令和元年度　卒業論文

言語表現に着目した Twitterデータ分析

千葉大学工学部都市環境システム学科
16T0244U 梶七夏実

指導教員：塩田茂雄
令和２年２月３日提出

目次

第 1 章

序論

1.1　研究背景

近年，インターネット上での情報伝達およびコミュニケーション手段として Facebook や Instagram 等の SNS(Social Networking Service) が広く普及し，世界中で日常的に使用されるようになった．　従来の情報伝達手段では，一般人が発信する情報のほとんどは発信者とその知人等ごく一部の人々の間でのみ共有される情報であり，一般個人が発する情報が広い範囲で共有され，社会に対して影響力を持つことは殆ど無かった．一方で SNS は従来の手段とは異なり，日常生活や趣味といった個人的な話題から，政治・事件など社会的に注目を集めうる話題まで，様々な情報を一般個人が世界中に対して容易に発信しうるという特徴を持っている．また SNS の多くはユーザーが自身が関心を持った他者の投稿を，自身の投稿を閲覧する人々に対して共有する機能を持ち，それによって従来の手段ではごく一部の人々にしか共有されなかった情報であっても，一度他のユーザーの注目を集めると急速に世界中に対して共有されるようになり，従来の手段では見られなかった経路で拡散が発生するようになった．

1.2　研究目的

本研究では，SNS の一種である Twitter を対象として，Twitter 上での情報拡散過程の特徴と使用言語ごとの特徴を明らかにすることを目的として，特定のキーワードを含むツイートデータの収集および分析を行う．

1.3　構成

本論文の構成は以下の通りである．

第１章 序論

研究背景および研究目的を示す．

第２章 準備

Twitter についての説明，既存研究，検索手法について示す．

第３章 リツイート数分布

全言語に共通するリツイート数の分布と，オリジナルツイート割合の関係について示す．

第４章 キーワードの種類ごとの傾向

キーワードをその内容に応じて 4 種類に分類し，種類ごとの傾向を示す．また，日常雨滴キーワードと非日常的キーワードに分類し，両者の傾向についても分析する．

第５章 言語ごとの傾向

ツイートデータを使用言語ごとに分類し，それぞれの傾向を検討する．

第６章 結論

本研究のまとめおよび今後の展望を示す．

第 2 章

準備

2.1 Twitter について

Twitter は SNS の一種であり，2018 年には 1 日にサービスを使用する人数の指標である DAU(Daily Active Usage) が役 1 億 2600 となるほど世界中で普及しており，日本国内においても利用率が約 37.3% と広く社会に普及しているサービスである．[1][2]Twitter においては 140 字以内のテキストおよび写真や動画を投稿することができ，Twitter 上での投稿はツイートと呼ばれ，ユーザーが自身で投稿したツイートであるオリジナルツイートと，他者のツイートを共有する機能であるリツイートに大別される．Twitter においてあるユーザーがオリジナルツイートを投稿すると，投稿者のフォロワーはタイムライン上でそのツイートを閲覧することが可能になる．そして，オリジナルツイートを閲覧したユーザーがそのツイートをリツイートすると，リツイートを行なったユーザーのフォロワーはタイムライン上でそのツイートが閲覧可能となる．リツイートが行われることによって，オリジナルツイートを投稿したユーザーとフォロー関係にないユーザーであっても，そのツイートを閲覧することが可能となり，より広い範囲に情報が拡散されることになる．

2.2 既存研究

Twitter における情報拡散過程について，特定のキーワードを含むツイートデータを収集した場合，リツイート数の分布に関してその補分布はべき分布をとり，そのべき指数は 0,5 から 1,0 程度の値をとることが示されている．[3]Twitter において投稿されたツイートのうち，特定のキーワードを含むものを収集すると，そのキーワードの内容によらず大半はリツイート数が 1 以下である一方で，最大リツイート数は平均リツイート数の数千から数万倍にもなる極端な分布をとることがわかっている． また，収集するキーワードの関しては，「美しい」といった形容詞や一般名詞などの日常的に使用されるキーワードと，現実の事件や出来事と関連してデータを収集した時点において一時的に話題となっているキーワードやデマ・炎上に関連する非日常的キーワードに分類してリツイート数について分析されている．[4] 両者について，全ツイートに占めるオリジナルツイート割合を求めると，日常的キーワードを含むツイートは非日常的キーワードの場合と比較して

オリジナルツイートの割合が大きくなることが示されており，このことから非日常的キーワードではリツイートによる情報拡散が発生しやすいことがわかっている．

2.3　分析手法

本研究では，Twitter API を用いて特定のキーワードを含むツイートについて，各ツイートに固有のツイート ID，ユーザー ID，ツイート時刻，お気に入り数，フォロワー数，リツイート数，認証の有無，ユーザー名，使用言語についてデータを取得し，リツイートの場合にはオリジナルツイートのツイート ID も同時に取得した．Twitter API では収集を開始した時点から最大で過去 7 日分程度のツイートを収集することが可能である．本研究では収集したツイートデータのうち，リツイート件数に焦点を当ててツイートデータの分析を行う．

2.4　ツイートの検索条件

本研究では，3 章・4 章で全ての使用言語によらない Twitter 全般の傾向，そして 5 章では言語ごとに固有の傾向を分析するため，２種類の条件でツイートデータを収集する．３章・4 章で用いるデータについては，使用言語を限定せずに複数言語のツイートを一度に収集するため，#(ハッシュタグ) を用いてデータを収集する．ハッシュタグは SNS 上で使用されるものであり，特定のキーワードに#を付してツイートすることでそのキーワードがタグ化され，SNS 上で容易に同じ内容の投稿を検索することが可能になる．世界中で広く話題になっている事象に関するキーワードの場合，ハッシュタグを用いることで使用言語によらずに表記が統一され，使用言語を指定せずとも一度に複数言語のツイートを収集することが可能になる．また，Twitter API においてはハッシュタグを使用した場合，タグづけされていない単純なキーワードとは異なるものであるとして扱われるため，ハッシュタグを用いなくても複数言語で同じ表記が用いられるキーワードに関しては，ハッシュタグと同時にタグ付けしていないキーワードも検索条件に加えてツイートの収集を行なった．５章においては，言語ごとの傾向を分析するため，3 章・４章で用いたデータに加えて使用言語を日本語と英語に限定してツイートの収集を行なった．3 章・４章で用いたデータに関しては，ツイートデータを収集する際に同時に取得した言語情報を基に, 使用言語ごとのリストを作成した．日本語と英語に限定したデータについては，最初に使用言語を日本語に限定し，日本語のキーワードを含むツイートを収集した後に，改めて使用言語を英語に限定し，同様の意味・用法を持つ英語のキーワードを含むツイートを収集し，分析を行なった．

第 3 章

リツイート数分布

3.1　リツイート数分布

次に, キーワードごとのリツイート件数の分布について分析を行う．リツイート件数の分布に関しては，既存研究よりべき分布を取ることが明らかになっており，さらにそのべき指数は 0.5 から 1.0 程度の値を取るとされている．[3]　各キーワードについて X 軸がリツイート件数，Y 軸がリツイート件数が X を超える確率とした両対数グラフを作成した結果，図 3.1 から図 3.8 に示す通りこれらのグラフ上でデータは全て直線状にプロットされた． ここで，一般に両対数グラフ上で直線状にプロットされる時，その分布はべき分布を取ることから，図 3.1 から図 3.8 に示すキーワードのリツイート件数はべき分布を取ることが確認される．さらに，他のキーワードについても同様にグラフを作成したところ，既存研究と同様に全てのキーワードにおいてリツイート件数の分布はべき分布であることが確認された．ここでべき分布の特徴より，全ツイートのうち多くのツイートはリツイート数が 10 にも満たないほど少ない一方で，ごく一部のツイートのみが極端に多くのリツイートを集めると言える．　べき指数については，一般に両対数グラフ上において直線状にプロットされるデータのべき指数はその直線の傾きと一致することから，各キーワードについて作成したグラフ上のプロットを最小二乗法を用いて直線で近似しその傾きを求め，分析した．その結果，図 3.1 から図 3.8 に挙げたキーワードに関しては表 3.1 に示した通り，べき指数は 0.5 から 1.0 の値をとった．さらに，全てのキーワードについて同様の方法でべき指数を求めたところ，全てのキーワードにおいて既存研究と同様に 0.5 から 1.0 程度の値を取ることが確認された．

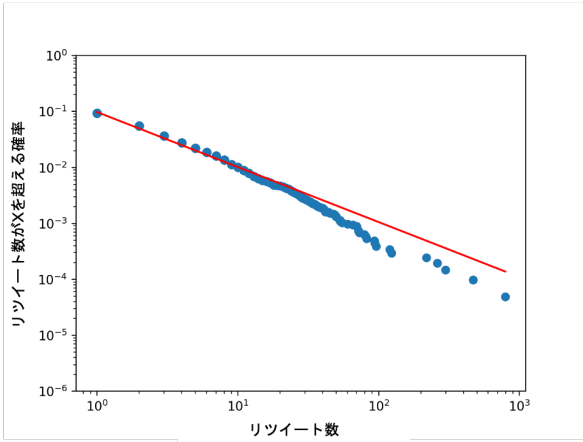

図 3.1　#Apple

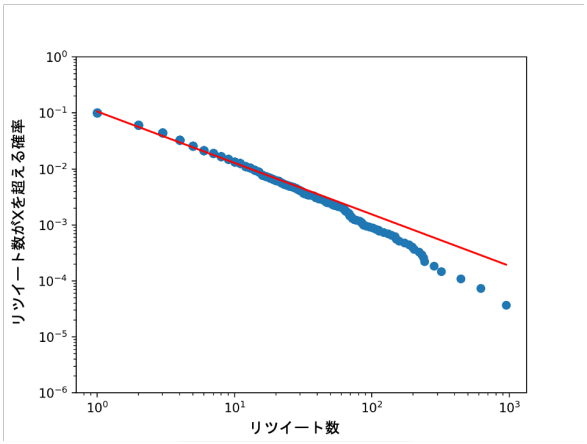

図 3.2　#Facebook

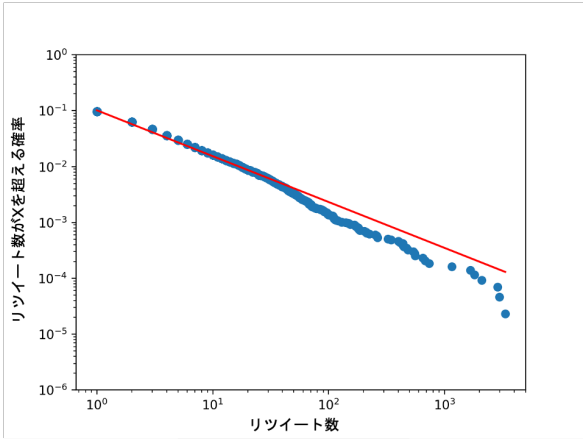

図 3.3　#Twitter

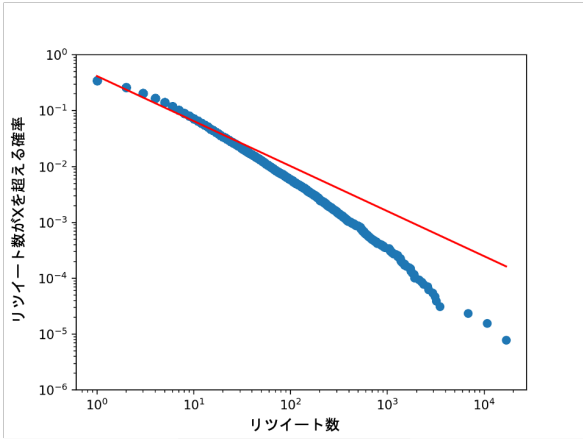

図 3.4　#COP25+COP25

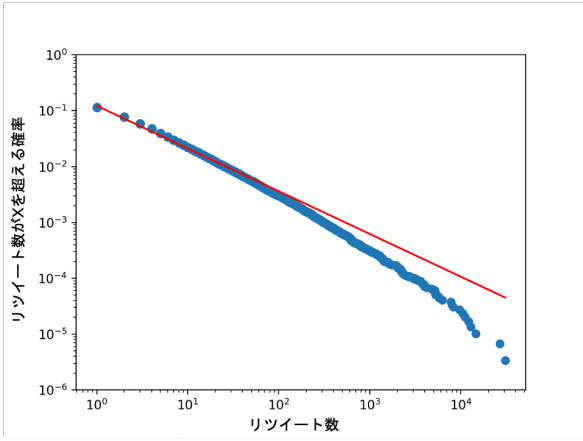

図 3.5　#HappyNewYear+#happynewyear

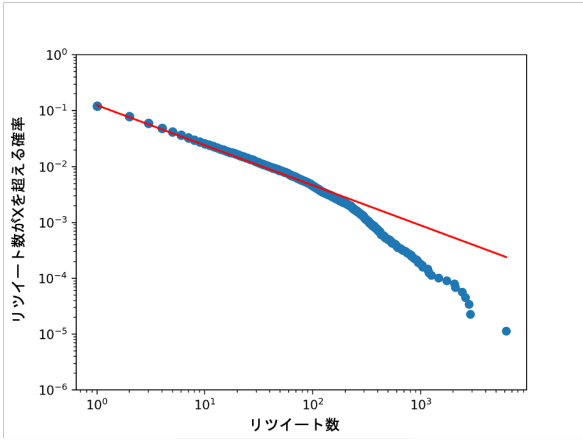

図 3.6　#TRUMP+#trump

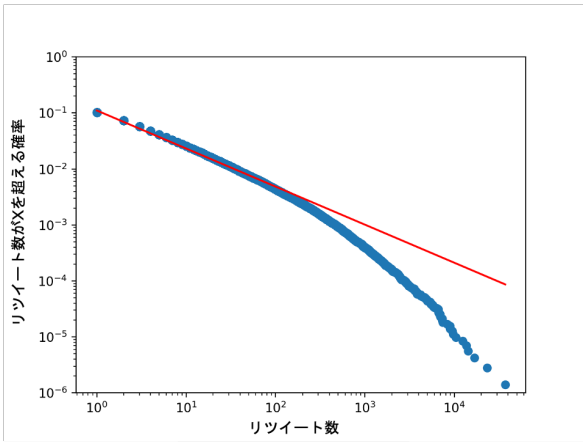

図 3.7　#Brexit

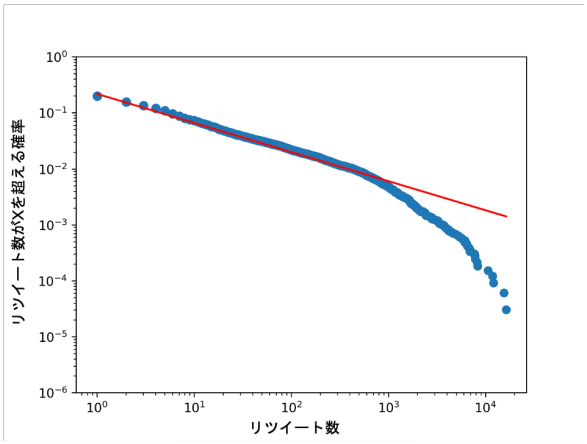

図 3.8　#clubWC

表 3.1　各キーワードのべき指数

キーワード	総ツイート数	オリジナルツイート件数	直線の傾き (べき指数)
#Apple	40,083	20,303	0.9833
#Facebook	53,123	27,110	0.919
#Twitter	112,847	43,071	0.8209
#COP25+COP25	796,385	126,989	0.8059
#HppyNewYear+#happynewyear	1,221,731	293,029	0.7651
#TRUMP+#trump	308,290	87,738	0.7164
#Brexit	2,781,263	705,891	0.6798
#clubWC	674889	32,371	0.5199

3.2　オリジナルツイート割合とべき指数

表 3.2 に各キーワードについてリツイート数分布のべき指数とオリジナルツイート割合を求め, べき指数が大きい順にソートした結果を示す．ここで，オリジナルツイート割合が大きいツイートほど, べき指数が大きくなっており, 両者の間には相関関係が観察される．そして，収集したキーワード全体のオリジナルツイート割合とべき指数の相関係数を求めると約 0.87 であり, 両者の間には強い正の相関が存在することが言語によらず確認された.

表 3.2　オリジナルツイート割合とべき指数

キーワード	オリジナルツイート割合 (%)	べき指数
#spotify	55.03	1.019
#Apple	50.65	0.9833
#YouTube	40.76	0.9822
#TGIF+#tgif	48.71	0.9781
#Instagram+#instagram+#insta	38.38	0.9375
#iPhone	60.99	0.9274
#Facebook	51.03	0.919
#notredame	24.19	0.8356
#google+#GOOGLE+#Google	38.05	0.8226
#Twitter	38.17	0.8209
#FridayThe13th	36.58	0.8074
#COP25+COP25	15.95	0.8059
#Netflix+#netflix	26.81	0.7977
#Amazon+#amazon	38.59	0.7888
#happynewyear+#HappyNewYear	23.98	0.7651
#StarWarsTheRiseofSkywalker+#StarWars9+#StarWars	32.66	0.7464
#JamesBond+JamesBond	26.00	0.7455
#trump+#TRUMP	28.46	0.7164
#Greta+#GretaThunberg	24.55	0.7122
#MissWorld2019	18.73	0.6964
#Nakamura+#nakamura	11.55	0.6883
#IranPlaneCrash	19.58	0.6801
#Brexit	25.38	0.6798
#NobelPrize	15.59	0.6717
#Iran	14.94	0.6657
#HBD+#hbd	8.95	0.6459
#OMG	24.67	0.5881
#HongKong	7.92	0.5447
#streamys	5.90	0.5266
#clubWC	4.80	0.5199

第 4 章

キーワードの種類ごとの傾向

4.1　分類ごとの傾向

ここでは，キーワードを (1) イベントに関連するキーワード，(2) 政治・事件に関連するキーワード，(3) 企業・商品・サービス名，(4) スラングの 4 種類に分類し，各キーワードについてオリジナルツイート割合・リツイート数補分布のべき指数を求めた．また，総ツイート数およびオリジナルツイート数の 1 分毎の変化を示したグラフを作成し，分類ごとの傾向を考察した．ツイート数の時間変化に関しては，既存研究より特定のキーワードについてツイートを検索した場合に，現実の事象に直接関わりのないキーワードでは昼夜変動を繰り返すものの，日変動はあまり観察されないことが示されているが [5]，[6] 本研究において収集したキーワードについても検討する．

4.1.1　イベントに関連するキーワード

イベントに関連するキーワードを含むツイートを収集・分析した結果が表 4.1 である．ここで挙げたキーワードは他の分類と比較してオリジナルツイート割合が低い傾向が見られた．図 4.1 から図 4.6 には，一部キーワードのツイート数時間変化を示す．ここに示すグラフではいずれも特定の日時でツイート数はピークを迎えているという特徴を持ち，そのピークは現実にこれらのイベントが発生する日時と一致している．図 4.1 の’#FridayThe13th’の例では 12 月グラフ上のピークは 12 月 14 日であるが，これは複数言語を同時に収集しているため，時差による影響受けているものとみられる．図 4.3 に示す’#StarWarsTheRiseofSkywalker’ や図 4.4 に示す’#JamesBond’ の例はいずれも映画の公開に合わせた使用されていたキーワードだが，これらも映画の公開日前後にピークを迎えている．図 4.5 に示す’#NobelPrize’ の例では，2019 年 12 月 11 日にノーベル賞の授賞式が行われたことによって，ツイート数はピークを迎えているものとみられる．図 4.6 に示す’#clubwc’ の例は，FIFA Club World Cup の開催に合わせて使用されていたキーワードであり，グラフ上では複数のピークを迎えている．これは，大会の期間中に試合開催の度にピークを迎えていることによるものと推測される．

表 4.1　イベントに関連する単語

キーワード	期間	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合 (%)	べき指数
#FridayThe13th	2019/12/5-2019/12/14	103100	37715	36.580989	0.8074
#COP25+COP25	2019/12/4-2019/12/12	796385	126989	15.9456795	0.8059
#HappyNewYear+#happynewyear	2020/1/1-2020/1/7	1221731	293029	23.9847	0.7651
#StarWarsTheRiseofSkywalker	2019/12/14-2019/12/21	879675	287299	32.6596754	0.7464
#JamesBond+JamesBond	2019/11/26-2019/12/5	114829	29853	25.997788	0.7455
#Greta+#GretaThunberg	2019/12/11-2019/12/19	156323	38379	24.5510897	0.7122
#MissWorld2019	2019/12/11-2019/12/19	156925	29388	18.7274176	0.6964
#NobelPrize	2019/12/5-2019/12/14	44143	6882	15.5902408	0.6717
#streamys	2019/12/7-2019/12/16	40057	2365	5.9040867	0.5266
#clubWC	2019/12/14-2019/12/22	674889	32371	4.7964925	0.5199

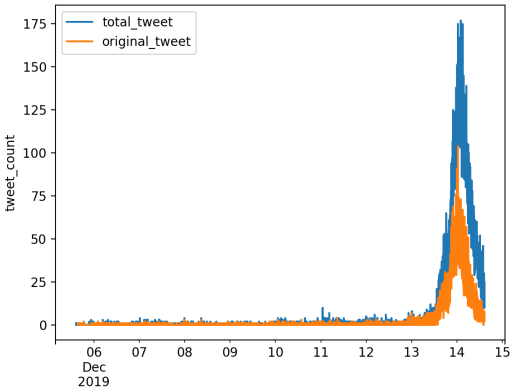

図 4.1　#FridayThe13th

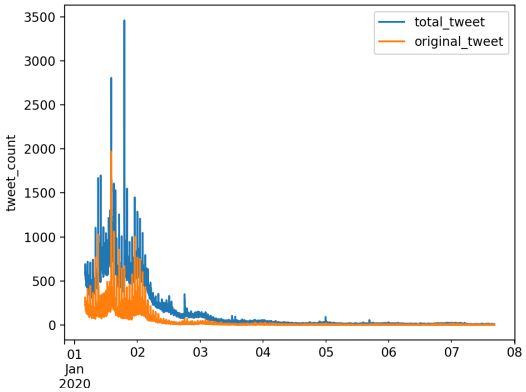

図 4.2　#HappyNewYear+#happynewyear

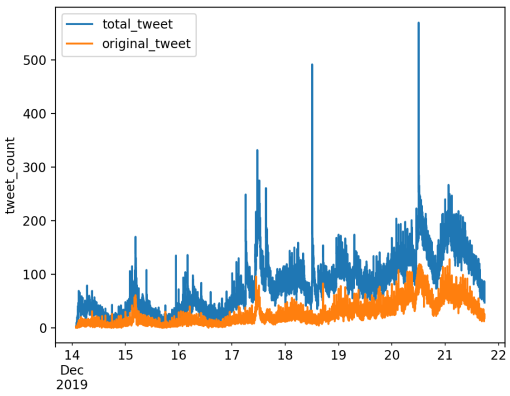

図 4.3　#StarWarsTheRiseofSkywalker

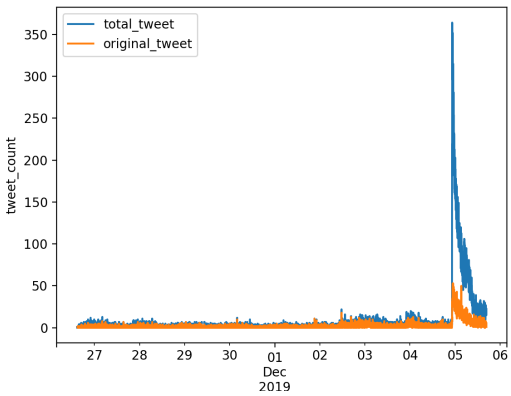

図 4.4　#JamesBond+#jamesbond

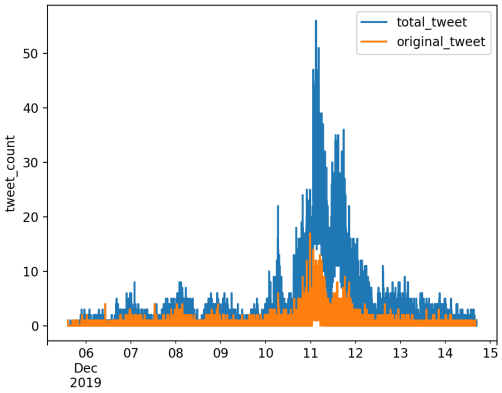

図 4.5　#NobelPrize

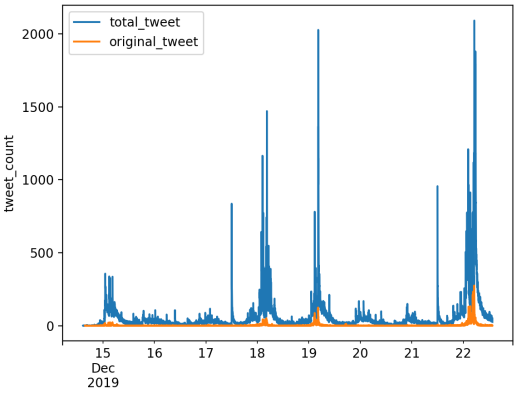

図 4.6　#clubwc

4.1.2　政治・事件に関連するキーワード

次に，政治に関連するキーワードを含むツイートを収集，分析した結果を表 4.2 に示す．図 4.7 に示す'#Brexit' の例では，この話題自体は長期的に議論されているものであるため，イギリスの現地時刻に合わせて昼夜変動が起こっており，12 月 13 日には，イギリス国内での選挙結果を受けてピークを迎えている．図 4.8 に示す'#Iran' の例と図 4.9 に示す'#IranPlaneCrash の例はいずれも，米国とイラン間の政情変化を受けて使用されているキーワードであるとみられ，2020 年 1 月 8 日にイランの首都テヘランを離陸した旅客機が墜落した事件を受けてその前後および関係国の政治的発言等を受けてツイート数は変化しているものと推測される．

表 4.2　政治・事件に関連する単語

キーワード	期間	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合	べき指数
#trump+#TRUMP	2019/12/1-2019/12/9	308290	87738	0.284595673	-0.7164
#IranPlaneCrash	2020/1/8-2020/1/14	215275	42161	19.58	0.6801
#Brexit+Brexit	2019/12/7-2019/12/14	2781263	705891	0.253802319	-0.6798
#Iran	2020/1/5-2020/1/11	1563915	233710	0.149439068	-0.6657
#HongKong	2019/12/7-2019/12/14	294402	23305	0.079160468	-0.5447

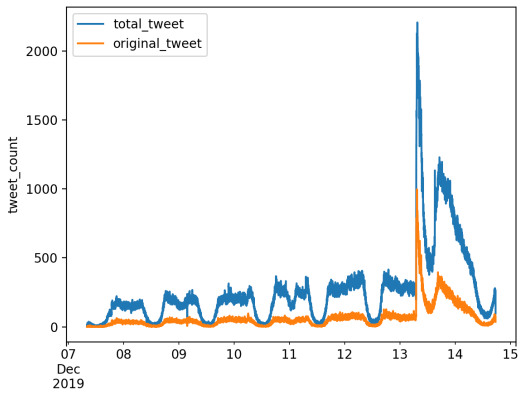

図 4.7　#Brexit

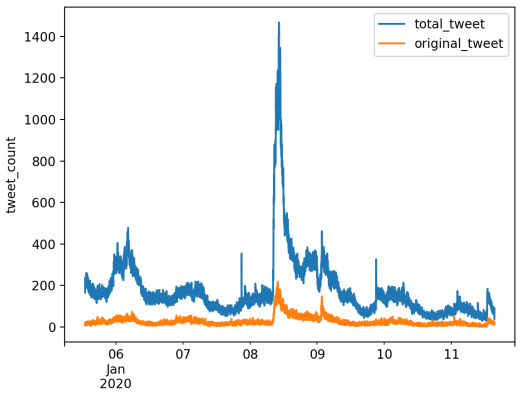

図 4.8　#Iran

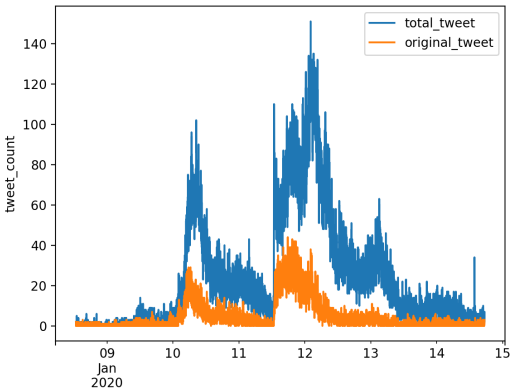

図 4.9　#IranPlaneCrash を含むツイート数の時間変化

4.1.3　企業・商品・サービス名

次に，企業・商品・サービス名を含むツイートを収集・分析した結果を表 4.3 に示す．ここであげるキーワードは他の分類に含まれるキーワードと比較してオリジナルツイートの割合が高い傾向があり，特に web サービス名である#spotify,#Apple,#iPhone,#Facebook が含まれるツイートはオリジナルツイートの割合が半数を超える．図 4.10 から図 4.14 に示すグラフはいずれも昼夜変動は見られるものの，それ以外に急激なピーク観察されなかった．

表 4.3　企業・商品・サービス名

キーワード	期間	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合	べき指数
#spotify	2019/11/26-2019-/12/5	55832	30723	0.550275827	-1.019
#Apple	2019/12/13-2019/12/21	40083	20303	0.506523963	-0.9833
#YouTube	2019/11/24-2019/12/3	236345	96333	0.40759483	-0.9822
#Instagram+#instagram+#insta	2019/11/29-2019/12/8	114084	43791	0.383848743	-0.9375
#iPhone	2019/11/24-2019/12/3	39145	23873	0.609860774	-0.9274
#Facebook	2019/12/5-2019/12/14	53123	27110	0.510325095	-0.919
#google+#GOOGLE+#Google	2019/12/1-2019/12/10	79481	30241	0.38048087	-0.8226
#Twitter	2019/12/3-2019/12/12	112847	43071	0.381676075	-0.8209
#Netflix+#netflix	2019/11/24-2019/12/3	110780	29705	0.268144069	-0.7977
#Amazon	2019/12/3-2019/12/10	177549	68514	0.385887839	-0.7888

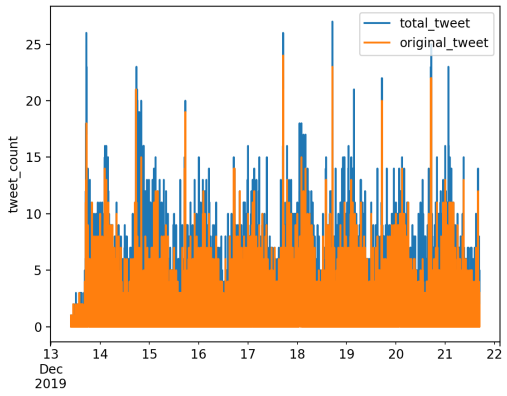

図 4.10　#Apple

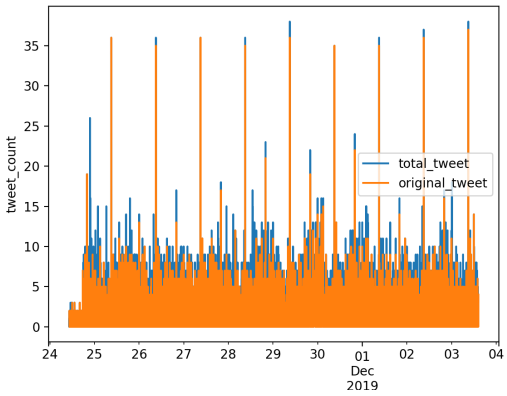

図 4.11　#iPhone

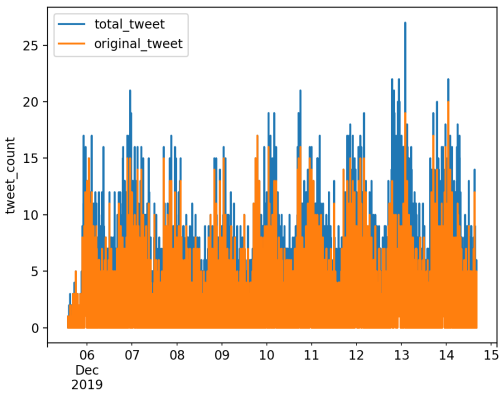

図 4.12　#Facebook

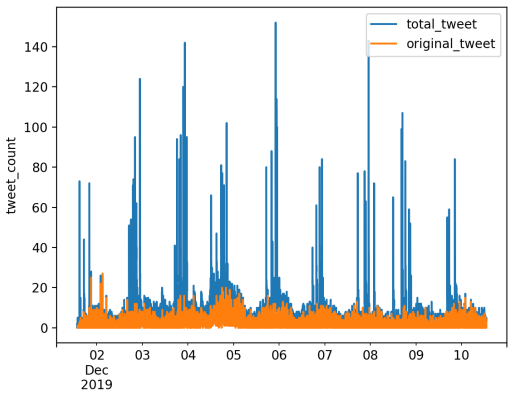

図 4.13　#google+#GOOGLE+#Google

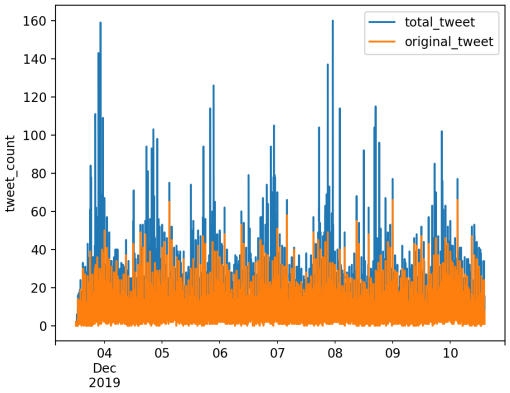

図 4.14　#Amazon

4.1.4　スラング

次に, スラングを含むツイートを収集・分析した結果を表 4.4 に示す. 図 4.15 に示す'#OMG'(Oh My Godness) の例では，昼夜変動以外にツイート数の大きな変動は観察されなかった．図 4.16 に示す'#TGIF'(Thanks God It's Friday) の例では，金曜日の前後でツイート数が増加しているものの，これは周期的な変化であると見られ，スラングに関しても昼夜変動以外などの周期的な変動を除くと，特定の日時における急激な変化は特に見られものと推測される．

表 4.4　スラング

キーワード	期間	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合	べき指数
#TGIF+#tgif (Thanks God It's Friday)	2019/11/29-2019/12/8	24857	12107	0.487066018	-0.9781
#HBD+#hbd (Happy BirthDay)	2019/11/20-2019/12/9	23579	2111	0.089528818	-0.6459
#OMG (Oh My Godness)	2019/12/1-2019/12/10	13353	3294	0.246686138	-0.5881

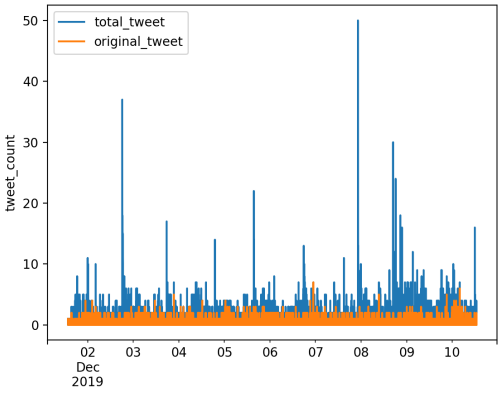

図 4.15　#OMG+#omg

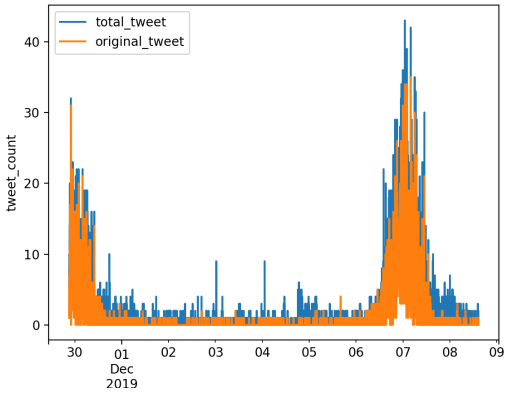

図 4.16　#TGIF+#tgif

4.2　日常的キーワード・非日常的キーワード

次に，収集したキーワードを日常的キーワードと非日常的キーワードに分類し，それぞれの傾向を分析する．両者の傾向については，既存研究より前者は後者に比較してオリジナルツイート割合が大きくなることが示されているが，[4] 本研究で収集したキーワードについてもオリジナルツイート割合について検討する．

4.2.1　日常・非日常の分類

次に，収集したキーワードを日常的キーワードと非日常的キーワードに分類し，分析する．ここで日常的キーワードとは時間依存性の低いキーワードのことを指し，一方で非日常的キーワードとは時間依存性が強く，現実世界の事象と連動して一時的に極端にツイート数が増加するキーワードとする．前節において，キーワードの分類ごとにツイート数の時間変化について分析を行なったがその結果，(1) イベントに関連するキーワード，(2) 政治・事件に関連するキーワードでは，その

キーワードに関連する現実の事象に連動して特定の時点でツイート数はピークを迎え，一方で (3) 企業・商品・サービス名，(4) スラングに分類されるキーワードでは昼夜変動や特定の曜日におけるツイート数の変化はあるものの，現実の事象と連動して急激にツイート数が変化する様子は観察されなかった．よって，(3) 企業・商品・サービス名 (4) スラングの 2 種を時間依存性の弱い日常的キーワード，(1) イベントに関連するキーワード (2) 政治・事件に関連するキーワードの 2 種を時間依存性の強い非日常的キーワードとして分析を行う．

4.2.2 日常・非日常の分類による傾向

次に，日常的キーワードと非日常的キーワードについて分類別に各キーワードのオリジナルツイート割合を求めた結果を表 4.6 および表 4.5 に示す．既存研究より両者のオリジナルツイート割合を比較した時に，前者の方がその割合が大きくなることが明らかになっているが [**?**]，表 4.6，表 4.5 に示す結果においても同様の傾向が確認された．さらに，表 4.7 に示す通り，全ての日常的キーワードと非日常的キーワードの合計から全体のオリジナルツイート割合を求めた結果，同様に前者の方がオリジナルツイート割合が大きくなることが確認された．以上より，ツイート数の時間依存性が低い日常的キーワードと比較して，時間依存性が強い非日常的キーワードではリツイートによる情報の拡散が発生しづらいという傾向が認められる．

表 4.5　非日常的キーワードのオリジナルツイート割合

キーワード	オリジナルツイート割合
#iPhone	60.99%
#spotify	55.03%
#Facebook	51.03%
#Apple	50.65%
#TGIF+#tgif	48.71%
#YouTube	40.76%
#Amazon	38.59%
#Instagram+#instagram+#insta	38.38%
#Twitter	38.17%
#google+#GOOGLE+#Google	38.05%
#Netflix+#netflix	26.81%
#OMG	24.67%
#HBD+#hbd	8.95%

表 4.6　日常的キーワードのオリジナルツイート割合

キーワード	オリジナルツイート割合
#FridayThe13th	36.58%
#StarWarsTheRiseofSkywalker+#StarWars9+#StarWarsr	32.66%
#trump+#TRUMP	28.46%
#JamesBond+JamesBond	26.00%
#Brexit	25.38%
#Greta+#GretaThunberg	24.55%
#notredame	24.19%
#happynewyear+#HappyNewYear	23.98%
#MissWorld2019	18.73%
#COP25+COP25	15.95%
#NobelPrize	15.59%
#Iran	14.94%
#HongKong	7.92%
#streamys	5.90%
#clubWC	4.80%

表 4.7　日常的キーワードと非日常的キーワードの合計

	日常的キーワード	非日常的キーワード
ツイート数合計	1,081,058	9,135,927
オリジナルツイート数合計	431,176	1,934,914
オリジナルツイート割合	39.88%	21.18%

第 5 章

言語ごとの傾向

5.1　日本語と英語の比較

次に，ツイートに使用される言語ごとの特徴を考察するために，日本語と英語に限定してツイートの収集・分析を行った．その結果をキーワードを (1) 状態を表す形容詞 (2) その他形容詞 (3) 挨拶に分類し, 表 5.1 から表 5.6 に示す．

表 5.1　状態を表す形容詞（日本語）

キーワード	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合
寒い	1,123,742	792,877	70.56%
うるさい	244,424	187,384	76.66%
静か	295,555	162,805	55.085%
眠い	679,334	585,573	86.20%

表 5.2　状態を表す形容詞（英語）

キーワード	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合
cold	1,869,936	839,316	44.88%
noisy	48,087	16,015	33.30%
quiet	519,573	167,468	32.23%
sleepy	337,521	126,375	37.44%

表 5.3　その他形容詞（日本語）

キーワード	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合
カラフル	111,987	33,809	30.19%
現実的	118,681	48,798	41.12%
主観	43,124	21,230	49.23%
垂直	16,001	8,231	51.44%

表 5.4　その他形容詞（英語）

キーワード	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合
colorful	85,735	30,973	36.13%
realistic	124,569	47,643	38.25%
subjective	36,000	23,921	66.45%
vertical	39,715	16,402	41.30%

表 5.5　挨拶（日本語）

キーワード	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合
おはよう	3,000,000	2,228,550	74.29%
おやすみ	998,389	786,530	78.78%
こんにちは	684,882	485,791	70.93%

表 5.6　挨拶（英語）

キーワード	総ツイート件数	オリジナルツイート件数	オリジナルツイート割合
good morning	1,494,467	638,514	42.73%
good night	747,316	326,943	43.75%
hello	1,272,541	482,854	37.94%

5.1.1　オリジナルツイート割合

キーワードごとに日本語とそれに対応する英語のキーワードを含む総ツイートに占めるオリジナルツイートの割合を比較したところ，ほとんどのキーワードで日本語の方がオリジナルキーワード割合が大きくなった．特に (1)，(3) に分類されるキーワードでは全てにおいて使用言語が日本語

である場合, 英語の場合と比較してオリジナルツイート割合が高くなった．
また，収集した全てのキーワードについて総ツイート件数とオリジナルツイート件数の合計を求め，そこから全体のオリジナルツイート割合を求めた結果が表 5.7 である．ここで示した通り，日本語のキーワードを含むツイートのオリジナルツイート割合が 70% を超えているのに対して英語のキーワードを含むツイートでは約 38.5% にとどまっている．このことから，キーワードによって差はあるものの，全体の傾向として日本語のキーワードを含むツイートは英語のキーワードを含むツイートに対してオリジナルツイート割合が高くなるという結果が得られた．ここで, オリジナルツイート割合が高いということは，それだけリツイートによる情報の拡散が発生しづらいということを示しており，つまり日本語を使用したツイートはオリジナルツイートを投稿したユーザーのフォロワーへの拡散が中心であるのに対して，英語を使用したツイートではリツイートを利用したオリジナルツイートを投稿したユーザーのフォロワー以外への拡散が日本語を使用した場合よりも発生しやすい傾向があるといえる．

表 5.7　全キーワード合計のツイート数とオリジナルツイート数

	日本語	英語
総ツイート件数	7,721,447	7,781,738
オリジナルツイート件数	5,468,269	2,998,890
オリジナルツイート割合 (%)	70.819226%	38.537535%

5.1.2　べき指数

次に，図 5.8，図 5.9 に日本語・英語それぞれについてオリジナルツイート割合とリツイート数補分布のべき指数を求め，べき指数の順に並べた結果を示す．ここで，図 5.8，図 5.9 両者においてべき指数が大きくなるほど，オリジナルツイート割合が大きくなっている傾向が確認された．そして，日本語・英語両者についてオリジナルツイート割合とべき指数の相関係数を求めたところ，日本語では約 0.84，英語では約 0.59 となり，ここから確かに日本語・英語それぞれについてオリジナルツイート割合とリツイート数補分布のべき指数の間には相関関係が存在すると認められる．

表 5.8　オリジナルツイート割合とべき指数 (日本語)

キーワード	オリジナルツイート割合	べき指数
眠い	86.20%	0.8051
おはよう	74.29%	0.7989
こんにちは	70.93%	0.7858
寒い	70.56%	0.7739
おやすみ	78.78%	0.7629
静か	55.08%	0.7501
うるさい	76.66%	0.7489
主観	49.23%	0.729
カラフル	30.19%	0.7238
現実的	41.12%	0.7167
垂直	51.44%	0.6911
国際	31.26%	0.6865

表 5.9　オリジナルツイート割合とべき指数 (英語)

キーワード	オリジナルツイート割合	べき指数
cold	44.88%	0.8189
subjective	66.45%	0.8002
vertical	41.30%	0.7952
goodnight	43.75%	0.7774
realistic	38.25%	0.7738
goodmorning	42.73%	0.7661
international	23.42%	0.7363
colorful	36.13%	0.7257
quiet	32.23%	0.713
sleepy	37.44%	0.7024
noisy	33.30%	0.6813
hello	37.94%	0.6793

以上より，日本語・英語両方に共通して第３章で分析を行なった結果と同様にオリジナルツイート割合とべき指数の間には相関関係が確認されたが，べき指数に関しては日本語と英語で大きな差は観察されず，各言語に固有の特徴は特に確認されなかった．よって，日本語を使用したツイートと英語を使用したツイートではリツイート数の分布に差は見られず，ごく一部のツイートが極端に多くのリツイートを集めている傾向は共通していることがここから確認された．

5.2　他言語との比較

次に日本語・英語以外の言語について，言語毎の傾向を分析する．ここでは第３章・４章で用いたデータについて取得した言語情報を基に使用言語毎に分類し，分析を行う．ここで収集したデータは，複数言語のツイートをを同時に収集することを目的としてキーワードを指定したものであるため,’＃’を付した世界的に使用されうるキーワードを選択した．

収集したツイートのうち，使用言語が多いものを中心に分析を行うため，日本語・英語・スペイン語・フランス語・ドイツ語・イタリア語の６言語について，第４章で述べた日常的キーワードと非日常的キーワードに分類し，オリジナルツイート割合とリツイート数の補分布のべき指数を求めた結果を表 5.10 から表 5.21 に示す．ただし，フランス語・ドイツ語・イタリアについては収集したキーワードのうち，総ツイート数が 100 未満のキーワードは除いて分析を行なった．

表 5.10　日本語の日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#iPhone	64.18%	0.8061
#spotify	22.91%	0.5986
#Facebook	72.08%	0.7958
#Apple	39.32%	0.7293
#TGIF+#tgif	47.44%	0.9965
#YouTube	31.84%	0.8373
#Amazon	28.18%	0.5038
#Instagram+#instagram+#insta	9.28%	0.8368
#Twitter	25.57%	0.6773
#google+#GOOGLE+#Google	11.20%	0.4059
#Netflix+#netflix	28.04%	0.6518
#OMG	5.56%	0.4227
#HBD+#hbd	6.58%	0.5984

表 5.11　日本語の非日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#FridayThe13th	48.14%	0.964
#StarWarsTheRiseofSkywalker+#StarWars9	27.13%	0.6804
#trump+#TRUMP	31.13%	0.6609
#JamesBond+JamesBond	51.24%	0.6396
#Brexit	33.78%	0.5986
#Greta+#GretaThunberg	64.56%	0.9013
#happynewyear+#HappyNewYear	7.26%	0.6193
#COP25+COP25	18.72%	0.6393
#NobelPrize	22.93%	0.7052
#Iran	18.86%	0.6706
#HongKong	5.04%	0.5277
#streamys	51.04%	0.697
#clubWC	62.35%	0.6351

表 5.12　英語の日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#iPhone	60.03%	1.083
#spotify	59.62%	1.101
#Facebook	55.62%	0.999
#Apple	52.27%	1.041
#TGIF+#tgif	46.22%	0.9647
#YouTube	46.16%	1.029
#Amazon	49.23%	0.942
#Instagram+#instagram+#insta	54.70%	0.9405
#Twitter	38.10%	0.8482
#google+#GOOGLE+#Google	45.34%	0.9017
#Netflix+#netflix	41.63%	0.8495
#OMG	34.09%	0.7645
#HBD+#hbd	7.09%	0.7003

表 5.13　英語の非日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#FridayThe13th	33.23%	0.7939
#StarWarsTheRiseofSkywalker+#StarWars9	29.82%	0.7215
#trump+#TRUMP	27.59%	0.6976
#JamesBond+JamesBond	21.00%	0.7224
#Brexit	24.54%	0.6686
#Greta+#GretaThunberg	24.95%	0.7048
#happynewyear+#HappyNewYear	26.58%	0.7761
#COP25+COP25	12.84%	0.8022
#NobelPrize	12.53%	0.6509
#Iran	13.94%	0.6444
#HongKong	6.97%	0.5284
#streamys	4.80%	0.5196
#clubWC	5.71%	0.4984

表 5.14　スペイン語の日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#iPhone	70.95%	1.135
#spotify	52.69%	1.142
#Facebook	52.96%	0.9653
#Apple	66.70%	1.151
#TGIF+#tgif	80.13%	1.249
#YouTube	43.38%	1.043
#Amazon	44.20%	0.7456
#Instagram+#instagram+#insta	55.94%	0.939
#Twitter	43.74%	0.7871
#google+#GOOGLE+#Google	67.15%	1.152
#Netflix+#netflix	56.09%	0.9635
#OMG	67.35%	0.8216
#HBD+#hbd	41.15%	0.8148

表 5.15　スペイン語の非日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#FridayThe13th	37.64%	0.7163
#StarWarsTheRiseofSkywalker+#StarWars9	41.09%	0.8283
#trump+#TRUMP	25.58%	0.7248
#JamesBond+JamesBond	51.18%	0.8998
#Brexit	33.44%	0.8024
#Greta+#GretaThunberg	25.22%	0.7682
#happynewyear+#HappyNewYear	45.95%	0.8611
#COP25+COP25	17.27%	0.815
#NobelPrize	7.34%	0.6143
#Iran	11.20%	0.6466
#HongKong	20.59%	0.7761
#streamys	11.72%	0.5858
#clubWC	8.85%	0.6271

表 5.16　フランス語の日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#iPhone	62.72%	0.9368
#spotify	52.50%	0.7433
#Facebook	43.82%	1.017
#Apple	50.33%	0.9574
#TGIF+#tgif	51.43%	0.9425
#YouTube	53.41%	1.187
#Amazon	49.19%	1.074
#Instagram+#instagram+#insta	48.83%	1.016
#Twitter	42.72%	0.8551
#google+#GOOGLE+#Google	58.56%	1.08
#Netflix+#netflix	37.91%	0.7893
#OMG	35.64%	0.7037

表 5.17　フランス語の非日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#FridayThe13th	50.00%	0.9936
#StarWarsTheRiseofSkywalker+#StarWars9	29.83%	0.8257
#trump+#TRUMP	18.11%	0.7532
#JamesBond+JamesBond	53.79%	0.7999
#Brexit	25.15%	0.6742
#Greta+#GretaThunberg	25.76%	0.7546
#happynewyear+#HappyNewYear	38.35%	0.8719
#COP25+COP25	26.91%	0.8712
#NobelPrize	37.41%	1.116
#Iran	11.51%	0.6747
#HongKong	18.07%	0.7586
#clubWC	15.94%	0.8066

表 5.18　ドイツ語の日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#iPhone	78.72%	1.358
#spotify	54.88%	1.187
#Facebook	42.34%	0.9872
#Apple	58.72%	0.9688
#TGIF+#tgif	66.22%	1.184
#YouTube	32.57%	1.284
#Amazon	48.58%	1.074
#Instagram+#instagram+#insta	51.73%	1.171
#Twitter	30.89%	0.9014
#google+#GOOGLE+#Google	59.13%	1.013
#Netflix+#netflix	68.72%	1.156

表 5.19　ドイツ語の非日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#FridayThe13th	51.17%	1.278
#StarWarsTheRiseofSkywalker+#StarWars9	42.87%	1.141
#trump+#TRUMP	33.55%	0.9084
#Brexit	36.66%	0.7689
#Greta+#GretaThunberg	17.91%	0.6871
#happynewyear+#HappyNewYear	34.44%	0.8606
#COP25+COP25	18.55%	0.7939
#NobelPrize	37.41%	0.8403
#Iran	26.93%	0.8018
#HongKong	3.02%	0.5061
#clubWC	67.87%	0.8498

表 5.20　イタリア語の日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#iPhone	44.37%	1.969
#spotify	56.21%	0.9726
#Facebook	16.68%	0.6315
#Apple	41.23%	1.509
#YouTube	68.34%	1.069
#Amazon	63.14%	0.9647
#Instagram+#instagram+#insta	59.09%	1.134
#Twitter	43.66%	0.7597
#google+#GOOGLE+#Google	68.77%	1.412
#Netflix+#netflix	74.38%	1.233

表 5.21　イタリア語の非日常的キーワード

キーワード	オリジナルツイート割合	べき指数
#StarWarsTheRiseofSkywalker+#StarWars9	66.10%	1.05
#trump+#TRUMP	31.36%	0.872
#JamesBond+JamesBond	69.91%	1.073
#Brexit	23.48%	0.6705
#Greta+#GretaThunberg	18.93%	0.6751
#happynewyear+#HappyNewYear	39.48%	0.8467
#COP25+COP25	28.94%	0.877
#NobelPrize	25.67%	1.16
#Iran	21.65%	0.7966
#HongKong	7.45%	0.7283
#clubWC	9.50%	0.4721

5.2.1　オリジナルツイート割合とべき指数の相関

ここで，全ての言語において，リツイート数補分布のべき指数が大きいほど，オリジナルツイート割合が大きくなっている様子が観察された．そして全 6 言語について両者の相関係数を求めた結果が表 5.22 であり，全言語に共通して両者の間には相関関係が認められた．よって，オリジナルツイート割合とべき指数の相関関係は第 3 章で述べた通り全言語に共通していることが確認されたが，一方で各言語に固有の特徴は見られなかった．

表 5.22　言語ごとのオリジナルツイート割合とべき指数の相関

使用言語	相関係数
日本語	0.615759351
英語	0.936032077
スペイン語	0.856885965
フランス語	0.623298038
ドイツ語	0.720538315
イタリア語	0.541021162

5.2.2　日常的キーワードと非日常的キーワード

次に，第 4 章で述べた日常的キーワードと非日常的キーワードのオリジナルツイート割合についても言語ごとに分析した．表 5.10 から表 5.21 に示すデータより，第 4 章で分析した結果と同様に日常的キーワードは非日常的キーワードと比較してオリジナルツイート割合が高くなる傾向が見られた．ここでさらに，日常的キーワードと非日常的キーワード両者について，全キーワード合計のオリジナルツイート割合を全 6 言語について求めた結果が表 5.23 および 5.24 であり，いずれの言語においても第 4 章で述べた傾向と同様に日常的キーワードは非日常的キーワードと比較してオリジナルツイート割合が高くなることが確認された．

表 5.23　言語毎の日常的キーワード合計オリジナルツイート割合

使用言語	総ツイート数	オリジナルツイート数	オリジナルツイート割合
日本語	311,290	82,556	26.52%
英語	476,568	220,052	46.17%
スペイン語	81,834	40,873	49.95%
フランス語	27,653	13,164	47.60%
ドイツ語	25,483	10,430	40.93%
イタリア語	17,060	7,285	42.70%

表 5.24　言語毎の非日常的キーワード合計オリジナルツイート割合

使用言語	総ツイート数	オリジナルツイート数	オリジナルツイート割合
日本語	166,331	27,263	16.39%
英語	6,493,743	1,461,373	22.50%
スペイン語	684,470	139,995	20.45%
フランス語	229,342	51,182	22.32%
ドイツ語	137,675	36,425	26.46%
イタリア語	119,068	29,243	24.56%

5.2.3　オリジナルツイート割合

次に，キーワードの種類によらず全キーワードについて言語毎のオリジナルツイート割合を求めた結果が表 5.25 である．ここに示す通り，特定のキーワードを含むツイートを使用言語を指定せずに収集し，言語毎に分類した時，オリジナルツイート割合は言語間で差は見られず，分析を行なった 6 言語の全てでオリジナルツイート割合は 20〜30% であった．以上より，使用言語を指定せずに収集したツイートデータについて第 3 章・第 4 章と同様の分析を行なった結果，各言語に固有の傾向は特にみられないと結論づけられる．

表 5.25　各言語の全キーワード合計オリジナルツイート割合

使用言語	総ツイート数	オリジナルツイート数	オリジナルツイート割合
日本語	477,621	109,819	22.99%
英語	6970311	1681425	24.12%
スペイン語	766,304	180,868	23.60%
フランス語	256,995	64,346	25.04%
ドイツ語	163,158	46,855	28.72%
イタリア語	136,128	36,528	26.83%

第 6 章

結論

6.1　まとめ

本研究では，特定のキーワードを含むツイートを収集したツイートデータについて，主にリツイートに焦点を当てて分析を行なった．結果として，使用言語によらず全言語に共通する傾向としてリツイート数の補分布はべき分布をとり，そのべき指数とオリジナルツイート割合の間には相関関係が見られた．使用言語毎の傾向としては，日本語と英語に限定して収集したデータで日本語と比較して英語のキーワードではオリジナルツイート割合が低くなる傾向が見られたが，使用言語を限定せずに複数言語について分析したデータでは，オリジナルツイート割合に大きな差は見られず，他の要素に関しても各言語に固有の特徴は観察されないという結果が得られた．

6.2　今後の展望

本研究では主にリツイートに焦点を当ててツイートデータの分析を行なった結果，各言語に固有の傾向は認めらないと結論づけたが，リツイート以外の要素に関しても分析を行い，それらについて言語毎に固有の傾向が見られるか検討したい．

謝辞

本論文作成にあたり，多大なるご指導をいただいた指導教員の塩田茂雄先生に深く感謝いたします．また，本研究に多大なるお力添えをいただきました塩田研究室の皆様にも深く感謝いたします．

参考文献

[1] Twitter,Inc(2019) 'Q4 and Fiscal Year 2018 Letter to Shareholders'. https://s22.q4cdn.com/826641620/files/doc_financials/2018/q4/Q4-2018-Shareholder-Letter.pdf 2020 年 2 月 1 日アクセス

[2] 総務省情報通信政策研究所, 平成 30 年度 情報通信メディアの利用時間と情報行動に関する調査報告書,https://www.soumu.go.jp/main_content/000644166.pdf,2020 年 2 月 1 日アクセス

[3] 塩田茂雄, 中島圭佑" Twitter データに見られる特徴と人間のリツイート行動," 日本人工知能学会全国大会, 2019.

[4] S. Shioda and M. Minamikawa, "Features found in Twitter data and examination of retweeting behavior," The International Workshop on Sentiment Analysis and Mining of Social Networks (SAMSN 2019), 2019.

[5] 南川雅人, 中島圭佑, 塩田茂雄, "Twitter データの特徴分析と人間の行動モデル," 電子情報通信学会 ネットワークシステム研究会, 2019

[6] 塩田茂雄, 南川雅人, 中島圭佑, "キーワード検索で収集される Twitter データの特徴と Twitter 上の情報拡散過程," 電子情報通信学会 情報ネットワーク研究会, IN2018-64, pp. 31-36, 2018