CO-002

通信状況の可視化によるネットワーク障害の早期検出・分析
Network Failure Detection, Analysis, and its Visualization

桂康洋 †
Yasuhiro Katsura
内田真人 †
Masato Uchida

1. はじめに

ネットワーク技術の発展、普及に伴う利用者の増加から、通信障害が社会生活に与える影響が深刻化している。総務省の調査 [1] によると、平成 29 年度に報告された通信障害の総件数は 6,205 件であり、そのうち通信障害の継続時間が 2 時間以上に達したものは 6,155 件 (総件数の 99%) であった。このような通信障害による影響を最小限に抑えるためには、ネットワーク管理者による迅速な復旧作業が急務である。しかし、通信規模の拡大、通信機器のマルチベンダー化、仮想化技術の発展等によりネットワーク管理が複雑になった昨今では、通信障害の検出や分析を早期に行うことは難しい。その結果、管理者による正式な情報公開には多くの時間を要する場合もある。その間、一般利用者は、周囲と情報を共有することでしか，通信障害の状況を把握することができない。したがって，可能な限り早期に通信障害の発生を検知・分析するために、管理者や一般利用者が障害状況を容易に確認できるシステムが必要である。

通信障害発生時の情報共有ツールとして、昨今では Twitter が多用される。Qui ら [2] は、通信障害発生時には通信状況の共有を目的としたツイート数が急増する傾向にあることを示した。本論文では、通信障害に言及しているツイートを元に通信状況をリアルタイムに可視化するシステムを構築する。これにより，ネットワーク管理者や利用者による容易な障害検知、および原因分析を可能にする。また、2017 年度に発生した 4 件の通信障害の状況を可視化することにより、本論文で構築したシステムの有用性について検証した。

本論文で構築したシステムでは、通信障害に言及するツイートに多く含まれる傾向にある単語を検索ワードとし、ツイートを収集する。そして、収集したツイートの集合から機械学習によって通信障害に関連したものを抽出し、それらを元に単位時間毎に発信されたツイート数の推移グラフ、および Word Cloud[3] によって通信状況を可視化する。2017 年度に発生した通信障害を可視化すると、通信事業者によって発表された障害期間前後でも多くのツイートが観測された。また、Word Cloud による可視化では、推移グラフよりも早期に異変を表示できる傾向にあることがわかった。

本論文の構成は以下のとおりである。第 2 節では関連研究について説明する。第 3 節では本論文の提案手法について述べる。第 4 節では構築した可視化システムの評価を行う。最後に第 5 節では本論文の結果、結論をまとめる。

†早稲田大学 基幹理工学研究科 情報理工・情報通信専攻
Department of Computer Science and Communication Engineering, Waseda University

2. 関連研究

Qui ら [2] は、通信障害が発生した際に発信されるツイートと通信事業者に対するカスタマーレビューを元に、ネットワークの利用状況を分析した。単位時間当たりの通信障害に言及したツイート数の増減が障害報告件数の増減におおむね従っている。このことから、通信障害に言及しているツイートからネットワークの利用状況を分析できることが示されている。また、ツイート数の増減の観測から、事業者が報告しないような小規模、もしくは短時間の通信障害を検知することが可能であることが示されている。さらに、通信障害が発生してから数分後という早い段階で通信障害に言及するツイートが見られる場合もあることから、カスタマーレビューよりも早い段階で通信障害を検知できることが明らかにされている。以上より、通信障害に関連するツイート数の増減から通信状況を分析できることがわかる。

また、竹下ら [4]、 [5] は単位時間当たりに発信されたツイート数を元に通信障害を自動検知する手法を提案した。この手法では，特定の検索ワードを含むツイートをリアルタイムに収集し、単位時間当たりのツイート数がある閾値を超えた場合に通信障害が発生したとみなしている。ただし、キーワードフィルタリングのみでは通信障害に言及していないツイートも誤って収集してしまうため、One Hot ベクトルを特徴量とする機械学習によって通信障害に関連しないツイートを除外している。そして、キーワードフィルタリングのみによる手法と機械学習を応用した手法を比較し、前者の手法では誤検知が多く見られることが示した。

これらの先行研究により、キーワードフィルタリングによるツイートの収集、機械学習を用いた無関連ツイートの排除による通信障害の検知が可能であることがわかる。しかし、障害内容によってツイート数は大きく異なるため、全ての通信障害に対して同一の閾値を設け、これを検知の判断基準とすることが適切でない場合も考えられる。例えば、インターネットの不通、データ通信の利用不可といった障害内容の場合、多くのユーザーが被害状況を Twitter 上で共有する傾向にある。一方で、音声通話の利用不可といった障害内容に言及するツイートの数は比較的少ない。同様に、通信障害が発生した時間帯もツイート数に大きな影響を与えると考えられる。実際、午前 2〜6 時には移動通信トラヒック数が少ないことが総務省 [6] により報告されており、この時間帯に通信障害が発生した場合、発信される関連ツイート数が減少するものと考えられる。このように、通信障害ごとに観測されるツイート数は条件によって大きく変動する。

そこで本論文では、閾値を超過するか否かにより通信障害を検知するのではなく、平常時との差異を容易

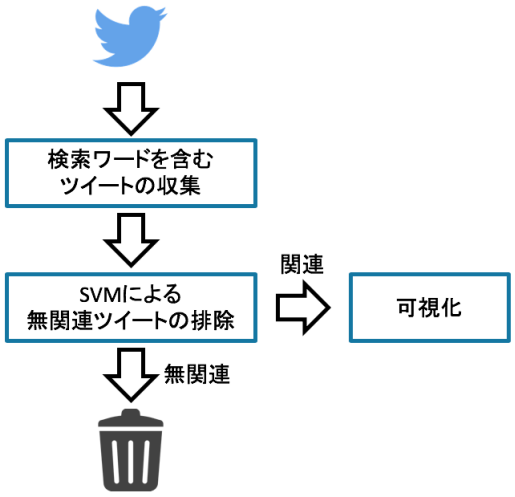


図 1: 提案手法の概略

に視認できる可視化システムの構築を行う。これにより、閾値による機械的な障害検知ではなく、ネットワーク管理者や利用者がシステムから得られた情報を元に障害分析を行うための一助となることを目的とする。

3. 提案手法

3.1. 提案手法の概要

本論文で提案する手法の概略を図 1 に示す．本論文では、関連研究 [4] と同様に特定の検索ワードを含むツイートを収集し、誤って収集した無関連ツイートを SVM (Support Vector Machine) によって排除する。その上で、SVM により排除されなかったツイート、すなわち SVM により通信障害の発生に関連すると判断されたツイートを可視化する手法を提案する。本論文では、単位時間毎に発信される通信障害に関するツイート数の推移グラフ、および Word Cloud による頻出名詞の表示、という 2 通りの可視化を行う。

推移グラフを用いた可視化では、5 分毎に得られた単位時間毎のツイート数の経時変化を表示する。平常時のツイート数はある一定の値を保っているが、通信障害が発生した際には大きく増加し、ネットワークの復旧に従い小さな値へと収束すると考えられる。よって、ツイート数の増減の監視を行うことで、通信障害の発生を容易に行えると考えられる。

また、Word Cloud を用いた可視化では、SVM により通信障害に関連すると判断されたツイートから名詞のみを抽出し、それらのうち出現頻度が高い名詞が目立つように表示する。頻出する名詞は障害内容に依存すると考えられるため、表示される名詞の傾向から大まかな障害状況を把握できると期待される。通信障害の影響を受けたサービス名や地域といった情報を表示することで、ネットワーク管理者や利用者による障害内容の推定の補助となることを目的とする。

3.2. 無関連ツイートの排除

“通信”、“回線” といった検索ワードでフィルタリングされたツイートが必ずしも通信障害に言及しているとは限らない。例えば、通信教育について言及したツイートは “通信” という単語は含むが、通信障害には言及していない。そこで、2 種類のベクトル表現による特徴量を用いた SVM 分類器を生成し、収集したツイートを分類する。本論文では、ツイートに含まれる単語の One Hot ベクトル、分散表現ベクトル [7] に変換する 2 通りの手法を試み、分類器の精度を比較した。なお、本論文中では、機械学習によって関連性があると分類されたツイートの集合を Class1、関連性なしと分類されたツイートの集合を Class0 と定義する。

3.2.1. データセット

通信障害の内容が異なる 3 件の大規模な通信障害 X、Y、Z に言及したツイートを別途収集した。本論文では、通信障害 X、Y、Z に言及したツイートが最も多く含む単語をそれぞれ 5 語ずつ選択し、それらをキーワードフィルタリングの検索ワードとして使用する。これらの単語を表 1 にまとめる。次に、2013 年度から 2015 年度に発信された検索ワードを含むツイートのうち、ランダムに 12,000 件のツイートを抽出する。また、これらの各ツイートに対し、通信障害に関連するか否かを表すラベルを著者の判断に基づき手動でつけた。全 12,000 件のうち、関連性を持つツイートは 1,807 件 (全体の約 15.1%) であった。ツイート t_i と、それに対して著者により付与されたラベル l_i の組 (t_i, l_i)、$(i = 1, 2, \cdots, n)$ からなる集合を学習用データセット D とする。ただし、$n = |D| = 12,000$ である。

なお、このようにして得られたデータセットをそのまま学習に使用した場合、多数派である Class0 に学習結果が傾いてしまうことが懸念される。そこで、SMOTE (Synthetic Minority Oversampling Technique) [8] によって両クラスのデータ数が均等になるように揃え、学習用データの均衡化を行なった。SMOTE とは、多数派ラベルを持つデータのアンダーサンプリング、および少数派ラベルを持つデータのオーバーサンプリングによりデータの均衡化を行う手法である。

3.2.2.One Hot ベクトルによる分類

各ツイートを、2 値の要素から構成される One Hot ベクトルへと変換する手順を述べる。

まず、学習用データセット D に含まれるツイートを構成する全ての単語の集合を W とする。次に、集合 W に含まれる単語のうち、分類精度に影響を及ぼすと考えられる単語の集合を W_c $(\subset W)$ とする。直観的には、“通信”、“障害”、“身体” といった単語は、ツイートが通信障害、もしくは身体障害に言及していると判断するための材料となり得るため、集合 W_c に含めるべきであると考えられる。しかし本論文では、より定量的に分類精度への貢献を評価した上で集合 W_c を決定するために、Filter Method[9] を採用した。Filter Method は各々の特徴量（本論文では単語に対応）が分類精度に与える影響力を、カイ二乗検定や ANOVA といった統計的手法で点数化し、点数が高い特徴量から順に N 個選択する。なお、分類精度への貢献の評価は各単語ごとに独立して行い、複数の単語の組み合わせによる評価は行わない。また、本論文中では Filter Method によ

表 1: 大規模通信障害 X、Y、Z に含まれた頻出名詞上位 5 語

通信障害	X	Y	Z
障害内容	データ通信の利用不可	メールの送受信不可	インターネットの不通
出現頻度 1 位	障害	メール	障害
出現頻度 2 位	LTE	障害	ネット
出現頻度 3 位	圏外	ダウン	接続
出現頻度 4 位	電波	エラー	回線
出現頻度 5 位	3G	ログイン	復旧

る特徴量の点数化の指標として ANOVA を用いる。最後に、ツイートの文中に集合 W_c に含まれる各単語が含まれるか否かによって、One Hot ベクトルへと変換する。したがって、このベクトルは $|W_c|$ 次元となる。

3.2.3. 分散表現ベクトルによる分類

本論文では、前節にて説明した One Hot ベクトルに加え、分散表現ベクトル [7] によってツイートを表現する手法も検討する。分散表現とは単語を 100〜200 次元のベクトルで表現する技術である。大容量のコーパス内で近い位置にある単語間の関係からモデルが構築されている。意味の類似した単語のベクトル間の距離が近くなるように設計されている。本論文では、ツイート文中の各単語の N 次元分散表現ベクトルを算出し、それらの各要素ごとの平均を求め、特徴ベクトルとする手法を試みる。文献 [10] で公開されている、日本語版 wikipedia 全文をコーパスとして作成された 200 次元の分散表現モデルを利用した。

4. 実験と考察

提案手法の評価実験を行い，その結果を考察する．ツイートの特徴ベクトルを入力とする分類器を SVM により作成し、分類精度を測定した。次に、2017 年度に発生した 4 件の通信障害を推移グラフ、および Word Cloud によって可視化し、それらを分析することで可視化システムの有用性について考察した。

4.1. 関連ツイートの分類器の性能評価

任意のツイートの One Hot ベクトルもしくは分散表現ベクトルを入力とし、通信障害への関連性を判断する分類器を SVM により作成した。それぞれの分類器について、10 交差検定と呼ばれる手法によって分類器の性能を評価する。学習用データを 10 分割し、そのうち 7 分割を訓練データとし、3 分割をテストデータとした。また、分類器の性能を評価するために Class0、Class1 のそれぞれに対する F 値を導出した。F 値とは分類器の信頼性を示す指標であり、Precision と Recall により以下のように求められる。

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recell} = \frac{TP}{TP+FN}$$

$$F = \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

なお、TP、FP、TN、FN はそれぞれ、True Positive、False Positive、True Negative、False Negative を表す。

One Hot ベクトルのサイズ N ($\leq$ 2,000)、および分散表現ベクトルのサイズ N ($\leq$ 200) を変化させた時の Class1、Class0 に対する F 値の変化を調査した。One Hot ベクトルを用いた場合、N が 450 以上であれば Class1、Class0 共に F 値は 0.90〜0.93 となった。これ対し、分散表現ベクトルを用いた場合、N が 120 以上であれば F 値は 0.90〜0.92 となった。両者を比較すると F 値に大きな差は生じないが、分散表現を用いた場合はるかに少ない特徴量で分類することができることがわかる。一方、ツイート文中に含まれる全ての単語ごとに分散表現ベクトルを算出する必要があるため、特徴ベクトル生成時に計算機にかかる負荷は One Hot ベクトルよりも分散表現ベクトルの方が大きい。

4.2. 可視化システムの性能評価

2017 年度に発生した 4 件の通信障害 A、B、C、D の発生期間中に発信されたツイートより、可視化システムの評価実験を行う。通信障害 A、B、C、D の詳細を表 2 に示す [1]。通信障害が発生している期間とそうでない期間の様子を比較するため、各通信障害の発生時間の 5 時間前から、収束時間の 5 時間後までの期間に発信されたツイートを収集し分析した。以後、通信障害 A、B、C、D が発生していた期間をそれぞれ P_A、P_B、P_C、P_D と定義する。同様に、通信障害 A、B、C、D が発生する 5 時間前から障害収束の 5 時間後までの期間をそれぞれ P'_A、P'_B、P'_C、P'_D と定義する。また、評価実験では 1,000 次元 One Hot ベクトルを入力とする SVM 分類器によってツイートを分類した。

4.2.1. 単位時間毎のツイート数推移グラフの評価

期間 P'_A、P'_B、P'_C、P'_D に収集された検索条件を満たす全ツイートついて、5 分毎に Class1 に分類されたツイート数の推移グラフをそれぞれ図 2a、図 2b、図 2c、図 2d に示す。なお、赤で示している部分が期間 P_A、P_B、P_C、P_D を表し、青で示している部分が通信障害が発生していない期間を表している。

図 2a、図 2c、図 2d を見ると、障害期間付近で 5 分毎の関連ツイート数が急増する傾向にあることがわかる。一方で図 2b に着目すると、平常時と比較して大きくツイート数が増加したということはなく、その変化はわずかである。表 2 にある通り、通信障害 B の継続時間は 139 分と短く、影響利用者数も 84,774 人と比較

的少ない。このことから、規模が小さい通信障害の場合は、明確な差異が観測できない可能性があると考えられる。

次に図 2a に着目すると、期間 P_A 以前からすでに多くの Class1 ツイートが観測される。期間 P_A 以前、および期間 P_A 内に観測されたすべてのツイートを確認、比較したところ、両者ともに LTE の利用不可やスマートフォンゲームの接続不良に関するツイートが多く見られた。期間 P_A 以前、および期間 P_A 内に確認された全 Class1 ツイートの一部を表 3 に示す。期間 P_A 以前、期間 P_A 内の Class1 ツイートが類似した障害内容に言及していることから、公式発表された障害期間以前から通信品質の低下等の何らかの状態変化が発生していた可能性があると考えられる。

また、図 2c、図 2d に着目すると、通信障害発生後にツイート数が増加し、平常時との明確な差異が視認できるようになるまでに約 250 分以上を必要としている。このように、平常時からの差異の確認に長時間を要する場合には、ツイート数の推移状況のみによる早期検出は難しいことがあると考えられる。さらに、図 2d において障害収束後にもツイート数が高い水準で推移していることがわかる。これは周囲との復旧情報の共有や、復旧したと公式に発表されているにも関わらず何らかの理由で接続不良を起こしている利用者によるツイートが原因だと考えられる。

以上のことから、障害規模が十分に大きければ単位時間毎のツイート数の急激な増加から通信状況を視認できることがわかる。通信事業者が発表した障害期間と本システムが示す障害期間が必ずしも正確に一致するとは限らないため、両者を比較することは通信障害の分析の一助となり得る。一方で、平常時との差異が生じるまでに長時間を要する場合がある。

4.2.2.Word Cloud の評価

期間 P'_A、P'_B、P'_C、P'_D について、SVM 分類器によって Class1 に分類されたツイートから名詞のみを抽出し、それらを Word Cloud で可視化した。出現頻度が高いほど目立つように表示することで、通信障害の大まかな内容や原因を視覚的に捉えることを目的とする。

通信障害 A、B、C、D について、障害期間前、障害期間内に Word Cloud 上に表示された名詞群をキャプチャしたものを図 3 に示す。これらの図を比較すると、“エラー”、“回線” といった名詞は全障害期間中で頻出することがわかる。また、通信障害 C の障害内容はネットワークの利用不可であるが、図 3f に注目すると、“ネット” という名詞が頻出していることがわかる。実際にデモ動画を撮影し、通信障害 C の Word Cloud の移り変わりを観察した。障害発生から数十分後には “ネット” をはじめとする障害内容毎に固有の名詞が多く出現し、平常時とは明らかに異なる様子を示した。前項の図 2c において示した通り、推移グラフでは通信障害の認識に約 250 分を要したが、Word Cloud 上ではそれよりもはるかに早い段階で平常時との差異を視認することに成功した。同様に、通信障害 D の障害内容は音声通話の利用不可であるが、“電話”、“固定 (電話) ” といった名詞が多く出現していることが確認できる。通信障害 C と同様にデモ動画を撮影、観察したところ、非常に早い段階で平常時との差異が確認できた。一方で、通信障害 B の障害内容はメールの送受信不可であったため “メール” といった名詞が頻出すると期待されたが、前項にて述べたように通信障害 B に言及しているツイートの母数が比較的少数であったため、平常時と比較しても大きな変化は見られなかった。

以上より、Word Cloud 上に障害内容に関連した名詞が多く出現する傾向にあり、単位時間毎のツイート数が急増するより早い段階で、表示状態に変化が生じる場合があることがわかった。

5. まとめ

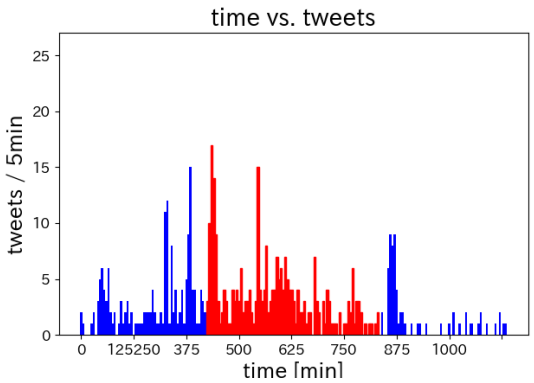
本論文では、通信障害の可視化により、ネットワーク管理者、利用者による通信状況の監視を容易にする手法を提案した。本提案手法においては、通信障害に言及したツイートをキーワードフィルタリング、および機械学習によって収集する。そして、得られたツイートから推移グラフ、Word Cloud を作成し、通信状態の経時変化を可視化する。

まず、One Hot ベクトル、および分散表現ベクトルを入力とするツイートの分類器を生成し精度比較を行った。その結果、分散表現を用いた場合は特徴ベクトルの生成に計算コストを必要とする一方で、比較的低次元な特徴ベクトルによって高精度の分類が行えることがわかった。次に、2017 年度に発生した 4 件の通信障害を実際に可視化、分析することで可視化システムの性能評価を行った。推移グラフを観測することで、通信障害発生時に単位時間あたりのツイート数が急増する傾向にあることが確認できた。ただし、公式に発表された障害期間と可視化システムにおいて状態変化が観測される期間が一致するとは限らない。特に公式発表された期間以前から可視化システム上で何らかの状態変化が観測される場合、両者を比較することにより新たな知見が得られると期待される。また、Word Cloud、推移グラフによる通信状況監視を比較したところ、前者の方がより早期に平常時との差異を表現することができた。

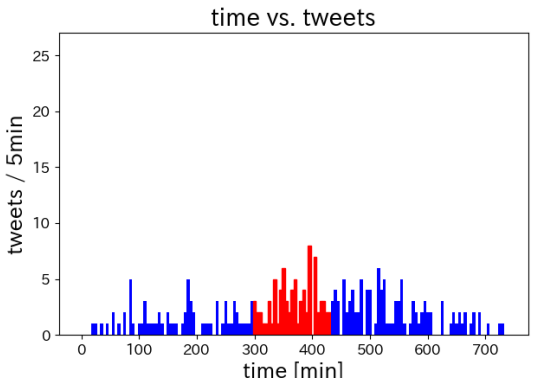
一方で、継続時間、影響利用者数が比較的小さい通信障害においては、平常時との差異を十分に可視化できないことがわかった。対策として、Twitter 以外の情報共有ツールに応用することがあげられる。通信障害時に取得できるデータの総数が増えるため、小規模な通信障害の可視化や、より正確な通信状況の可視化が実現できると期待される。また、本論文では国内で発生する全通信障害を可視化の対象としたが、 通信事業者名やサービス名で可視化対象を絞るといった利便性の追求も今後の課題としてあげられる。

謝辞

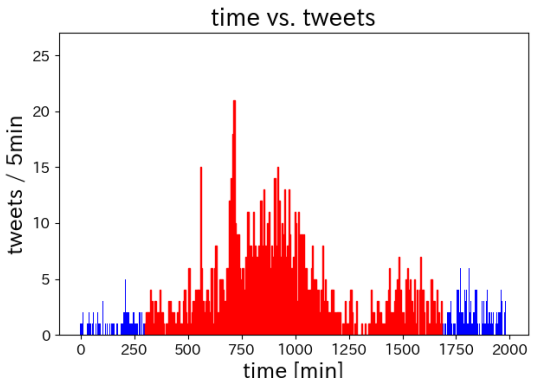
本研究では、東北大学 乾・岡崎研究室が公開している訓練済み分散表現モデルを利用した。また、本研究の一部は、日本学術振興会における科学研究費補助金



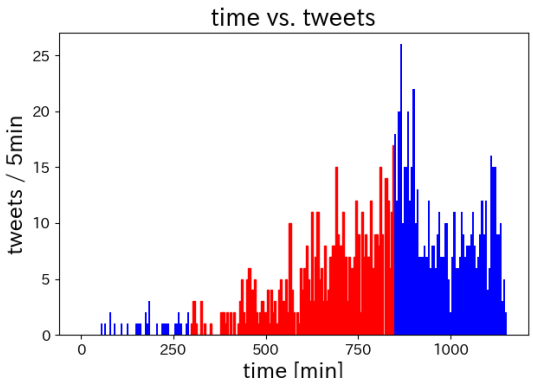
(a) 通信障害 A の 5 時間前後に発信されたツイート数の推移



(b) 通信障害 B の 5 時間前後に発信されたツイート数の推移



(c) 通信障害 C の 5 時間前後に発信されたツイート数の推移



(d) 通信障害 D の 5 時間前後に発信されたツイート数の推移

(a) 通信障害 A 発生前

(b) 通信障害 A 発生期間中

(c) 通信障害 B 発生前

(d) 通信障害 B 発生期間中

(e) 通信障害 C 発生前

(f) 通信障害 C 発生期間中

(g) 通信障害 D 発生前

(h) 通信障害 D 発生期間中

図 3: 通信障害発生前、発生期間中の Word Cloud のキャプチャ

表 2: 2017 年度に発生した 4 件の通信障害の詳細 [1]

	A	B	C	D
事業者名	楽天コミュニケーションズ (株)	(株) 朝日ネット	(株) ジェイコムウェスト	ソフトバンク (株)
発生時刻	2017 年 4 月 7 日 19:53	2017 年 4 月 13 20:06	2017 年 7 月 3 日 11:50	2018 年 2 月 19 日 09:30
収束時刻	2017 年 4 月 8 日 02:45	2017 年 4 月 13 22:25	2017 年 7 月 4 日 10:58	2018 年 2 月 19 日 18:44
継続時間 (min)	412	139	1,388	554
障害内容	データ通信の利用不可	メールの送受信不可	インターネットの不通	音声通話の利用不可
影響利用者数 (人)	220,300	84,774	52,792	670,000
Class1 ツイートの数	504	258	1,278	1,143
収集したツイート数	119,335	94,457	230,146	159,976

表 3: 通信障害 A 発生前、発生中のツイート例

通信障害発生前	しばらく電波繋がらんやったけどやっと通じた 日田駅なうンゴ
	あらぁ?? Wi-Fi の接続切れてる？と思って ON にしても繋がらないんだけどなんだ???
	勝ったのに通信エラーで下がるとかマジクソですよね、(言葉悪くてすみません)
	あれれー？おかしいなー？また通信障害が起きてるみたいだー
	森のホール、非常に電波状態が悪い。アンテナ立っていても繋がらない
通信障害発生中	通信エラー起こした… 何が悪いんだろ
	電波はつかんでるけど通信できないくそ状態。 LINE ができない。
	ネット繋がらねぇ!! モデムの不具合ってどういう事だよぉ!!
	楽天モバイル、繋がらないと思ったら通信障害か。Wi-Fi がある今なら大丈夫だけど、明後日までに直ってもらわないと困るなぁ
	首都圏で大規模な通信障害が起きてるよーな噂を聞いたけど、何がソースなんだろうか。

基盤研究 (B)(課題番号 17H01742) による支援を受けている。ここに記し謝意を表す。

参考文献

[1] 総務省. **電気通信サービスの事故発生状況** *(平成 29 年度)*. http://www.soumu.go.jp/main_content/000499953.pdf.

[2] Tongqing Qiu et al. “Listen to Me if You Can: Tracking User Experience of Mobile Network on Social Media”. In: *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*. IMC ’10. Melbourne, Australia: ACM, 2010, pp. 288–293.

[3] *WordCloud for Python documentation — wordcloud 1.5.0.post34+g64ff55e documentation*. https://amueller.github.io/word_cloud/.

[4] K. Takeshita, M. Yokota, and K. Nishimatsu. “Early network failure detection system by analyzing Twitter data”. In: *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. 2015.

[5] M. Yokota, K. Takeshita, and K. Nishimatsu. “Demonstration of carrier network failure detection by analyzing Twitter”. In: *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. 2015.

[6] 総務省. **総務省|平成 *24* 年版 情報通信白書**. http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/html/nc245320.html.

[7] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[8] Nitesh V. Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *J. Artif. Intell. Res.* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953. URL: https://doi.org/10.1613/jair.953.

[9] Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, and María Tombilla-Sanromán. “Filter Methods for Feature Selection – A Comparative Study”. In: *Intelligent Data Engineering and Automated Learning - IDEAL 2007*. Ed. by Hujun Yin et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 178–187. ISBN: 978-3-540-77226-2.

[10] 東北大学 乾・岡崎研究室. **日本語 *Wikipedia* エンティティベクトル**. http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/.