位の「死ぬ」まで加えて 60%に達する。このことについて、リツイートやリプライの絶対数はそれ以外に比べて少ないので、リツイート・リプライ以外のツイートでは様々な表現が用いられると捉えることもできる。逆に、リツイートやリプライで使われる語彙はあまり多くないともいえる。

次に、4 位から 8 位までに着目し、リプライと他の発信タイプを比較すると、リツイートでは「笑う」「死ぬ」「泣く」「面白い」の使用割合が、リツイート・リプライ以外では加えて「痛い」の使用割合が増えている。これらの語は、特定のユーザを相手に発信するリプライよりも、ユーザを限定しないでより広く発信できるリツイートや通常のつぶやきに多いからだと考えられる。特に、「死ぬ」「痛い」に関しては、何かつらい状況を訴えていると考えられるが、特定の個人を相

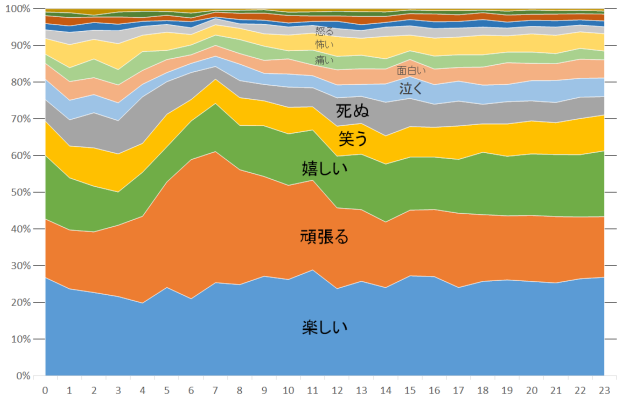
手にしないのは、誰でも良いから気づいて欲しいという感情の表れではないかと考えられる。

4.3 感情表現語を含むツイートの日内変動

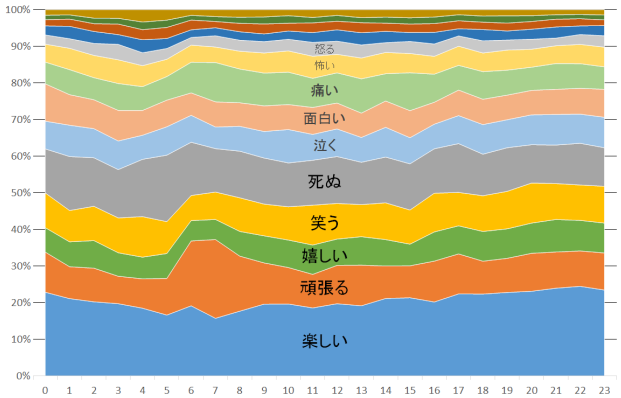
図 5 は、発信タイプごとに、感情表現語がそれぞれの全体に占める割合を時系列に並べたものである。



(a) リツイート



(b) リプライ



(c) リツイート・リプライ以外

図 5: 各感情表現語が全体に占める割合の日内変動

これらの図は感情表現語を含むツイートのみのデータで作図するものであるため、サンプル数が足りなくなる。そこで、今度は計 14 日の全ツイートから感情表現語を含むツイート全 677,093 件を抜き出して新たにコーパスを作り、分析を行った。計算の負荷を減らすため、5 分の 1 にあたる計 135,425 件で作図した。

まず、全体的に午後のツイートについては割合の変化が少なく、早朝 5 時頃から 9 時頃までにかけて変化が多いといえる。この変化の要因は、リツイート以外における「頑張る」という語の影響とみて良いと考える。この時間帯は、図 1 より、もともとリツイートが少ない時間帯である。図 4 で「頑張る」の割合の変化に影響していたのは、この時間帯のツイートが原因と考えられる。

5 まとめと今後の課題

本研究は、一般公開されている日本語ツイートにおいて、発信タイプ別に、14 の感情表現語の使用割合の日内変動を調査したものである。特定のトピックに絞らずに全体の傾向をつかむことで、この先に感情表現の詳細な研究をする上で留意すべき点を見出すことができた。特に深夜帯～早朝帯のツイートについては、全体の偏りが大きいため、留意すべきであるという知見が得られた。

現時点ではリツイート・リプライ以外のツイートを 1 つのタイプとしてみているので、自動ツイート機能である bot の影響を除くなど、他に排他的な発信タイプを設定し、より純粋なつぶやきを区別することで、見えていない傾向がまだ残されていると考えられる。

本研究ではこれらの語がどのような形で使われているのかは調査できていない。本研究は正規表現とのマッチングで言語解析を行ったが、形態素解析や係り受け解析を応用して、より厳密に感情を分析することが必要である。過去の出来事の報告か（映画を見て感動！マジ泣けた）、今の状況を知らせているのか（小指ぶつけた、痛い）、未来へ向けた願望なのか（明日から頑張る！／死にたい）、という視点で調べてみると面白そうである。

現在、本研究で用いた属性情報以外にもさまざまな属性情報を収集している。これらのデータとも組み合わせると、不思議な現象を発見できるかもしれない。

参考文献

[1] 山本 湧輝, 熊本 忠彦, 瀧本 明代, ”ツイートの感情の関係に基づく Twitter 感情軸の決定”, 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（DEIM 2015）, 2015

[2] 奥谷 貴志, 山名 早人, ”メンション情報を利用した Twitter ユーザプロフィール推定”, DBSJ Jabanese Journal, Vol.13-J, No.1, 2014

[3] 加藤 慶一, 秋岡 明香, 村岡 洋一, 山名 早人, ”ミニブログにおける注目語抽出手法の提案と注目語を用いたメディア間での話題追跡”, 情報処理学会研究報告 Vol.2010-DBS-151 No.22, 2010

[4] 上岡 由征, 若宮 翔子, 張 建偉, 白石 優旗, 河合 由紀子, 熊本 忠彦, ”ニュースとツイート分析による話題に対する相関感情俯瞰グラフ”, 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（DEIM 2015）, 2015