

# ソーシャルボットの検出：言語非依存性の特徴量と ボット集団の定量化

Detection of Social Bots: Quantification of Language Independent Features and Bot Clusters

杉森 真樹 \*<sup>1</sup> 笹原 和俊 \*<sup>1</sup> \*<sup>2</sup> 時田 恵一郎 \*<sup>1</sup>  
Masaki Sugimori Kazutoshi Sasahara Kei Tokita

\*<sup>1</sup> 名古屋大学大学院情報学研究科 \*<sup>2</sup> JST さきがけ  
Graduate School of Informatics, Nagoya University JST, PRESTO

With the advent of social media, social bots—computer programs that automatically post news—have become popular, and some of these have been purposefully used to disseminate fake news. This is an increasingly serious problem in the digital era. Although a lot of bot detection methods have been proposed, such methods often cannot overcome their language dependencies and fail to find bot clusters. This research aims at constructing a language-independent method for detecting social bots and their clusters on Twitter using machine learning. The results provide new information available for the development of such methods.

## 1. 導入

近年, Twitter や Facebook といったソーシャルメディアが普及し, 単なるコミュニケーションのツールにとどまらず, 様々な目的に利用されるようになり, 社会において重要な役割を果たすようになってきている. このような背景のもと, コンピュータープログラムによって自動化された「ソーシャルボット」と呼ばれるアカウントによって引き起こされる様々な悪影響が大きな問題となっている [Ferrara 16].

ソーシャルボットが引き起こす問題には次のようなものがある.

- **スパムの拡散**: 従来のスパムと同様に, 一般ユーザーに対して有害な Web ページへのリンクをクリックさせ, 特定の Web ページにアクセスさせたり, マルウェアに感染させることを目的とする.
- **人気の操作**: ソーシャルボットの投稿によってソーシャルメディアのトレンド (話題となっているキーワード) が影響を受ける可能性がある. また, 特定のアクティビティを集中的に行うことで, 特定のユーザーや商品などが人気, 影響力があるように見せかけることができる. これは, マーケティングなどに悪影響を及ぼすことが考えられる.
- **意見の操作**: ソーシャルボットによるフェイクニュースなどの拡散により, 特定の対象に関する人々の印象を操作しようとする可能性がある. 具体的な例としては選挙が挙げられ, 2016 年のアメリカ合衆国大統領選挙では, フェイクニュースなどが大量に発信されており, 選挙に影響を及ぼしていた可能性がある [Bessi 16].
- **市場の操作**: ソーシャルボットによるフェイクニュースの拡散により, 株式市場などの動向が操作される可能性がある. 2013 年には, テロに関するフェイクニュースによって, 一時的にダウ工業株 30 種平均が約 140 ドル急落した. また, ほとんど資産価値のない企業の株価を高騰させようとするソーシャルボットの動きが確認されている [Ferrara 16].

このような様々な悪影響をもたらすソーシャルボットを検出して, ユーザーに注意を喚起し, ソーシャルメディアから排除するために, ボットの特性や検出手法, 情報の拡散過程などについて様々な研究を行っていく必要がある.

## 2. 関連研究

ソーシャルメディアの普及と発展に伴い, ボットがもたらす悪影響は大きな問題を引き起こしており, 様々な先行研究が行われている.

### (1) 単体のアカウントを対象とした観察的手法

Varol らは, Twitter の英語アカウントを対象として, 1150 個の特徴量と機械学習を用いて, AUC 値が 0.94 という高精度な分類モデルを構築した. 彼らは, この分類モデルを応用して, アカウントがボットであるかどうかのスコアを [0~1] の範囲で算出する “Botometer” というシステムを開発し, 公開している [Varol 17, Davis 16]. Chu らは, 半自動化されているアカウントを “サイボーグ” と定義し, 機械学習を用いて, 人間, ボット, サイボーグの 3 クラスの分類に取り組んでいる [Chu 10, Chu 12].

### (2) ボットの集団を対象とした観察的手法

Abokhodair らは, シリア内戦に関する情報を拡散しているボットの集団を対象として, ボットの集団の構造や成長過程, ソーシャルメディア上の議論に与えた影響などを調査した [Abokhodair 16]. Chavoshi らは, アカウント同士の過度に同期的な活動に注目して, 学習用のデータを必要としないボット検出手法を提案するとともに, “DeBot” というボット検出システムを開発, 公開している [Chavoshi 16a, Chavoshi 16b, Chavoshi 17a, Chavoshi 17b].

### (3) 単体のアカウントを対象とした実験的手法

Freitas らは, それぞれ異なる特徴と戦略を持つ 120 個のボットアカウントを作成し, これらが Twitter 上のコミュニティに侵入する過程や結果について調査を行った. その結果, 単純なメカニズムを採用しているボットであってもコミュニティに侵入することができ, 必ずしも複雑なメカニズムを必要としないことを示した [Freitas 15].

連絡先: 杉森 真樹, 名古屋大学大学院情報学研究科,  
〒464-8601 名古屋市千種区不老町, sugimori@nagoya-u.jp

#### (4) ボットの集団を対象とした実験的手法

Boshmaf らは、ボットの集団による活動を想定して、102 個のボットアカウントと 1 個のマスターアカウントからなるボットの集団のプロトタイプを構築し、これを Facebook 上で稼働させて、ボットの集団がもたらす影響とソーシャルメディアの脆弱性を調査した [Boshmaf 11, Boshmaf 13].

ソーシャルボットの性質の調査や検出方法については様々な先行研究が行われている。しかし、ソーシャルメディアには世界中に多くのユーザーが存在し、多様な言語が利用されているにもかかわらず、ボット検出に関する研究において、言語の違いに注目した研究はほとんど行われていない。また、単体のボットよりも大きな影響を及ぼす可能性があるボット集団を検出する手法の研究はほとんど行われていない。

そこで本研究では、Varol らによる Botometer に関する研究 [Varol 17] を主に参考にして、Twitter を対象として英語と日本語それぞれの言語において、ボット検出に関連する重要な特徴量について比較、分析を行う。そして、言語に依存しないユニバーサルなボット検出手法について検討を行う。さらに、特徴量の組の分布に注目して、ボット集団を対象とする機械学習を用いた検出手法について検討を行う。

### 3. データ

本研究では、Twitter アカウントが持つ様々な特徴量と機械学習を用いて、分類モデルを構築する。分類モデルを訓練するために、学習用のデータとして、各アカウントのメタデータと連続した一定量のツイートから算出した特徴ベクトルと人間かボットかを事前に判別した正解ラベルが必要となる。先述した Botometer による分類が正しいと仮定し、正解ラベルの作成には Botometer を利用した。日本語アカウントの分類には、言語的な特徴を無視したスコアを利用した。

データを収集するにあたり、英語または日本語で投稿を行なった Twitter アカウントの最新 200 件の投稿を収集し、特徴ベクトルと正解ラベルを算出した。最新の投稿 200 件を対象とするのは、API 制限のためであり、先行研究においても最新の 200 件以降の投稿は対象とされていない。2017/12/7~2018/1/4 の期間、英語についてはボット 12193 件と人間 73616 件、日本語についてはボット 8951 件と人間 63866 件のアカウントのデータを収集した。

英語については、ボット 12193 件と無作為抽出した人間 12193 件の合計 24386 件を用いて、言語ごとの分類モデルを構築する。日本語については、ボット 8951 件と無作為抽出した人間 8951 件の合計 17902 件を用いて、言語ごとの分類モデルを構築する。

### 4. 結果

#### 4.1 分類モデル

本研究では、先行研究 [Varol 17, Chu 10, Chu 12] で利用されている特徴量と独自の特徴量の中から、次のリストに示した 4 つの特徴クラス、326 個の特徴量を利用する。

分類モデルの構築には “scikit-learn” [Pedregosa 11] を利用し、機械学習のアルゴリズムはランダムフォレストを採用する [Breiman 01]。これは、先行研究においてランダムフォレストを利用した場合の精度が最も高いことに加え [Varol 17, Lee 11], 特徴量の重要度を導出することができるためである。

英語と日本語それぞれについて、アカウントが人間であるかボットであるかを判別する分類モデルを特徴クラスごとに構築する。精度の指標には AUC を採用する。

#### 1 コンテンツクラス

- 1-1 品詞の出現頻度<sup>1,2</sup>
- 1-2 品詞の割合<sup>1,2</sup>
- 1-3 単語の総数<sup>1</sup>
- 1-4 単語の種類の数<sup>1</sup>
- 1-5 単語のエントロピー<sup>1</sup>

#### 2 メタデータクラス

- 2-1 スクリーンネーム (登録名) の長さ
- 2-2 スクリーンネームに含まれる数字の個数
- 2-3 タイムオフセット
- 2-4 アカウントの日齢
- 2-5 自己紹介文の長さ
- 2-6 自己紹介文中に URL を含むか
- 2-7 フォロー数
- 2-8 フォロワー数
- 2-9 お気に入りの数
- 2-10 プロファイルがデフォルトのままか
- 2-11 プロファイルイメージがデフォルトのままか
- 2-12 投稿数 (総数, 1 時間あたりの数)
- 2-13 通常のツイート数 (最新 200 件) (総数, 1 時間あたりの数)
- 2-14 リツイート数 (最新 200 件) (総数, 1 時間あたりの数)
- 2-15 メンション数 (最新 200 件) (総数, 1 時間あたりの数)
- 2-16 リプライ数 (最新 200 件) (総数, 1 時間あたりの数)
- 2-17 リツイートされた数 (最新 200 件) (総数, 1 時間あたりの数)
- 2-18 リツイートされた数 (最新 200 件)<sup>1</sup>
- 2-19 お気に入りされた数 (最新 200 件)<sup>1</sup>
- 2-20 メンションした数 (最新 200 件)<sup>1</sup>
- 2-21 URL 数 (最新 200 件)<sup>1</sup>
- 2-22 ハッシュタグ数 (最新 200 件)<sup>1</sup>
- 2-23 メディア数 (画像数+動画数) (最新 200 件)<sup>1</sup>
- 2-24 画像数 (最新 200 件)<sup>1</sup>
- 2-25 動画数 (最新 200 件)<sup>1</sup>
- 2-26 投稿元に関する指標 (最新 200 件)<sup>1,3</sup>
- 2-27 フォロー比
- 2-28 フォロワー比
- 2-29 リストされている数

#### 3 タイミングクラス

- 3-1 通常のツイートの投稿時間間隔 (分)<sup>1</sup>
- 3-2 リツイートの投稿時間間隔 (分)<sup>1</sup>
- 3-3 リプライの投稿時間間隔 (分)<sup>1</sup>
- 3-4 メンションの投稿時間間隔 (分)<sup>1</sup>
- 3-5 投稿の時間間隔 (分)<sup>1</sup>

#### 4 ネットワーククラス (ハッシュタグ共起ネットワーク)

- 4-1 ノード数
- 4-2 エッジ数
- 4-3 ネットワークの密度
- 4-4 ネットワークの強度<sup>1</sup>
- 4-5 クラスタリング係数<sup>1</sup>

#### 注釈

<sup>1</sup> [最小値, 最大値, 中央値, 平均, 標準偏差, 歪度, 尖度, エントロピー] を算出し、特徴量とする。

<sup>2</sup> 品詞は、英語 [名詞, 動詞, 形容詞, 副詞, 助動詞, 限定詞, 間投詞, Wh-, 代名詞], 日本語 [名詞, 動詞, 形容詞, 副詞, 助動詞, 助詞, 接統詞, 感動詞, 接頭詞] を考慮する。

<sup>3</sup> 投稿元に関する指標については、各投稿において、自動化が行われていないと思われる投稿元の値を 1, 自動化が行われていると思われる投稿元の値を -1, それ以外の値を 0 とする。

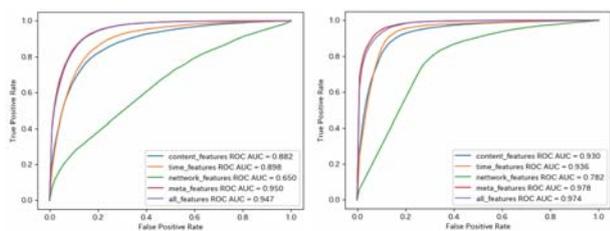


図 1-a : 分類モデルの精度 (英語)

図 1-b : 分類モデルの精度 (日本語)

図 1-a と b を見ると、英語と日本語の双方のアカウントにおいて、メタデータクラスが最も高い精度を示しており、コンテンツクラスとタイミングクラスがこれに続いていることが分かる。一方で、ネットワーククラスはある程度の精度を示しているが、他のクラスと比較すると精度が低いことが分かる。

また、コンテンツクラスの特徴量よりもメタデータクラスとタイミングクラスの特徴量が高い精度を示していることから、言語に依存しないユニバーサルな分類モデルにおいて、コンテンツクラスの特徴量を考慮せずとも高い精度を示すことは可能であると考えられる。

## 4.2 重要な特徴量

主に 2 つの手法を用いて、英語と日本語それぞれの分類モデルにおいて重要な特徴量の選択を行う。

### (1) 分類モデルにおける特徴量の重要度の測定

ランダムフォレストを用いた分類モデルにおける特徴量の重要度を、決定木ごとにその特徴量によって減少する基準値 (ジニ不純度) の平均として算出する。

### (2) 人間とボットの特徴量の分布が異なるかどうかの判定

コルモゴロフ・スミルノフ検定 (KS 検定) によって、人間とボットで特徴量の分布が異なることを確認する。

重要な特徴量を特定するための手順としては、分類モデルにおいて重要度が高い特徴量を 50 個選択し、その中で人間とボットで分布が異なっていることが確認できたものを重要な特徴量とする。

また、双方の言語の分類モデルで共通している重要な特徴量の中でも、2 つの言語を統合した場合でも、KS 検定によって人間とボットで分布が異なっていることが確認できた特徴量を言語に依存しない重要な特徴量とする。

## 4.3 言語非依存の分類モデル

言語に依存しないユニバーサルな分類モデルの構築について検討を行うために、英語と日本語について言語に依存しない分類モデルの構築を行う。また、この分類モデルが学習用のデータ以外に有用であるか検証を行う。

英語と日本語について言語に依存しない分類モデルの構築にあたり、言語に依存しない重要な特徴量の中で、後に行う分類モデルの検証において、精度を大きく低下させていた特徴量や、共通したデータやアクティビティに基づく特徴量の中で重要度が低い一部の特徵量は利用しないものとする。最終的には、次のリストに示した 20 個の特徴量を用いて分類モデルを構築する。

この分類モデルの構築には、学習用のデータから英語と日本語それぞれについて無作為抽出した、ボット 8000 件、人間 8000 件の計 32000 件を用いる。精度については、クロスバリデーション (10 分割) を利用して、平均 AUC を測定する。

結果として、この分類モデルの平均 AUC の値は約 0.95 となり、先行研究や全 326 個の特徴量を考慮した言語ごとの分類モデルとほぼ同等の精度を有することが確認できた。

- コンテンツクラス
  - 単語のエントロピーの標準偏差
- メタデータクラス
  - リプライの総数
  - リツイート of 総数
  - 被リツイート of 総数
  - 被リツイート数の最大値, 平均, 標準偏差
  - 被お気に入り数の平均
  - メンション数の標準偏差
  - URL 数の平均
  - ハッシュタグ数の平均
- タイミングクラス
  - リツイートの時間間隔の最大値, 中央値, 平均, エントロピー
  - リプライの時間間隔の最大値, 中央値, 平均, エントロピー
  - 全投稿の時間間隔のエントロピー

分類モデルが学習用のデータ以外のデータに対しても、同様の精度で分類を行えることを確認するために、検証用のデータとして英語と日本語それぞれについて手で正解ラベルを作成した、ボット 250 件、人間 250 件ずつ、計 1000 件のデータを用意する。(手動による分類は正しいと仮定する。)

検証の結果、1000 件のうち約 95% のアカウントを正しく分類できることが確認でき、英語と日本語について言語に依存しない分類モデルを構築できたことと結論づけることができる。

このように、英語と日本語について言語に依存しない分類モデルを構築することができたことから、他の言語についても同様の研究を行うことで言語に依存しないユニバーサルな分類モデルの構築は可能であると考えられる。

## 4.4 ボット集団

ソーシャルボットの検出に関する先行研究の多くは、単体のボットを対象とするものであった。しかし、ソーシャルボットの集団は相互作用することで、単体のボットよりも大規模な問題を引き起こす可能性がある。そのため、ボット集団の検出についても検討を行う必要がある。

ボット集団は、マスター (製作者または操作者) によって、共通の目的のために製作者または操作されることから、同一のコンピュータープログラムに基づいて活動していると考えられる。そのため、複数の特徴量が近い値をとることが予測される。本研究では、この複数の特徴量が近い値をとることに注目して、特徴量の組の分布を確認するとともに、クラスタリングを行うことで特徴量の組について分析を行い、ボットの集団が検出できるか検証を行う。なお、クラスタリングのアルゴリズムには、密度 (距離) に注目したアルゴリズムである DBSCAN [Ester 96] を採用する。

図 2 は、学習用データに含まれる全てのボットアカウントを対象とした、投稿に含まれる URL 数の平均と投稿間隔のエントロピーの 2 つの特徴量の組の散布図とクラスタリング (DBSCAN) の結果を示したものであり、特徴量の組の分布が集中している複数の領域が存在していることが確認できる。

これらの分布が集中している領域の中でも、本来は分布が集中しにくい領域がボット集団の存在を示唆していると考え、実際に確認を行った。その結果、日本の芸能、ゴシップニュースを拡散しているボットの集団が確認された (図 2: 青色)。

このように、特徴量の組の分布に注目してクラスタリング (DBSCAN) を行い、不自然に分布が集中している領域を特定することで、ソーシャルボットの集団を検出することができた。このことから、複数の特徴量に注目することで、ボット集団の検出が可能であると考えられる。

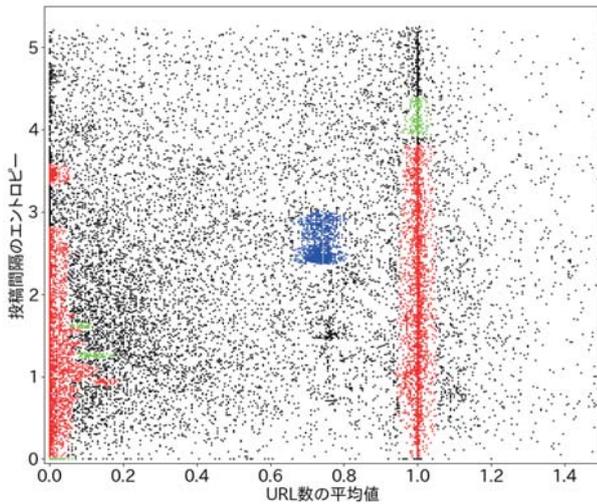


図 2: URL 数の平均と投稿間隔のエントロピー (散布図)

赤: 1000 個以上のアカウントからなるクラスター  
 緑: 1000 個未満, 50 個以上のアカウントからなるクラスター  
 青: 緑のクラスターの中でボットの集団の存在が確認できたもの

今後は, ボットの集団の検出に利用可能な特徴量やクラスターリングを行う上で最適なパラメーターなどについてさらなる調査を行うことで, 複数の特徴量とクラスターリングを用いたソーシャルボットの集団を対象とする検出手法の構築に取り組む。

## 5. 結論

ソーシャルメディアには世界中に多くのユーザーが存在し, 多様な言語が利用されているにもかかわらず, ソーシャルボットの検出手法において, 言語の違いに注目した研究が行われていなかった。また, ソーシャルボットの集団は相互作用し, 単体のボットよりも大規模な問題を引き起こす可能性があるにもかかわらず, ソーシャルボット集団の検出手法に関する研究がほとんど行われていなかった。

そこで本研究では, 言語に依存しないユニバーサルな分類モデルの構築について検討を行うとともに, ソーシャルボットの集団を対象とする新たな検出手法について検討を行った。結果として, 英語と日本語を対象とした分類モデルにおいて, 言語に依存しない重要な特徴量を特定し, これらの特徴量を利用して英語と日本語の双方の言語に対して精度の高い分類モデルを構築することで, 言語に依存しないユニバーサルな分類モデルを構築できる可能性を示した。また, 単体のソーシャルボットの検出手法で利用されている特徴量の組の分布に注目し, クラスターリングを行うことでソーシャルボットの集団を検出できることを確認した。

今後は, 言語に依存しないユニバーサルな分類モデルの構築に取り組むとともに, ソーシャルボットの集団を対象とする検出手法について, より詳細な調査と検討を行う。

## 謝辞

本研究は JSPS 科研費 (JP16K16112, JP17H06383), JST さきがけ (JPMJPR16D6), JST CREST (JPMJCR17A4) の助成を受けたものです。

## 参考文献

- [Abokhodair 16] Abokhodair, N., Yoo, D., and McDonald, D. W.: Dissecting a Social Botnet: Growth, Content and Influence in Twitter, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2016)
- [Bessi 16] Bessi, A. and Ferrara, E.: Social bots distort the 2016 U.S. Presidential election online discussion, in *First Monday* (2016)
- [Boshmaf 11] Boshmaf, Y., Musluhkov, I., Beznosov, K., and Ripeanu, M.: The Socialbot Network: When Bots Socialize for Fame and Money, in *Proceedings of the 27th Annual Computer Security Applications Conference* (2011)
- [Boshmaf 13] Boshmaf, Y., Musluhkov, I., Beznosov, K., and Ripeanu, M.: Design and Analysis of a Social Botnet, in *Computer Networks: The International Journal of Computer and Telecommunications Networking* (2013)
- [Breiman 01] Breiman, L.: Random Forests, in *Machine Learning* (2001)
- [Chavoshi 16a] Chavoshi, N., Hamooni, H., and Mueen, A.: DeBot: Twitter Bot Detection via Warped Correlation, in *IEEE 16th International Conference on Data Mining* (2016)
- [Chavoshi 16b] Chavoshi, N., Hamooni, H., and Mueen, A.: Identifying Correlated Bots in Twitter, in *International Conference on Social Informatics* (2016)
- [Chavoshi 17a] Chavoshi, N., Hamooni, H., and Mueen, A.: On-Demand Bot Detection and Archival System, in *Proceedings of the 26th International Conference on World Wide Web Companion* (2017)
- [Chavoshi 17b] Chavoshi, N., Hamooni, H., and Mueen, A.: Temporal Patterns in Bot Activities, in *Proceedings of the 26th International Conference on World Wide Web Companion* (2017)
- [Chu 10] Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S.: Who is Tweeting on Twitter: Human, Bot, or Cyborg, in *Proceedings of the 26th Annual Computer Security Applications Conference* (2010)
- [Chu 12] Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S.: Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?, in *IEEE Transactions on Dependable and Secure Computing* (2012)
- [Davis 16] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F.: BotOrNot: A System to Evaluate Social Bots, in *Proceedings of the 25th International Conference Companion on World Wide Web* (2016)
- [Ester 96] Ester, M., Kriegel, H. P., Sander, J., and Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996)
- [Ferrara 16] Ferrara, E., Varol, O., Davis, C. A., Menczer, F., and Flammini, A.: The Rise of Social Bots, in *Communications of the ACM* (2016)
- [Freitas 15] Freitas, C. A., Benevenuto, F., Ghosh, S., and Veloso, A.: Reverse Engineering Socialbot Infiltration Strategies in Twitter, in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2015)
- [Lee 11] Lee, K., Eoff, B. D., and Caverlee, J.: Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter, in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (2011)
- [Pedregosa 11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine and Learning in Python, in *Journal of Machine Learning Research Scikit-learn: Machine and Learning in Python* (2011)
- [Varol 17] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A.: Online Human-Bot Interactions: Detection, Estimation, and Characterization, in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (2017)