PONTIFICAL CATHOLIC UNIVERSITY OF MINAS GERAIS

Graduate Program in Informatics

Josemar Alves Caetano

CHARACTERIZING POLITICALLY ENGAGED USERS DURING THE 2016 US PRESIDENTIAL CAMPAIGN USING TWITTER

Belo Horizonte

2018

Josemar Alves Caetano

CHARACTERIZING POLITICALLY ENGAGED USERS DURING THE 2016 US PRESIDENTIAL CAMPAIGN USING TWITTER

Dissertation presented to Graduate Program in Informatics at Pontifical Catholic University of Minas Gerais, as partial requirement to obtain Master's degree in Informatics.

Advisor: Prof. Dr. Humberto Torres Marques Neto

Belo Horizonte

2018

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Caetano, Josemar Alves
C128c Characterizing politically engaged users during the 2016 us presidential campaign using Twitter / Josemar Alves Caetano. Belo Horizonte, 2018.
56 f. : il.

Orientador: Torres Marques Neto
Dissertação (Mestrado) - Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Twitter (Rede social on-line). 2. Emoções - Análise. 3. Campanha eleitoral - Redes sociais. 4. Candidatos a presidência - Redes sociais. 5. Interação social. I. Marques Neto, Torres. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 301.175.1
Ficha catalográfica elaborada por Rosane Alves Martins da Silva – CRB 6/2971

Josemar Alves Caetano

CHARACTERIZING POLITICALLY ENGAGED USERS DURING THE 2016 US PRESIDENTIAL CAMPAIGN USING TWITTER

Dissertation presented to the Graduate Program in Informatics as a partial requirement for qualification to the Master's degree in Informatics from the Pontifical Catholic University of Minas Gerais.

Prof. Dr. Humberto
Torres Marques-Neto – PUC Minas (Advisor)

Prof. Dr. Jussara
Marques de Almeida – UFMG (Examining Bank)

Prof. Dr. Thyago
Mota – Moravian College (Examining Bank)

Prof. Dr. Luis
Enrique Zárate - PUC MINAS (Examining Bank)

Belo Horizonte, February 26, 2018.

ABSTRACT

Political campaigns have frequently used the online social network as an important environment to exhibit the candidate ideas, their activities, and their electoral plans if elected. Some users are more politically engaged than others. As an example, we can observe intense political debates, especially during major campaigns on Twitter. In such context, this work presents a characterization of politically engaged user groups on Twitter during the 2016 US Presidential Campaign. Using a rich dataset with 23 million tweets, 115 thousand user profiles and their contact network collected from January 2016 to November 2016, we identified four politically engaged user groups: advocates for both main candidates, political bots, and regular users. We present a characterization of each group analyzing which features highlight each group, the mean sentiment analysis, language patterns, popular users, mood variation analysis, and homophily analysis. Our study contributes to a better understanding of the political engagement of users on online social networks, particularly during the 2016 US presidential campaign. We believe that the methodology proposed here can be replicated for identifying groups of advocates, bots, and regular users during other elections or major events in general. It also sheds some light on how candidates may influence their voters (particularly their mood) using a platform such as Twitter, thus suggesting how they can approach the public on online social networks and, using homophily, we analyze how Twitter users engaged in a political campaigns interact with each other.

Keywords: Twitter Characterization. Sentiment Analysis. American Presidential Election.

RESUMO

As campanhas políticas frequentemente usam as redes sociais online como um ambiente importante para exibir as idéias do candidato, suas atividades e seus planos eleitorais caso eleitos. Alguns usuários são mais envolvidos politicamente do que outros. Por exemplo, podemos observar intensos debates políticos, especialmente durante as principais campanhas no Twitter. Nesse contexto, este trabalho apresenta uma caracterização de grupos de usuários politicamente envolvidos no Twitter durante a Campanha Presidencial dos EUA de 2016. Usando um rico conjunto de dados com 23 milhões de tweets, 115 mil perfis de usuários e sua rede de contatos coletados de janeiro de 2016 a novembro de 2016, identificamos quatro grupos de usuários envolvidos politicamente: *advocates* de cada um dos candidatos principais, *bots* políticos e usuários normais. Apresentamos uma caracterização de cada grupo analisando quais características destacam cada grupo, sentimento médio, padrões de linguagem, usuários populares, variação de humor e a homofilia. Nosso estudo contribui para uma melhor compreensão do envolvimento político dos usuários em redes sociais online, particularmente durante a campanha presidencial dos EUA de 2016. Acreditamos que a metodologia aqui proposta pode ser replicada para identificar grupos de *advocates*, *bots* e usuários normais durante outras eleições ou grandes eventos em geral. Este trabalho também lança alguma luz sobre como os candidatos podem influenciar os eleitores (especialmente o seu humor) usando uma plataforma como o Twitter, sugerindo assim como eles podem se aproximar do público em redes sociais online e, utilizando homofilia, analisamos como os usuários do Twitter envolvidos em uma camapanha política interagem entre si.

Palavras-chave: Caracterização do Twitter. Análise de sentimentos. Eleição presidencial americana.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUCTION

Online social networks are currently one of the main platforms for debates, discussions, and exchange of information among people in different contexts. Politics, in particular, is a recurring topic in the content shared on several social networks. As an example, Twitter has become an environment of intense political debates and opinion confrontation, especially during major campaigns, such as the 2016 US presidential campaign. This election process was characterized by intense online activity, especially after the political party primaries which resulted in the dispute between Donald Trump, representing the Republican Party, and Hillary Clinton, representing the Democratic Party. Tweets from both candidates (or their representatives) were often shared, reaching millions of users, though how such content affected (if at all) each such user is unclear. Indeed, as prior work observed, there may be great variation in how different users perceive a particular piece of content, whether towards a positive or negative sentiment [Ranganath et al. 2016].

We here take the challenge of characterizing the political behavior of politically engaged users, like the language patterns and sentiment. We also investigate how the mood of Twitter users, as expressed by the general sentiment of the content they share on the system, may be affected as they receive tweets from different candidates (i.e., from their official Twitter accounts) running a political campaign. We focus on the two main candidates running the 2016 US presidential campaign, namely Hillary Clinton and Donald Trump. We emphasize that we *cannot* establish a strong causal relationship between the candidates' tweets and the mood variation of Twitter users, as such variation might have been influenced by a collection of related and unrelated online and offline events. Yet, we analyze the Subjective Well-Being (SWB) [Diener 2000] of each user aiming at assessing whether they usually changed their mood after receiving (and often retweeting) candidates' tweets. SWB has been used before to study emotion prediction on online social networks, to identify patterns of emotions in streams of data, and to create models to predict the mood tendency on political events [Jin e Zafarani 2017, Choudhury, Counts e Gamon 2012]. Finally, we investigate how politically engaged users interact in an online social network through an homophily analysis. Homophily is the tendency of individuals to have characteristics and behavior similar to their peers. This social phenomenon has been already perceived on online social networks [Easley e Kleinberg 2010]. The characteristics of peers, for instance, friends have in common non-mutable characteristics, such as ethnicity, even as mutable characteristics such as beliefs, professions [McPherson, Smith-Lovin e Cook 2001], and sentiments towards a topic [Yuan et al. 2014] or towards political candidates [Caetano et al. 2017].

To take political biases each user may have into account, we identify four groups of users with distinct behavior and attitude towards each candidate. The first two groups are formed by *advocates* of each candidate. We consider as an advocate a Twitter user highly engaged in promoting the

campaign of her particular candidate and who would hardly change her sentiment towards any of the candidates during the campaign [Ranganath et al. 2016]. The third group is composed by users referred as *political bots*, that is, algorithm-driven accounts that aim to promote or demote a candidate's campaign [Ferrara et al. 2016]. The fourth group, referred as *regular users*, is composed of users who participate in political debates but do not present a highly biased political behavior towards any particular candidate (such as advocates), and thus may change their opinion (and sentiment) towards each candidate during the campaign as result, for example, of tweets from the candidates. These political biases lead to different reactions, behaviors and network characteristics.

We collected 23 million tweets published by about 115,000 users. To analyze how each one of the four groups behaves, we present six user characterizations. Each of these analyses allowed us the validate and characterize each Twitter user group found.

- **Which features highlight each group:** We used the K-means clustering algorithm [MacQueen 1967] and a tool called BotOrNot [Varol et al. 2017] to identify four distinct groups of users. To identify the characteristics of each group, we perform statistical analysis on their features.
- **Mean Sentiment Analysis:** We analyze the mean user sentiment towards candidates and the mean sentiment in non-political tweets during the last three months of US presidential campaign.
- **Language Patterns:** We analyze the usage of hashtags about politics in the timeline of each group as well as the most frequent words in each timeline.
- **Popular users of each group:** We analyze the top 5 most popular users of each group. Here we consider as a metric of popularity on Twitter, the total of shares (retweets) that a user had on her timeline.
- **Mood Variation Analysis:** We analyze how the tweets posted by Donald Trump and Hillary Clinton may have influenced users that retweeted them. For this we consider the mood expressed through tweets posted before and after the candidate's retweet during a time window of two hours.
- **Homophily Analysis:** We present the political homophily among users of each group. We analyze homophily considering as connections between users the Twitter follow, mention, and retweet interactions and that a connection can be both unidirectional and reciprocal.

Our study contributes to a better understanding of the political engagement of users on online social networks, particularly during the 2016 US presidential campaign. We believe that the methodology proposed here can be replicated for identifying groups of advocates, bots, and regular users during other elections or major events in general. It also sheds some light on how candidates may influence their voters (particularly their mood) using a platform such as Twitter, thus suggesting how they can approach the users on online social networks. Also, this work sheds some light on how Twitter users interact among themselves using different types of connection.

1.1 Objectives

1.1.1 Main Objective

Characterizing the behavior of users in an online social network taking into account the political biases of them.

1.1.2 Specific Objectives

1. Identifying user groups that publish texts during a political campaign (advocates of each of the candidates, political bots, and regular users);
2. Analyzing the mean user sentiment towards candidates and the mean sentiment in non-political tweets of each group;
3. Analyzing the language patterns of each group;
4. Analyzing the most representative users of each group;
5. Analyzing users' mood variations due candidates' messages;
6. Analyzing homophily among users;

1.2 Master Thesis Structure

This master thesis is organized as follows. We first discuss the theoretical reference of this work. Next, we present the related work. We describe each step of our methodology. We then discuss the results obtained when applied it to our dataset, finishing with a discussion on threats to the validity of this work. We then present our main conclusions and possible directions for future work.

2 THEORETICAL REFERENCE

In this chapter, we present concepts and techniques used throughout this work. We start presenting basic Twitter concepts and information about how candidates on the American campaign used Twitter. Next, we present statistical concepts and techniques that we use in the preparation of the dataset. Finally, we introduce data mining techniques that we use to identify and validate politically engaged groups on Twitter.

2.1 Twitter and the 2016 American Presidential Election

Twitter is a social network that allows the publication of small messages, which, in 2016, were up to 140 characters that are called tweets [Giachanou e Crestani 2016]. Released in 2006, it had about 328 million active users in 2017, and it was a popular online social network during the 2016 American Presidential Election[1].

On Twitter, when user A creates a link with user B, we say that A is following B, or that B has A as a follower. Unlike other social networks, on Twitter, the connections between users are not necessarily reciprocal because when one user follows another one, not necessarily this one will follow his/her follower.

A Twitter user profile is composed of the following attributes: name, profile description, photo, and location. User's timeline is the set of tweets that he/she published. The Republican candidate Donald Trump has a Twitter account identified by the username @realDonaldTrump and in November 2016 had about 17.1 million of followers. On the other hand, the Democrat candidate Hillary Clinton, whose account is @HillaryClinton, was followed by approximately 11.6 million users in November 2016.

A retweet is a tweet published by user A that has been shared by user B. Hashtags are one of the most common features used by Twitter users. These consist of expressions beginning with the character "#" and has the function of labeling or summarizing a topic under discussion [DeMasi, Mason e Ma 2016]. For example, Donald Trump's supporters used hashtags like #VoteTrump and #TrumpWon, while Hillary Clinton's supporters used hashtags like #VoteHillary and #NeverTrump.

Finally, the basic concept of mention is a reference to a Twitter user in the tweet text. This reference starts with the character "@". One way to target a tweet to a specific user is putting a mention to him/her at the beginning of it. Thus, users mentioned at the beginning of a tweet are called targets.

[1] http://www.statista.com/statistics/272014/global-social-networks-ranked-by- number-of-users

2.2 Data Preparation

Data preparation is an important step in data mining [Sebastiani 2002]. It allows us to eliminate noise, to treat inconsistencies and to select the best attributes of a database. In this section, we present feature selection techniques that we use in this work. Firstly, we discuss the techniques of variance and entropy. They are used to select attributes with the least possible noise, and redundancy. Next, we discuss the correlation measure. We show that it can be used to identify linear dependencies between attributes and, in this way, to find redundant attributes.

2.2.1 Variance and Entropy

The variance of a set of values indicates whether they are near or far from the mean μ. Features that have high variance indicate a greater dispersion of values in relation to the mean [Mingoti 2005], that is, they are attributes that must be considered in a statistical analysis. The variance is given by:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1} \tag{2.1}$$

Where σ^2 is the variance of an attribute x, x_i is the i-th value of the attribute x, μ is the mean of the x, and n values is total of values of x.

The entropy of a set of values indicates the degree of uncertainty (or disorder) of the set. The entropy states that the lower the probability of an event occurring, the greater the amount of relevant information that event has. The entropy is given by:

$$H(x) = -\sum_{i=1}^{n}\rho_i \log\rho_i \tag{2.2}$$

Where $\log\rho_i$ indicates the probability of the ith value occurring and ρ_i indicates the number of times the i-th value occurred in the set of values. In data mining, entropy is used to select attributes with minimum noise and high expressiveness [Brand 1998].

2.2.2 Correlation

Correlation is a statistical measure used in data mining to evaluate the degree of relationship between two attributes. One of the most common forms of association between attributes is linear, but there are other forms of association such as exponential, quadratic, or logarithmic relationships [Bishop 2006].

According to DeGroot e Schervish 2012, the first step in correlation analysis is the construction of scatter plots. Through these graphs, it is possible to verify if is there a correlation between a given

pair of attributes and in case it exists, what is its behavior (linear or non-linear). For example, a linear relationship between two attributes translates graphically into a scatter diagram where the points are arranged on a line. If two attributes are independent, the scatter diagram is graphically expressed as a random point distribution or in some cases as a set of points arranged on a horizontal line [Bishop 2006].

Also according to DeGroot e Schervish 2012, the second step in correlation analysis is the calculation of a correlation coefficient. The Pearson correlation coefficient is used to measure the linear dependence between attributes. This coefficient is given by:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.3}$$

Where x_i e y_i correspond to the i-th value of the attributes x and y with means $\bar{x}$ and $\bar{y}$, respectively. The value n corresponds to the total number of values (or observations) of the attribute. The correlation can vary between -1 and 1. The closer to 1, the greater the indication of a positive linear relationship, the closer to -1, the greater is the indication of a negative linear relationship. A correlation coefficient equal to zero or close to zero indicates that there is no relationship between the two attributes [Mingoti 2005].

The third step in the correlation analysis consists of performing a hypothesis test that aims to verify if the values of association measures observed in the data are significant [DeGroot e Schervish 2012]. The square of the coefficient, called the coefficient of determination, is used to evaluate the degree of dependence between two attributes. Since $-1 \leqslant \rho \leqslant 1$, the coefficient of determination is always between 0 and 1. The closer to 1 is the coefficient of determination, the higher degree of dependency between the two attributes.

Therefore, by calculating the correlation coefficient, if two or more attributes are highly correlated, they can be reduced to only one. In data mining, correlation can be used to reduce the dimensionality of data.

2.2.3 Normalization

The normalization of attributes allows you to transform values of different scales into values that can be compared to each other [Zaki e Jr 2014]. In this way, the clustering of objects is facilitated, since the values of the attributes of the objects of the population can be grouped using the same distance measure. In this work, we chose to use the normalization by scope, which is given by:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{2.4}$$

Where z_i is the normalized value, x_i is the original value of an attribute x. In normalization by scope, each value of an attribute is subtracted by the minimum value of the attribute x, and the

result of the subtraction is divided by the reach of the attribute x, that is, the largest value of x minus the smallest value of x. The normalized value (z_i) will be in the interval [0,1].

2.3 Data Mining

In this section, we present K-Means and Silhouette Index. We use these techniques to identify and validate politically engaged groups on Twitter.

2.3.1 K-Means

The K-Means algorithm starts by randomly selecting the initial centroids (center of the cluster) of K groups, assigning new objects to these groups based on the similarity between the registers until a convergence criterion is reached [MacQueen 1967].

Similarity can be calculated by the Euclidean distance (when the attributes of the objects are numerical) or by the Hamming distance (when the attributes of the objects are categorical). The criterion is met when there is no need to move an object from one *cluster* to another *cluster*.

The algorithm has two initial problems, the choice of K and the random selection of *centroids*. Therefore, optimization techniques such as Silhouette Index should be used to find out what the optimal K is. Also, attribute values must be normalized, and all attributes must have a high degree of importance [MacQueen 1967].

2.3.2 Silhouette Index

The Silhouette Index is a measure of cohesion and coupling of clusters [Zaki e Jr 2014]. This measure is based on the difference between the average distance of the records belonging to the nearest cluster and the records belonging to the same cluster. In this way, it is possible to validate if a cluster (*clustering*) is trustworthy. The Silhouette Index s_i is calculated for each record x_i of a data set and is defined by the following equation:

$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\left\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\right\}} \tag{2.5}$$

Where $\mu_{in}(x_i)$ is the average distance of x_i to the points of its own *cluster* and $\mu_{out}^{min}(x_i)$ is the average of the distances between the register x_i and the nearest *cluster* registers.

The Silhouette Index can range from -1 to 1. The closer the coefficient is to -1, the more indication that the grouping is unreliable. The closer to 1, the greater indication that the grouping found is reliable. That is, the value 1 indicates that the records are perfectly positioned in the respective clusters and the value -1 indicates that the records were wrongly grouped and could be in any of the

clusters found [Zaki e Jr 2014].

Rousseeuw [Rousseeuw 1987] define that if the coefficient is less than 0.25, then no substantial grouping was found. If the coefficient varies between 0.26 and 0.50, then the grouping is weak and may be artificial. If the coefficient ranges between 0.51 and 0.70, then a reasonable grouping was found. If the coefficient is greater than 0.7, then a strong grouping was found.

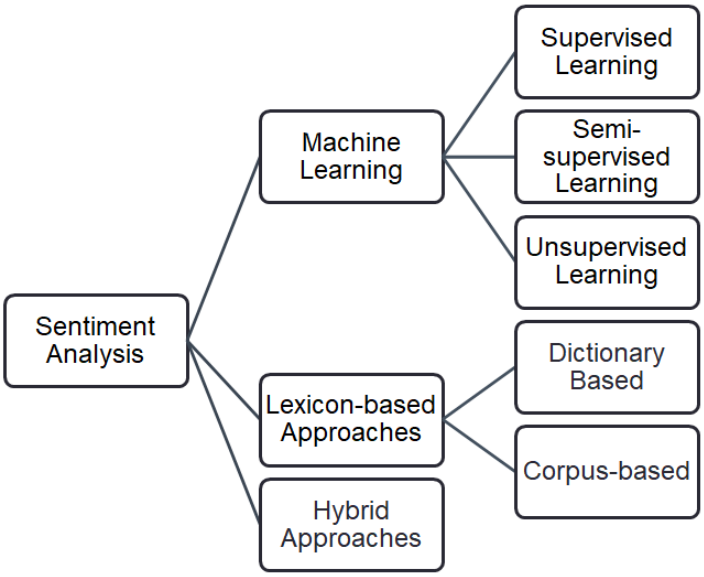
3 RELATED WORK

In this section, we present related work on sentiment analysis, identifying user groups, mood variation analysis, and homophily analysis on Twitter during political campaigns.

3.1 Sentiment Analysis on Twitter

There are several approaches to the sentiment analysis of tweets, but they can all be classified into two categories: machine learning and lexemes [Silva, Coletta e Hruschka 2016]. Figure 1 presents the possible approaches to sentiment analysis using machine learning. Supervised machine learning requires a training set consisting of labeled data. Labels represent classes (positive, neutral, negative) of each tweet. The labeling process depends on the human ability to be considered correct, but when done correctly, the training process generates a high-performance classifier. The unsupervised machine learning used the all tweets for training since there are no labeled textures for the training set. According to [Silva, Coletta e Hruschka 2016], unsupervised learning is inexpensive and easy to implement machine learning process. Semi-supervised machine learning aims to have the best of both approaches: to be robust and at the same time inexpensive and easy to implement. In the semi-supervised approach a small set of tagged tweets is used combined with the remainder of the non-labeled tweets for the training.

Figure 1: Sentiment analysis approaches



Source: Prepared by the author with data extracted from Silva, Coletta e Hruschka 2016

Sentiment analysis using lexemes can be divided into two categories: (a) dictionary-based dictionaries that use dictionary words as lexeme sources and (b) corpus-based, which use statistical or semantic methods to determine the polarity of sentiments. A dictionary contains primarily adjectives and adverbs but may contain punctuation marks, hashtags, slang, profanity, and differentiate uppercase and lowercase letters for better performance. Hybrid approaches integrate lexemes based mining techniques and machine learning techniques. In this approach, writing style, linguistic resources and semantics can be used as characteristics for classification.

Ribeiro et al. [Ribeiro et al. 2016] proposed a state of practice of strategies of sentiment analysis in different contexts and database. The study presents a benchmark of 24 methods applied to 18 different databases. The databases used in the benchmark included data from: social networks, movie repository, product review and reviews and reviews on content provider news, lectures on Ted [1] and on blogs. Of the 18 databases used in the benchmark, 8 bases were composed of tweets. In most databases (5), the sentiment of tweets was labeled by humans, usually non-specialists. To evaluate the performance of the tools and methods were used precision metrics, recall and F1 metrics. Benchmarking has shown that in the tweets database the SentiStrength tool [Thelwall et al. 2010] gives the best results. SentiStrength tool is an implementation of the dictionary-based lexeme classification technique. The tool returns the positive and negative values associated with each sentence in the text and a value *scale* which is the difference between the two values. SentiStrength relies on a dictionary consisting of 700 sentiment lexemes and uses lists of emoticons, negations and *boosting words* (very, most, worst, best) to improve the performance of analysis [Giachanou e Crestani 2016].

3.2 Identifying User Groups on Twitter

[Ranganath et al. 2016] characterized advocates in Twitter and proposed a framework to identify them. The authors modeled the message and propagation strategies used by advocates in Twitter using sociology and psychology definitions. According to the authors, advocates use message strategies to write opinion formation tweets, have persuasive language, publish tweets with a high positive sentiment, and they are determined to share language topics and patterns with as many users as possible. [Ferrara et al. 2016] proposed a framework to identify radical users who participate in political campaigns and also which users are more likely to consume extremist content. Additionally, they characterized the behavior of each user type. [Mitra, Counts e Pennebaker 2016] proposed a characterization of advocates of the anti-vaccination movement in the US. The authors divided users into three categories: long-term pro and anti-vaccination advocates and users who have recently adopted anti-vaccination attitudes and then characterized them.

Advocates use various strategies to reach people and promote their campaign. One of them is the use of a community language evoking certain words that can summarize an opinion, sentiment or will toward people and things. On Twitter, these words are hashtags, and they are used to promote

[1] https://www.ted.com/

social movements, political campaigns, and connect people. [DeMasi, Mason e Ma 2016] proposed a framework to cluster users hashtags types based on their use through time.

[Ferrara et al. 2016] and [Dickerson, Kagan e Subrahmanian 2014] characterized social bots on Twitter. The authors compared legitimate user profiles with known social bots profiles and identified that social bots tend to publish a higher number of retweets, are most recently created accounts, and publish more tweets when compared to legitimate users.

In this paper, we identified user communities using sentiment analysis features, political features, syntactic features and user activity features. In addition to the features suggested by the previous works, we also used sentiment analysis features. We applied clustering processes using these features to identify four clusters: Hillary's Advocates, Trump's Advocates, Political Bots and Regular Users.

3.3 Mood Variation Analysis

Sentiment Well Being (SWB) [Diener 2000] has been used before to study emotion prediction on online social networks, to identify patterns of emotions in streams of data, and to create models to predict the mood tendency on political events [Jin e Zafarani 2017, Choudhury, Counts e Gamon 2012]. Some papers have already proposed SWB analysis and mood analysis on Twitter. For instance, [Bollen et al. 2011] analyzed the SWB of Twitter users using sentiment analysis and proposed a metric to get the Twitter user's SWB from his/her tweets. In the work [Jin e Zafarani 2017], authors analyzed emotions of Twitter users to create a model to perform emotion prediction and to understand the types of emotions that are expressed online. They used a dataset of a network of users and their emotions aiming to identify patterns of emotions. [Choudhury, Counts e Gamon 2012] analyzed the mood landscape on social media, in the light of individuals' behavioral and social attributes. They identified more than 200 moods on Twitter using crowdsourced analysis and psychology literature. [Hernandez-Suarez et al. 2017] proposed a mood analysis methodology for predicting political bias during the United States 2016 presidential elections. The authors collected Twitter data, prepared the data for processing and classification and, then, they created a model to predict if there is a positive or negative tendency to users have mood variations during political events.

In this paper, we characterized the mood variation of Hillary's Advocates, Trump's Advocates, and Regular Users to understand how Donald Trump's tweets and Hillary Clinton's tweets may have influenced their retweeters. We applied SWB on our dataset the same way as [Bollen et al. 2011].

3.4 Homophily Analysis

Homophily is a phenomenon in which people tend to have frequent social relationships with people with similar characteristics to them [McPherson, Smith-Lovin e Cook 2001]. Several studies have observed the occurrence of the phenomenon of homophily through attributes such as ethnicity,

age, gender, religion, profession, education, interests, opinions, among others [McPherson, Smith-Lovin e Cook 2001, Currarini, Jackson e Pin 2009, Easley e Kleinberg 2010]. [Bhattacharyya, Garg e Wu 2011] showed that the level of similarity of Facebook user profile attributes is higher among friends than between random pairs of people. [Kwak et al. 2010] showed the occurrence of homophily on Twitter considering the geographic location and the user popularity. [Mislove et al. 2010] considered the phenomenon of homophily to develop a method for inference of missing attributes of social network users. Homophily makes people build more similar social networks where most people look like themselves. In this way, this phenomenon limits the individual's social context, since it strongly implies restrictions on the information that the person would receive, the opinions he/she would form, and the new social interactions that he/she would experience [McPherson, Smith-Lovin e Cook 2001].

Some papers have already proposed political homophily analysis on Twitter. In this section, we present some previous work related to the present paper. [Colleoni, Rozza e Arvidsson 2014] investigated the political homophily in an American Democratic and Republican voters database. They used a machine learning and social network analysis approach to classify users' party preference. The authors noticed that generally, the Democrats show a higher level of political homophily when comparing with the Republicans. They also found that homophily levels are higher when considering reciprocal connections. Additionally, they performed a second experiment with users that follow the official party accounts, and the Republicans political homophily also had higher rates than expected by chance and the Democrats had lower homophily than expected by chance.

[Huber e Malhotra 2017] investigated the political homophily on online dating sites. Different from the other previous work they did not use an algorithm for defining the political preference of the user, they did an analysis applying a questionary with questions such as "How do you think of yourself politically?" and "How would you describe yourself politically?". They found that people are more favorable to other people that have similar political characteristics and are more welcome to reach them out. The political ideology homophily was half as extensive as racial homophily and higher than the educational homophily.

[Halberstam e Knight 2016] investigated how political homophily influences the dissemination of information on social networks. The authors used a politically engaged Twitter users' database and identified that users linked to major political groups have more connections than users connected to minority political groups, are exposed to more information than users connected to minority political groups and receive the information faster than users connected to minority political groups.

[Brady et al. 2017] researched how the propagation of polemic content happen on Twitter. The authors did not focus on homophily itself although their findings are related to the present work. They found that the polemic discourse is much more likely to be retweeted when analyzing users of the same group. This investigation used a network with retweet as edges and the user profiles as the nodes. They estimated each user's political ideology with an algorithm based on followers network proposed by [Barbera et al. 2015].

[Barbera et al. 2015] proposed a statistical model for political ideology estimation based on the ideological position, they suggested that ideological identification is a predictive feature of the following decision. They evaluated the model with 12 political and non-political subjects on a database of 150 million tweets. They found that explicitly political users are likely to share information that comes from similar ideological users than share information from different ideological users. Conservatives are less likely than the liberals to take part in the heterogeneous dissemination of political and non-political information. They did not investigate the homophily of this groups, although this finding can also be related to a higher homophily once the groups are composed of more similar individuals that the expected by chance.

In a preliminary paper [Caetano et al. 2017], we identify user groups on Twitter using their mean sentiment towards the candidates. We then analyze homophily considering follow connections. We noted the existence of homophily among users that expressed negative sentiment towards Donald Trump and Hilary Clinton. The homophily was higher among users that had a mean negative sentiment towards Donald Trump. There was heterophylly among users that didn't publish tweets about the candidates and among users with neutral mean sentiment toward them.

4 METHODOLOGY

Recall that our present goal is to investigate how tweets sent by each of the two main candidates of the 2016 US presidential campaign may have affected the mood of politically engaged Twitter users, as expressed by the content (tweets) they shared on the system. Towards that aim, we performed 5 main steps, as described next. We started by collecting a dataset from Twitter. Next, we categorized all tweets in our dataset into either political or non-political content. For all collected tweets, we assigned a sentiment (positive, negative or neutral). Based on such sentiment and on a number of other features extracted from our dataset, we identified four main groups of politically engaged users, which differ in terms of their general behavior and sentiment towards each candidate. We then analyze the the mood variation of given user once she received a tweet from a given candidate. The final step of our methodology consists of analyzing the political homophily of each group.

4.1 Collecting Twitter Data

Using the official Twitter API, we collected a dataset, composed of tweets, user profiles, and their contact networks, comprising the period between January 1st and November 30th 2016. This period covers the dates of the three televised debates between Donald Trump and Hillary Clinton (September 26th, as well as October 9th and 19th), and election day (November 8th). This API provides up to 200 tweets published by each user (the most recent ones). Additionally, the API has a limit of 300 requests per 15 minutes time window.

We started data collection by identifying *seed users*, that is, users who published content about the US presidential campaign. The seed users were identified using the API's streaming method, which enables the real-time collection of tweets. Specifically, we considered as a seed user one who retweeted at least once a piece of content posted by one of the two candidate accounts between August 1st and November 30th[1]. Our hypothesis on doing so is that if a user retweets a tweet from a candidate account, then she has at least read it and made reference to some comment by the candidate. As such, this user is using Twitter to discuss or promote political discussions, and thus should be considered in our analyses. All users who fell into this category during the collection period were considered as *seed users*.

For each seed user, we collected his/her Twitter profile, tweet timeline (covering from January 1st to November 30th), and contact network (followers and friends). We also collected the same information from each user in each seed user's contact network, up to 2 hops from the origin. We kept only users with English as the standard language in their profiles and with 200 tweets in their

[1]We started our collection on August 1st. However, since the API provides the most recent 200 tweets posted by each user, our dataset goes back to January 1st as initial date.

timelines. In total, we collected data from 115,664 users (37,468 seed users and 78,196 seed users' network). Figure 2 shows the steps followed in the data collection.

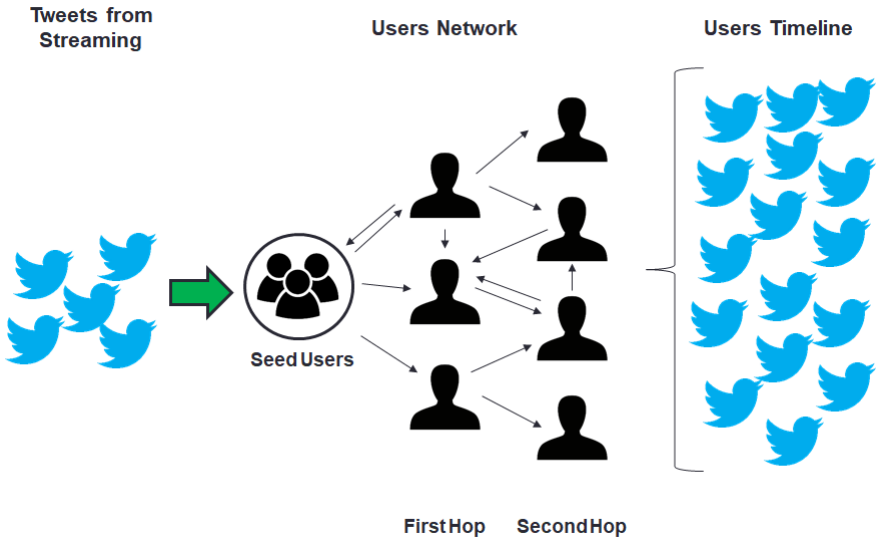


Figure 2: Data collection steps

Table 1 summarizes our collected dataset[2], showing the total numbers of tweets, user profiles, and follower/friend relationships among these users.

Table 1: Summary of our Dataset

Number of tweets	23,132,800
Number of users	115,664
Number of relationships among users	1,762,451

4.2 Identifying Political Tweets

One key step of our methodology is to categorize all tweets into either political or non-political content. This categorization is required in the next two steps of our methodology, namely performing sentiment analysis of each tweet and identifying politically engaged user groups. We used syntactic and semantic features to identify whether a tweet has political content or not [Mitra, Counts e Pennebaker 2016, DeMasi, Mason e Ma 2016]. Specifically, we analyzed the hashtags and mentions that occurred in each tweet[3]. We defined the following conditions for the identification of political tweets:

- The tweet is a retweet of a post by one of the two candidates;
- The tweet has a reference to at least one candidate;

[2]Our dataset will be publicly available if this paper is accepted
[3]A mention is a reference to a Twitter user staring with "@".

- The tweet contains at least one hashtag associated with the political campaign of one of the candidates;

We considered as a reference to a particular candidate a mention to the candidate's official Twitter account or the presence in the tweet text of the name or name abbreviation of the candidate. Thus, the presence of words "@realDonaldTrump","Trump" and "DT" (for candidate Donald Trump), or "@HillaryClinton", "Hillary" and "HC" (for candidate Hillary Clinton) was considered a flag of a political tweet. We performed the syntactic and semantic analysis with all tweet words converted to lower case.

We also considered as a criterion for the identification of political content the use of at least one of the 10 most popular hashtags related to each candidate's campaign, that is, the 10 hashtags that most often co-occurred (in our dataset) with a slogan commonly used by each candidate. One of the Donald Trump's campaign slogans was "Make America Great Again" (MAGA), while one of the Hillary Clinton's slogans was "ImWithHer". The hashtags used in our study are shown in Table 2. The assumption is that candidates and other Twitter users would use such hashtags to publish a tweet about the presidential campaign.

Table 2: *Hashtags* related to the candidates

	Donald Trump	**Hillary Clinton**
1	#Trump	#ImWithHer
2	#MAGA	#NeverTrump
3	#TrumpTrain	#Hillary
4	#TrumpPence16	#HillaryClinton
5	#DrainTheSwamp	#Hillary2016
6	#tcot	#UniteBlue
7	#Trump2016	#VoteBlue
8	#GOP	#HillaryBecause
9	#PJNET	#OHHillYes
10	#cco	#HillYes

A tweet that had at least one of the conditions above satisfied (for one or both candidates) was considered as a political tweet. We refer to the set of political tweets in each user's timeline as the user's political timeline.

4.3 Tweet Sentiment Analysis

The next step of our methodology consists of assigning a sentiment to each tweet in each user's timeline. We do so for all (political and non-political) tweets. On the one hand, we intend to assess the sentiment expressed on a particular (political) tweet as regarding one candidate or both candidates. On the other hand, we also want to assess the general mood of each user during a given time window based on all tweets she posted during that period.

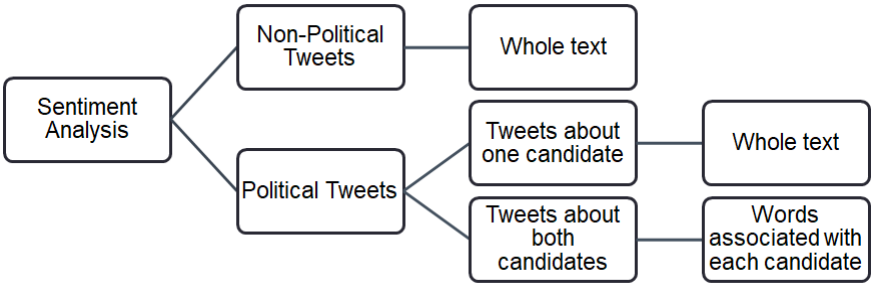
For political tweets specifically, if the tweet is associated with a particular candidate *A*, the tweet's sentiment reflects a sentiment towards *A*. However, the analysis becomes challenging for political tweets that refer to both candidates in the text. For example, a very positive comment about one candidate could cancel out a negative expression about the other, leading to an overall neutral sentiment score.

We took here three approaches when performing sentiment analysis on a tweet. For non-political tweets, we built a bag of words containing all words in the tweet. For a political tweet referring to a single candidate, the whole text of the tweet was used to build a bag of words for the candidate. For a political tweet referring to both candidates, we first identified the words related to each candidate in the tweet text and then built one bag of words for each candidate.

We identified the words associated with Hillary Clinton and Donald Trump in a tweet using the Stanford Parser tool [Klein, Manning et al. 2003]. It is a natural language parser that processes the grammatical structure of the sentences, allowing us to group words into phrases. We can use it to identify which words are phrase subjects and which ones are associated with each subject. We identified the subjects related to each candidate considering the Twitter username, candidate's name and name abbreviation (as defined in the previous section). We also explored the use of third-person personal pronouns. That is, we defined the pronouns "he" and "him" as related to Donald Trump and the pronouns 'she' and "her" as related to Hillary Clinton. The words associated with the candidates' subjects and pronouns were used to build each candidate's bag of words.

Thus, for each political tweet, we performed a sentiment analysis in each candidate's bag of words to identify the general sentiment towards the candidate expressed in that tweet. For non-political tweets, we analyzed the sentiment of the whole text of the tweet. We used the SentiStrength tool [Thelwall et al. 2010] to perform the sentiment analysis. This tool returns a score that varies from -4 to +4. As in [Ribeiro et al. 2016, Giachanou e Crestani 2016], we assume a positive sentiment for positive scores, negative sentiment for negative scores, and neutral if the score is 0. Figure 3 presents a diagram with the two strategies that we foll to perform sentiment analysis on tweets.

Figure 3: The two sentiment analysis followed approaches



4.4 Identifying Politically Engaged User Groups

In this section, we describe our steps to identify four groups of politically engaged users, namely, advocates of either candidate (Hillary or Trump), political bots and regular users.

4.4.1 Removing Outliers

Since our focus is on politically engaged users, we consider users who did not publish any tweet related to either candidate (i.e., political tweet) in their timelines as *outliers*. These users did not express their political position in their tweets during the analyzed period, and thus should be disregarded from our analyses. To identify outliers, we defined the Political Discourse of user u (PD_u) as the average fraction of tweets related to either candidate in u's timeline.

$$PD_u = \frac{T_{u,Trump} + T_{u,Hillary}}{2} \tag{4.1}$$

where $T_{u,A}$ is the fraction of all tweets in u's time that makes reference to candidate A[4]. Note that PD_u varies from 0 (user u did not publish any political tweet) and 1 (all tweets published by u were political tweets towards both candidates).

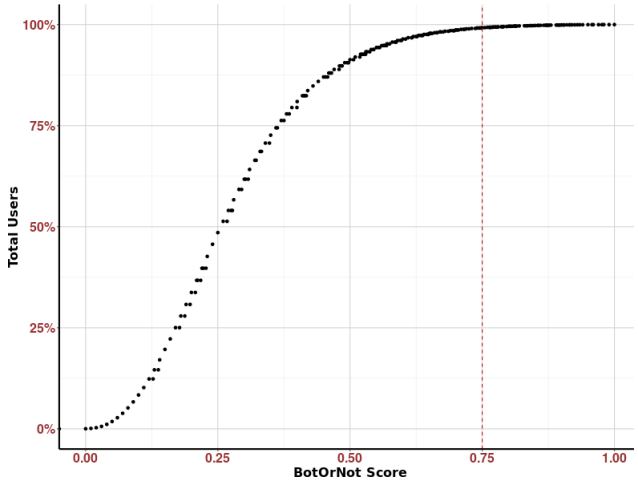
We identified 1,318 users (out of 115,664) with political discourse equal to zero, and removed them. Thus, we considered the remaining 114,346 politically engaged users in the following procedures.

4.4.2 Identifying Political Bots

We used the BotOrNot tool [Varol et al. 2017] to identify bots in our dataset. This tool combines machine learning algorithms, sentiment analysis, and other features to define whether a given profile has more characteristics of being legitimate or being a false account [Varol et al. 2017]. The tool returns a score between 0 and 1. The closer to 1, the higher the probability of that user profile being a bot. However, there is no recommendation on a particular score threshold that distinguishes bots from legitimate users [Varol et al. 2017]. It is up to the users (this case, us) to interpret the results and decide whether a Twitter user is a bot or not. To make such decision, we first analyzed the Cumulative Distribution Function of the BotOrNot scores computed for all users (Figure 4). However, we found no clear break nor change in the behavior of the distribution. Thus, we chose to be conservative in our choice. We considered as political bots all users who had BotOrNot scores greater than or equal to 0.75. Considering this threshold, we identified 4,053 (out of 114,346) users in this group.

[4] A tweet that makes reference to both candidates is counted twice (in both $T_{u,Trump}$ and $T_{u,Hillary}$).

Figure 4: Users BotOrNot scores CDF



Next, we present the procedures to categorize the remaining 110,293 users into three different groups. We start by discussing the features exploited in such categorization and then describe the method adopted.

4.4.3 Feature Set Engineering

We processed the user timelines to build a feature set for each user. In total, we considered 44 features, as shown in Table 3, divided into four categories: user metadata, syntax, political bias and sentiment analysis. The user metadata, syntax and political bias features were engineered according to [Ranganath et al. 2016, Ferrara et al. 2016, Mitra, Counts e Pennebaker 2016]. The use of sentiment analysis features along with the others is a contribution of our work, compared to previous efforts exploring community/advocacy and extreme user behavior identification.

The *User Metadata* category contains 6 features related to the user activity and profile characteristics [Ferrara et al. 2016, Subrahmanian et al. 2016]: fraction of geolocalized tweets in timeline, number of hashtags used in profile description; number of characters in Twitter username; network popularity (total number of followers/friends); tweet interaction (ratio of number of published tweets to number of favorited tweets) and the BotOrNot score.

Syntax category contains 10 features related to how the user writes her tweets [Ferrara et al. 2016, DeMasi, Mason e Ma 2016, Mitra, Counts e Pennebaker 2016]. Syntactic features include the average (avg) and standard deviation (std) – distinct metrics – of the numbers of hashtags, URLs, mentions, targets[5] and words per tweet in the user's timeline.

The *Political Bias* category contains 11 features related to the user's political bias expressed in her tweets [Ranganath et al. 2016, Ferrara et al. 2016, Mitra, Counts e Pennebaker 2016]: the political

[5]The term target refers to a mention "@" to a user at the beginning of the tweet, used as a means to direct the post to that particular user, as opposed to only citing her (as in regular mentions).

Table 3: User's feature set (*A* refers to Trump or Hillary)

Category	Name
User Metadata	percentage geolocalized tweets in timeline
	number of hashtags in description
	number of characters in username
	network popularity
	tweet interaction
	BotOrNot score
Syntax	number of hashtags per tweet (avg,std)
	number of URLs per tweet (avg,std)
	number of mentions per tweet (avg,std)
	number of targets per tweet (avg,std)
	number of words per tweet (avg,std)
Political Bias	political Discourse
	percentage of tweets targetting *A*
	percentage of tweets with mentions to *A*
	percentage of tweets with reference to *A*
	percentage of retweets from *A*
	avg number of political hashtags of *A* per tweet
Sentiment Analysis	all tweets (avg and std of sentiment score)
	tweets targetting *A* (avg,std)
	tweets with mentions to *A* (avg,std)
	tweets with reference to *A* (avg,std)
	positive/negative bias towards *A*
	positive/negative bias (non-political)

discourse of user u (PD_u), the fractions of tweets targetting, mentioning or making some reference to either candidate, the fraction of all retweets in user's timeline of some tweet from either candidate, the average number of political hashtags of either candidate (Table 2) per tweet. Note that, on Table 3, *A* refers to eitheer candidate. Thus, we account each metric generates two features, one for each candidate.

Finally, the *Sentiment Analysis* category contains 17 features related to the sentiment expressed towards either candidate and the sentiment expressed in non-political tweets. It includes the mean and standard deviation of the sentiment scores assigned to all tweets and only tweets targetting/mentioning/referencing either candidate. It also includes a measure of the positive/negative bias of the user, calculated as the number of negative tweets to the number of positive tweets, computed separately for each candidate and non-political tweets posted by the user.

4.4.4 Identifying Regular Users, Hillary's Advocates, and Trump's Advocates

We used the K-means algorithm [MacQueen 1967] to perform the clustering of the non-bot users in our dataset. We hierarchically applied K-means, in two phases: we first separate regular users from political advocates and then distinguish Trump's and Hillary's advocates within the second group.

The reason for this approach is that we might be able to distinguish regular users from advocates better if we consider only these two profiles first. Similarly, we may be able to better distinguish advocates for either candidate if regular users are not considered in the analysis. We used Silhouette Index [Zaki e Jr 2014] to evaluate the clustering results (Section 2.3.2).

For both phases of the clustering procedure, we performed a feature selection to reduce the feature set. This procedure, based on discussion in [Zaki e Jr 2014,Bishop 2006], consists of three steps, performed separately for each clustering phase. First, we computed the pairwise correlation across all 44 features to identify pairs of highly correlated features. Highly correlated features are redundant and do not need to be simultaneously used. Using Pearson clustering coefficient [Bishop 2006] (Section 2.2.2), we identified several groups of highly correlated features (correlation coefficient above 0.7). For each such group, we retained only one feature, selecting the one with higher variability (and thus less bias towards fewer values), as this could better distinguish different user behaviors [Bishop 2006]. Second, we further analyzed the distributions of each retained feature, removing those whose distributions were highly concentrated around few values and thus poorly discriminative of different user groups (Section 2.2.1). As a final step, we performed a greedy feature selection approach: taking one feature (out of the remaining ones) at a time, we compared the clustering results with and without it, choosing to keep the feature, only if the results were worse without it.

The result of the feature selection was as follows. The clustering of users into regular users and political advocates considered the features: political discourse, average number of political hashtags related to Trump/Hillary per tweet, and positive/negative bias towards Trump/Hillary. The further clustering into Trump's and Hillary's advocates used: number of hashtags in user's description, average number of words per tweet, the fraction of political tweets with some reference to Trump/Hillary and standard deviation of the sentiment score of tweets with some reference to Trump/Hillary.

For both clustering phases, we normalized the feature values using the scope normalization (Section 2.2.3) so that their numeric values range from 0 to 1. Recall that the number of clusters k is an input parameter of K-means. For each clustering procedure, we tested for the optimal value of k by evaluating the results of K-means (in our selected feature set) with increasing values of k, starting from $k=2$. We found that the Silhouette index was worse for all values of $k>2$. Therefore, 2 clusters were found in both clustering phases, as expected.

4.5 Analyzing Mood Variation

This step of our methodology consists of analyzing how tweets from either candidate may have affected the mood of individual users in each of the four identified groups. To that end, we make use of the concept of Subjective Well-Being (SWB) [Diener 2000]. SWB consists of a person's cognitive and affective evaluations of her life. SWB has various separable components such as life satisfaction, satisfaction with important domains (e.g., work satisfaction), positive and negative affects,

etc. SWB has been widely used in various domains, including the analysis of people's emotions and happiness based on data from online social networks [Jin e Zafarani 2017, Bollen et al. 2011].

We here apply the SWB concept to evaluate a user's mood variation, as expressed by the general sentiment of tweets the user shared, during a time window δ before/after the user received a tweet from one of the candidates. We assess mood variation around a candidate's tweet that the user retweeted, as a retweet is a strong indication that the user actually read and to some extent internalized the candidate's tweet content.

First, borrowing from [Jin e Zafarani 2017, Bollen et al. 2011], we define the subjective well being (SWB) of user u during the time interval $[t_1,t_2]$, $S_u(t_1,t_2)$ as the fractional difference between the number of positive tweets and the number of negative tweets posted by u in $[t_1,t_2]$:

$$S_u(t_1,t_2) = \frac{N_{p_u}(t_1,t_2) - N_{n_u}(t_1,t_2)}{N_{p_u}(t_1,t_2) + N_{n_u}(t_1,t_2)} \tag{4.2}$$

where $N_{p_u}(t_1,t_2)$ and $N_{n_u}(t_1,t_2)$ are the numbers of tweets posted by u during time interval $[t_1,t_2]$. The value of S_u is in the $-1 \leq S_u \leq 1$ range.

Let's then assume that user u received a tweet from candidate A and retweeted it at time t. We assess u's mood variation around this tweet as:

$$\Delta S_u = S_u(t,t+\delta) - S_u(t,t-\delta) \tag{4.3}$$

The value of ΔS_u is in the $-2 \leq \Delta S_u \leq 2$ range. Positive values imply in mood change towards a more positive trend after the retweet, while negative values imply in a change towards more negative sentiment.

We emphasize, once again, that we cannot claim any direct causality effect between the candidate's tweet and a possible mood variation of a user, as many other factors may have contributed to the change. However, by considering a reasonably short time window δ (e.g., δ = 2 hours), we increase the chance of some correlation between them.

4.6 Homophily Analysis

In the final step of our methodology, we analyze the homophily among groups considering three connection types on Twitter: follow, retweet, and mention. To mathematically represent the homophily level of the group i, [Colleoni, Rozza e Arvidsson 2014] applied the following equation:

$$H_{\mathrm{i}} = \frac{s_{\mathrm{i}}}{s_{\mathrm{i}} + d_{\mathrm{i}}} \tag{4.4}$$

Where H_i is the homophily index, s_i represents the number of connections among users of group *i* (homogeneous connections), d_i represents the number of connections among users of group *i* with users of other groups (heterogeneous connections).

[Currarini, Jackson e Pin 2009] also used the Equation 4.4 to calculate the homophily level. However, they underline the difficulty of measuring homophily simply by considering H_i. The authors presented the following example: suppose a group A that corresponds to 95% of a network population and a group B that corresponds to the remaining 5%. Assume that each group has a percentage of 96% of homogeneous friendships ($H_i = 0.96$). When comparing the group A with the group B, although both have the same H_i, the homophily level among the group B is higher than the group A, since the probability of friendships between the group B members is smaller than the probability of the group A members.

In this way, [Currarini, Jackson e Pin 2009] recommended to use inbreeding homophily index, developed by [Coleman 1958] to normalize the H_i. This measure is given by:

$$IH_i = \frac{H_i - w_i}{1 - w_i} \tag{4.5}$$

Where H_i is the homophily index defined in Equation 4.4 and w_i is the probability of the occurrence of *i* individuals. The w_i consists of the total of *i* individuals divided by the total number of individuals in a network.

Returning to the previous example, the IH_i value for groups A and B, is 0.2 and 0.96, respectively. This result demonstrates that the inbreeding homophily index can be used to compare relative homophily between different populations. The higher the value of IH_i, stronger is the homophily occurrence.

The opposite of homophily is Heterophily since there is a predominance of relationships among individuals of different kinds. When the IH_i is zero, it corresponds to the homophily baseline determinant. In this work, we defined the occurrence of homophily or heterophily using the following condition:

$$\begin{cases} IH_i > 0 & homophily \\ IH_i < 0 & heterophily \end{cases}$$

4.6.1 Network Description

To perform the homophily analysis, we define a network where the nodes represent Twitter users, and the edges represent the interactions between them. We here select a sample of 68,863 users from the 114,346 users since we only collected the network data from these 68,863 users (Section 4.1). Therefore, all following homophily analysis will consider the connections among the 68,863

users instead of the 114,346 users.

We define the connection types as follow, mention, and retweet. We define that there is following connection between two users when at least one of them follows the other. There is retweet connection with two users when at least one of them retweets the other, and there is mention connection with two users when at least one of them mentions the other. Therefore, there is reciprocal follow connection with two users when both follow each other. There is reciprocal retweet connection with two users when both retweet each other, and there is reciprocal mention connection with two users when both mention each other. In our network, there are 18,122 (26%) Hillary's Advocate nodes, 10,866 (16%) Trump's Advocate nodes, 2,675 (4%) Political Bot nodes, and 37,200 (54%) Regular User nodes.

4.6.2 Calculating Homophily Level of each Group

In this work, we analyze political homophily on Twitter considering uniplex connections. That is, we perform single analysis considering follow connections, mention connections, and retweet connections. We then measure homophily in two contexts: (i) considering only unidirectional edges and (ii) considering only reciprocal edges.

We calculate the H_i (Equation 4.4) for each one of the groups. Using H_i, we also calculate the index IH_i (Equation 4.5). The index IH_i is useful for comparing the homophily level among different classes. For example, to compare whether homophily between Hillary Supporter users is greater than the homophily between Trump Supporter users, even if their number of users is different.

Note that we consider the variable w_i as one of the parameters in the IH_i calculation for a class i, which corresponds to the probability of the occurrence of a user from class i in the network under analysis. This way, w_i values are 0.23, 0.12, 0.04 and 0.61 in all scenarios that we analyze for Hillary's Advocates, Trump's Advocates, Political Bots, and Regular Users, respectively.

5 EXPERIMENTAL RESULTS

In this section, we characterize the user groups by first discussing clustering results and analyzing how the features exploited in the clustering approach differ across groups. Next, we analyze the sentiment towards each candidate as well as the sentiment expressed in non-political tweets, the usage of hashtags and the most frequently used words in each user group. Next, we illustrate each user group by manually inspecting the top 5 most retweeted users in each of them. Then, we perform a mood variation analysis of each group and finally, a homophily analysis of each group. In the daily mean sentiment analysis and the language patterns analysis, we use only tweets published during the last three months of the presidential campaign (09/01/2016 to 11/30/2016), since it was the period of more user activity in our dataset.

Collectively, all analyses validate and characterize the groups found. In particular, we do expect that a candidate's advocate has a more positive sentiment towards the candidate she supports while expressing a more negative sentiment towards the other candidate. Differences in the language patterns used by each candidate's advocates can also be expected [DeMasi, Mason e Ma 2016], as well as differences in mood variation between advocates of each candidate and regular users after receiving tweets from either candidate.

5.1 Clustering Results

Recall that we applied a hierarchical clustering approach by first applying K-means to separate regular users from advocates, and then using K-means a second time to break the latter into each candidate's advocates. In the first clustering process, we achieved a Silhouette Index of 0.81, which indicates clusters of very good quality, i.e., very clearly separated. One cluster has 70,290 users while the other has 40,003 users. Table 4 presents the mean (μ) and standard deviation (σ) of each feature used in the clustering process. We define the following alias for each feature: feature 1: political discourse; feature 2/3: average number of political hashtags related to Trump/Hillary per tweet; feature 4/5: Positive/negative bias towards Trump/Hillary. Users in the second cluster (4^{th} and 5^{th} columns), tend to have a much stronger political discourse as well as use more political hashtags and have stronger biases towards either candidate (especially Hillary). Thus, we labeled the first cluster Regular Users and the second one Advocates.

Table 4: Feature Analysis: Regular Users vs Advocates

	Regular Users (70,290)		**Advocates (40,003)**	
	μ	σ	μ	σ
Feature 1	0.0871	0.4083	0.4614	1.5802
Feature 2	-0.0005	0.0088	-0.0080	0.0297
Feature 3	-0.0066	0.0141	-0.0318	0.0385
Feature 4	0.0759	0.0617	0.3431	0.1050
Feature 5	0.0833	0.4276	0.6592	2.1534

Table 5: Features Analysis: Trump's Advocates vs. Hillary's Advocates

	Hillary's Advocates (26,230)		**Trump's Advocates (13,773)**	
	μ	σ	μ	σ
Feature 6	0.4030	0.1494	0.3516	0.1886
Feature 7	0.2787	0.1934	0.3429	0.1961
Feature 8	0.5578	0.2349	0.7702	0.2532
Feature 9	0.8355	0.1864	0.6504	0.2624
Feature 10	3.7241	4.8296	7.8192	5.2273
Feature 11	0.4692	1.4341	0.7009	1.7545

In our second clustering process applied to the advocate users, we obtained two clusters with a high Silhouette Index of 0.72. Once again, these results show very good separation between the two groups. One cluster has 26,230 users, and the other has 13,773 users. Table 5 presents the same statistics for each feature used in the clustering. We define the following alias for each feature: feature 6: number of hashtags in user's description; feature 7: avg number of words per tweet; feature 8/9: % political tweets with reference to Trump/Hillary; feature 10/11: standard deviation of sentiment score of tweets with reference to Trump/Hillary.The main differences between the two clusters are the fractions of political tweets concerning one of the candidates (features 8 and 9) as well as the variability in the sentiment score of those tweets (features 10 and 11). Users in one of the clusters (2^{nd} and 3^{rd} columns) tend to post tweets regarding Hillary more often, and have lower variability in the sentiment expressed by tweets regarding either Hillary or Trump. Users in the other group mention Trump more often, although the sentiment scores of all their political tweets tend to vary more. We label the former Hillary's advocates and the latter Trump's advocates.

5.1.1 Testing Hypothesis

We conduct a testing hypothesis to validate the users' features of each cluster identified in the two clustering moments. To perform this test, we use the null hypothesis of the two-tailed test of the population mean. According to this analysis, each feature means lies in its most significant values. Moreover, when comparing a two-tailed test in both clusters of the two clustering processes, we observe that all features' means are different. Therefore, we reject the null hypothesis that says that both clusters have equal means.

5.1.2 Examples of Tweets

In this section, we show some examples of tweets posted by users of each group to validate and show a typical user behavior related to each group. We randomly selected three tweets from each group. We expect that Hillary's Advocates have negative tweets about Donald Trump and positive tweets about Hillary Clinton. We expect the opposite when analyzing tweets of Trump's Advocates. Political Bots' tweets Regular User's tweets may have various characteristics since they tend to talk about politics and non-political content and have different sentiments towards the candidates. Using the following examples, we observe that each group has true positive users.

Hillary's Advocates

- Hillary unlike Trump is Americas best hope. Hillary is extremely positive Trump is extremely negative. Vote for Hillary.
- RT @ChuckNellis: Trump is a vile despicable human, Hillary is no better. #VoteYourConscience https://t.co/3GYeqYD8Cw
- Yo if anyone wants to know why I believe @HillaryClinton will be good for our country and Mr. Trump will be bad, I've got a list going.

Trump's Advocates

- @politico Win! Hillary has already destroyed the FBI, DOJ , Senate & now has her eyes set on the presidency so to destroy the USA vote Trump
- #Trump loves America and he loves Americans. #Hillary hates America and she hates Americans. https://t.co/NWEUGPE4cE
- @Anna_Giaritelli @dcexaminer Hillary despises her husband, people in,general: Trump has wonderful family: Who's more dangerous?#bliss

Political Bots

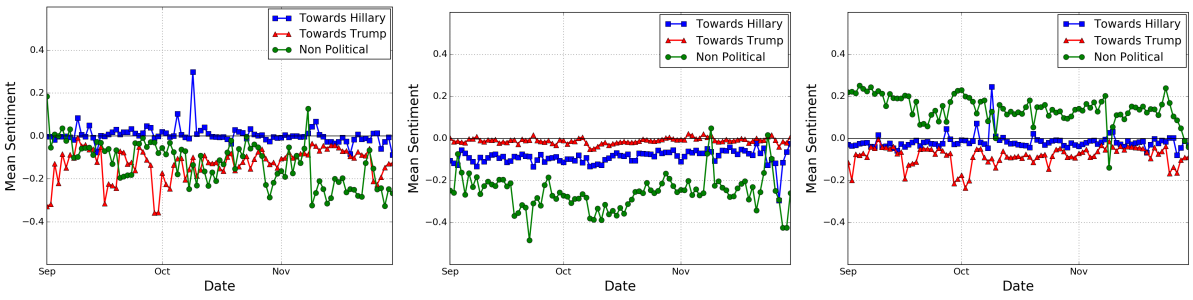
- @HillaryClinton OMG, Hillary is so offended, Trump said "men".
- @foxnewspolitics he only saying that now, he feels the debate was not so bad, he is pleased to hear Trump say Hillary is tough and strong etc
- @realDonaldTrump so many Trump haters but yet when asked what @HillaryClinton has accomplished as a politician, they remain silent...

Regular Users

- The point is, Trump still is an awful choice for President Hillary is our best hope for a USA to continue on a forward path of unity
- RT @PaulRudnickNY: Trump's rallies are vicious lynch mobs,
  while Hillary's are passionate celebrations. Why would anyone choose hatred?
- @paulkrugman You mean Trump was the best candidate against Hillary, who is also horrible. He's a Pied Piper candidate if there ever was one.

5.2 Daily Mean Sentiment

The sentiment expressed in tweets published by a user helps to understand whether she agrees with a specific topic [Wong et al. 2016], supports a political campaign [Wang et al. 2012] or is involved in social movements [Mitra, Counts e Pennebaker 2016]. In the present analysis, we calculate the daily mean sentiment score of political and non-political tweets posted by users in each group. Figure 5 shows three time series of the average sentiment scores towards the candidate Donald Trump, towards the candidate Hillary Clinton and in non-political context for each user group. We omit results for political bots since their sentiment is artificial and is not influenced by candidate's tweets or any online and offline event [Ferrara et al. 2016].



(a) Hillary's Advocates (b) Trump's Advocates (c) Regular Users

Figure 5: Daily mean sentiment towards both candidates and in non-political context for each user group

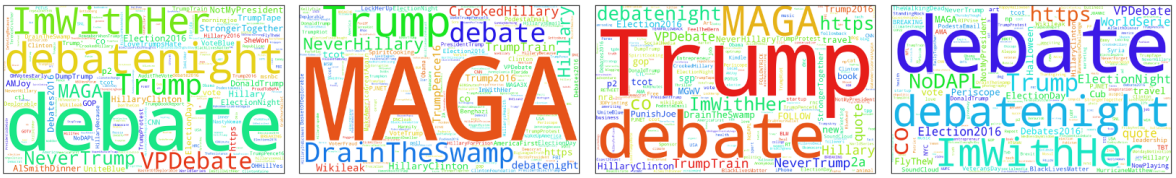
For both candidate's advocates, the daily mean sentiment is negative in almost all analyzed days regardless of the topic of the tweet, that is, regardless of whether the tweet is towards Hillary, towards Trump or about non-political content. Moreover, in absolute terms, the negative sentiment of tweets from Trump's advocates tend to be stronger. There are a few exceptions, however. For example, Trump's advocates experienced a positive mean sentiment in their non-political tweets on November 9^{th} (the day after the election day). Hillary's advocates, on the other hand, had a peak of positive sentiment in their non-political tweets on the day before the presidential election (November

8^{th}), followed by a sharp drop (on non-political sentiment) in the following day. These observations suggest a positive/negative boost in the general mood of these users to some extent possibly related to the election results

For regular users, on the other hand, the mean daily sentiment of non-political tweets remain positive in all days, except election day. The other two time series are similar to the corresponding ones for Hillary's Advocates, except that the sentiment towards Trump tends to be somewhat less negative among regular users.

5.3 Language Patterns

Hashtag usage patterns may indicate a user engagement on a particular topic as well as her participation in communities. In [DeMasi, Mason e Ma 2016], the authors demonstrated that hashtags contain more information than pure text and thus are a valuable feature to be analyzed. Inspired by such observation, we processed each user's timeline to retrieve all hashtags used in her tweets. We then built a hashtag cloud with the hashtags used by all users in each group, as shown in Figure 6. In a hashtag cloud, the font size captures the number of times a hashtag was used. The following discussion is based on the characterization of hashtags related to each candidate shown in Table 2.



(a) Hillary's Advocates (b) Trump's Advocates (c) Political Bots (d) Regular Users

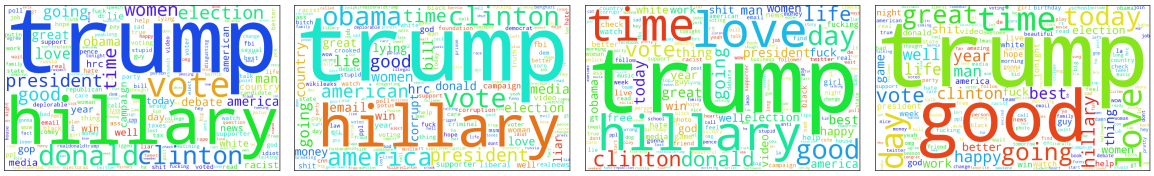
Figure 6: Hashtag cloud of each user group

As expected, both candidates' advocates used hashtags related to the campaign of the corresponding candidate more often. Moreover, Hillary's advocates used hashtags related to TV debates more often than Trump's advocates. Interestingly, the hashtags more often used by political bots were about Trump's campaign (much more often than Hillary's) and TV debates.

On the other hand, regular users tend to use more hashtags related to TV debates, about the election in general (#ElectionNight) and topics not directly related to the presidential campaign (#NoDAPL, #WorldSeries). The hashtag #NoDAPL (*On the Dakota Access Pipeline*), for instance, is related to an advocacy movement of indigenous and environmental rights in the United States. The use of such off-topic hashtags indicates that regular users, unlike advocates and political bots, also engage in other non-political matters, even during the campaign. Only two hashtags directly related to Donald Trump and Hillary Clinton (#Trump and #ImWithHer) were among the most common hashtags used by regular users.

One interesting analysis is whether user communities can be identified within each user group. We leave this investigation for future work.

We also identified the top 15 most frequently used words in each user's timeline. We then built a word cloud with these words for each group, as shown in Figure 7.



(a) Hillary's Advocates (b) Trump's Advocates (c) Political Bots (d) Regular Users

Figure 7: Word cloud of each user group

The word "trump"appears as the most frequently used word in all four groups, suggesting great attention to the candidate in general. The word "hillary" was the second most popular word for users in the three more politically engaged groups (i.e., all but regular users). Moreover, for both advocate groups, the most frequent words are associated with politics. On the other hand, regular users and political bots tend to frequently use also more general words such as "good", "love"and "time". Based on hashtags used by both groups, one might speculate that such words are often used jointly with other politically oriented words by political bots, whereas regular users may use then in a more diverse set of contexts.

5.4 Popular Users

We manually inspected five users from each group. We selected for inspection the 5 most popular users in each group, estimating user popularity by the number of retweets, i.e., by the number of times a tweet by the user was retweeted by someone else. Table 6 lists the five most popular users of each cluster, as well as the numbers of retweets received.

We note that four of the 5 most popular Hillary's advocates are official US media user accounts (@CNN, @nytimes, @thehill, @ABC). The last one is the official user account of Democrat Senator Tim Kaine (@timkaine), who was Hillary's vice presidential candidate in the election campaign. Only @timkaine had an explicit statement supporting Hillary's campaign. We also note that all five Hillary's advocates have verified Twitter accounts.

The most popular Trump's advocate is @mitchellvi, Twitter account of Bill Mitchell, who is the owner of a website and a YouTube channel dedicated to supporting Donald Trump's campaign. Indeed, @mitchellvii profile description presents him as a Trump's supporter. @TEN_GOP, @Linda-Suhler, and @bfraser747 also present themselves, in their profile descriptions, as unofficial supporters of Donald Trump's campaign. Profile descriptions of these users contain statements such as "I'm

Table 6: Top 5 Popular User Profiles of each Group

Cluster	Username	Retweets Received
Hillary's Advocates	@CNN	52,872
	@timkaine	33,903
	@nytimes	32,979
	@thehill	29,817
	@ABC	21,459
Trump's Advocates	@mitchellvii	34,981
	@TEN_GOP	9,380
	@LindaSuhler	9,210
	@bfraser747	9,154
	@LouDobbs	9,117
Political Bots	@Pamela_Moore13	3,550
	@OnMessageForHer	2,778
	@Crystal1Johnson	2,644
	@_xM_G_W_Vx_	2,370
	@__xM_G_W_Vx__	2,003
Regular Users	@JoyAnnReid	53,896
	@kurteichenwald	48,571
	@wikileaks	48,566
	@SenSanders	27,730
	@chrislhayes	15,810

a Trump supporter" and "Proud Trump Supporter." Only @LouDobbs did not have such statement in the profile. This user account belongs to the host of Fox News "Lou Dobbs Tonight" program. Moreover, only @mitchelvii and @LouDobbs are verified Twitter accounts.

The most popular political bots had more recent account creation dates when compared to the top users of the other clusters. Indeed, all five accounts of political bots were created between 2015 and 2016. We note that @_xM_G_W_Vx_ and @__XM_G_W_Vx__ have very similar usernames as well as the same account date creation and profile description. The profile description presented an offer to buy 5,000 followers within a week. The accounts of all 5 political bots were *not* verified accounts. Moreover, we notice that Twitter had banned all five accounts, strengthening the argument that they were indeed bots [Varol et al. 2017, Ferrara et al. 2016].

The two most popular regular users (@JoyAnnReid and @kurteichenwald) are the accounts of writers Joy Reid and Kurt Eichenwald, respectively. The third one is the official WikiLeaks Foundation account, while @SenSanders is the official account of Senator Bernie Sanders, who was a pre-candidate in the presidential campaign and supported Hillary Clinton's campaign. User @chrislhayes is the official account of MSNBC host Chris Hayes. The accounts of all five regular users have been verified.

As a final comment, we note that the five selected Hillary's advocates are in general more popular (i.e., received a larger number of retweets) than the Trump's. Indeed, this holds for the whole

clusters, as the average numbers of retweets received by Hillary's and Trump's advocates are 37.99 and 29.06, respectively. Four of the selected Trump's advocates received fewer than 10,000 retweets, a number more comparable to the retweets received by political bots. Regular users, on the other hand, tend to be more popular in general, receiving a large number of retweets.

5.5 Mood Variation Analysis

Here we analyze the perceived mood variation (ΔS_u) in users of each group as they receive *and retweet* tweets from either candidate. We analyze mood variation for advocates (both Hillary's and Trump's) as well as regular users (bots are disregarded for obvious reasons). Recall that ΔS_u varies from -2 to +2 with negative (positive) values representing a variation towards more negative (positive) sentiment. Our analysis considers time windows before/after the retweet from the user with duration $\delta = 2$ hours.

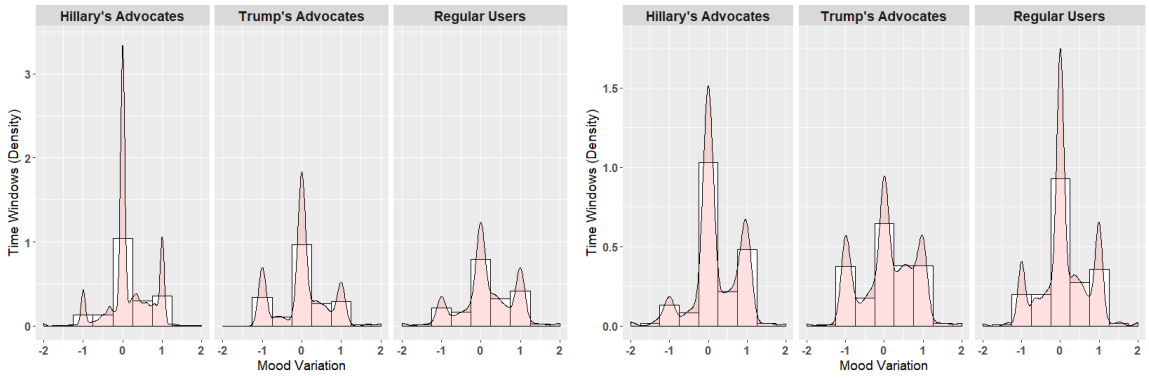
Figure 8 shows the distributions of mood variation for users in each of the three groups. Figure 5.8(a) considers the mood variation associated with tweets posted by Hillary Clinton, while Figure 5.8(b) shows mood variation for tweets from Donald Trump. Overall, all distributions are more concentrated on small values (in absolute terms) of ΔS_u, implying that no mood variation or, at best, only a small variation occurs more often. This situation happens for all user groups and tweets from both candidates. However, some differences across the user groups and both sources of tweets are worth noting.

For instance, in general, both Hillary's advocates and regular users tended to experience a positive mood variation more often than a negative one after retweeting a tweet from Hillary. Note, for example, the peaks around $\Delta S_u = +1$ in the left and right graphs of Figure 5.8(a)). They are higher than the corresponding peaks around $\Delta S_u = -1$ in the same graphs. Trump's advocates, on the other hand, tend to experience a negative mood variation more often in the same scenario (see the middle graph of Figure 5.8(a)).

On the other hand, the mood variation around tweets from candidate Donald Trump had a different, somewhat unexpected, effect. For advocates of the candidate, positive mood variations occurred only slightly more often than negative ones (see for instance the large peak around $\Delta S_u = -1$ in the middle graph of Figure 5.8(b)). However, for both Hillary's advocates and regular users, the distributions are much more skewed towards positive variations. It is hard to explain such behavior. One could speculate that perhaps the frequent use of sarcasm and irony could affect the sentiment analysis of tweets from such users (especially Hillary's advocates).

5.6 Homophily Analysis

As a final analysis of the user groups, we present the homophily analysis considering the three connection types. Figure 9, shows the results. Each chart refers to a specific connection type and the



(a) Hillary Clinton's tweets (b) Donald Trump's tweets

Figure 8: Mood variation of users for time windows δ=2 horas before and after a candidate's tweet.

group's IH_i value. For each group, we calculate two IH_i values: (i) considering only unidirectional connections and (ii) considering only reciprocal connections among users. In the following sections, we discuss the results referring to follow connections, mention connections, and retweet connections. In Section 5.6.4 we perform a homophily comparative analysis considering the three different connection types.

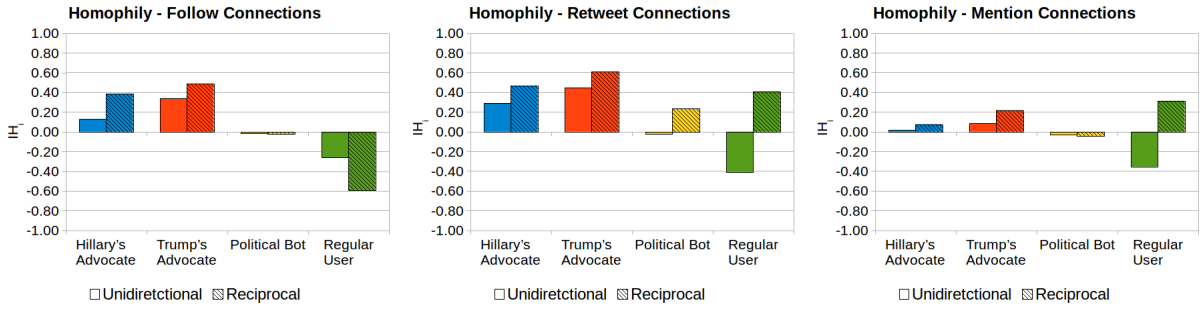


Figure 9: Different connections homophily

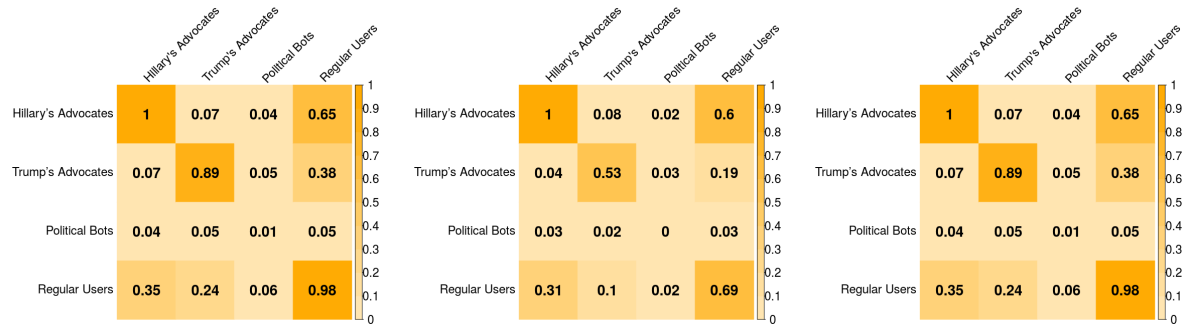
Table 7 presents the percentage of connections between same group's users considering only unidirectional values and reciprocal connections values referring to following, mention and retweet types.

Table 7: Connections among groups

Group	**Follow**		**Retweet**		**Mention**	
	Uni.	**Rec.**	**Uni.**	**Rec.**	**Uni.**	**Rec.**
Hillary's Advocates	36%	55%	48%	61%	28%	32%
Trump's Advocates	44%	57%	54%	67%	23%	34%
Political Bots	3%	2%	2%	27%	1%	0%
Regular Users	42%	27%	35%	73%	38%	68%

Figure 10 presents the total connections among groups normalized by scope: follow (Figure

5.10(a)) - the horizontal lines represent the followers, and the vertical lines represent the followed; retweet (Figure 5.10(b)) - the horizontal lines represent the retweeters, and the vertical lines represent the retweeted; and mention (Figure 5.10(c)) - presents the horizontal lines represent the mentioners, and the vertical lines represent the mentioned.



(a) Follow Connections (b) Retweet Connections (c) Mention Connections

Figure 10: Connections among groups

The reciprocity in a relationship is a factor that indicates a higher level of proximity between users [Colleoni, Rozza e Arvidsson 2014]. There are 419,969 reciprocal follow connections, 8,331 reciprocal retweet connections, and 8,725 reciprocal mention connections. Therefore, the proportion of reciprocal follow connections is more meaningful than in other types of connections among users since there is a 31.28% reciprocity and 0.69% retweet reciprocity and 0.47% mention reciprocity.

5.6.1 Follow Connections

There is homophily only in Trump's Advocates and in Hillary's Advocates. The homophily level in these groups is even higher when analyzing only reciprocal connections among users. The Regular Users has strong heterophily, indicating that unlike the advocates, they tend to follow users of others groups more frequently. Political Bots has heterophily, however, very close to zero (baseline that determines homophily), which demonstrates that Political Bots tends to follow one another close to the expected by chance in the network [McPherson, Smith-Lovin e Cook 2001].

The highest IH_i happens among Trump's Advocates, though Hillary's Advocates also demonstrate a significant homophily level. Note that 55% of the reciprocal Hillary's Advocates connections occurs among their peers (Table 7). It is a value close to the reciprocal follow connections of Trump's Advocates, where 57% of the connections occur among them.

Even though the Trump's Advocates and Hillary's Advocates have almost the same H_i among themselves, Trump's Advocates IH_i value is higher than Hillary's Advocates IH_i value. It happens because Trump's Advocates total is almost half of the Hillary's Advocates total.

5.6.2 Mention Connections

Figure 9 shows that unidirectional connections has a significant difference comparing with the reciprocal connections. When analyzing the unidirectional links, there is low homophily among the candidates' advocates, heterophily close to the baseline among the Political Bots, and heterophily among the Regular Users. However, the scenario is entirely different when we analyze only reciprocal mention connections. Occurs homophily when considering reciprocal mentions among Trump's Advocates, Hillary's Advocates, and Regular Users. We also note that occurs heterophily among Political Bots. Regular User have IH_{i} equal to -0.35 when analyzing only unidirectional mention connections, which indicates a strong heterophily. When examining only reciprocal mentions, IH_{i} reaches 0.31, which represents a high homophily level. Both Trump's Advocates and Hillary's Advocates homophily level increases when considering only the reciprocal mention connections. Nevertheless, the Political Bots reciprocal mention homophily level does not increase.

5.6.3 Retweet Connections

Note that occurs a change from heterophily in unidirectional connections to homophily in reciprocal connections among Regular Users and Political Bots. Trump's Advocates and Hillary's Advocates have homophily either using unidirectional connections or reciprocal connections, though the homophily is higher in the reciprocal connections.

Table 7 shows that this scenario presents the highest levels of homophily for all groups when analyzing only reciprocal connections. In this scenario, the highest levels of homophily occur for advocates when considering the unidirectional connections.

5.6.4 Discussing Homophily Results

After analyzing the homophily considering follow, mention, and retweet connections, we find some general and specific characteristics that describe the different interaction types among Twitter users. A phenomenon that became evident is that in all scenarios homophily becomes stronger when we analyze only the reciprocal connections, corroborating the work of Colleoni et al. [Colleoni, Rozza e Arvidsson 2014] who found similar results for political homophily on Twitter. However, they considered only the follow connection type. Therefore, we demonstrate that mention and retweet connections also exhibit similar characteristics of follow connections.

Another common characteristic is that both Trump's Advocates and Hillary's Advocates have homophily in all analyzed scenarios. The same does not happen among Regular Users since they only have homophily in reciprocal mention and retweet connections. Political Bots only have homophily considering reciprocal retweet connections. Trump's Advocates have higher levels of homophily than Hillary's Advocates.

Once in an election campaign, one of the advocate's objective is to promote their candidate to as many people as possible [Ranganath et al. 2016], homophily may represent a problem because it highlights a limit in reaching people with different political behavior. However, homophily may also indicate high advocates' political activism since they usually tend to have more contact with people who share their political bias. According to [Colleoni, Rozza e Arvidsson 2014], we observe that there is an echo chamber among Trump's Advocates and among Hillary's Advocates.

Once one of the Political Bots' goals is to reach a large number of real users in a social network [Ferrara et al. 2016], they are more likely to succeed when they do not have homophily. This way, the Political Bots in our database reach one of their objectives since there is no occurrence of homophily for them when analyzing either unidirectional connections or reciprocal connections except considering reciprocal retweet connections. However, we note that they do not usually establish reciprocal retweet connections with real people [Ferrara et al. 2016].

The Regular Users connect in a diversified way when analyzing only unidirectional connections. In other words, this type of user often creates relationships with Trump's Advocates, Hillary's Advocates, and Political Bots (Figure 10). The occurrence of heterophily in mention and retweet unidirectional connections and the high level of homophily in the mention and retweet reciprocal connections indicate that Regular Users have unidirectional interactions with users of different groups and more interactions among themselves considering mentions and reciprocal retweets connection types (Table 7). We understand that the change of heterophily in the unidirectional connections for homophily in the reciprocal cases of mention and retweet connections are explained by non-political features, for example, company, city, personal preferences, and others. Although Regular Users mention and retweet more advocates and Political Bots, the reciprocity of mention and retweet may indicate closer proximity between them on Twitter, even though they are not directly engaged in a political campaign.

5.7 Threats to Validity

We acknowledge that one limitation of our work is that the Twitter API provides at most the last 200 tweets published by a user, which could compromise the representativeness of the analyzed data set. Another limitation of this work is the risk to have many tweets that contain sarcastic political content since the SentiStrength does not recognize sarcasm. This way, some users that had a high mean sentiment toward the candidates, in fact, had a low mean sentiment toward them. However, detecting sarcasm is an open problem in the sentiment analysis research area [Silva, Coletta e Hruschka 2016].

6 CONCLUDING REMARKS

In this master thesis, we characterize user groups, their mood variation due to candidates' tweets on Twitter during the 2016 US presidential election, and also their homophily. We identify Trump's Advocates, Hillary's Advocates, and Regular Users applying the K-means algorithm; and the Political Bots using BotOrNot tool [Varol et al. 2017]. Next, we present a characterization of the users of each group considering six aspects of users behavior: which features highlight each group, the daily mean sentiment towards candidates, language patterns, the top 5 most retweeted users of each group, mood variation analysis, and homophily analysis. Our characterization results uncover some interesting observations, in each of these analyses.

- **Which features highlight each group:** Advocates tend to have a much stronger political discourse as well as use more political hashtags and have stronger biases towards either candidate (especially Hillary) than Regular Users. Hillary Advocates tend to post tweets regarding Hillary more often, and have lower variability in the sentiment expressed by tweets regarding either Hillary or Trump. Trump's Advocates mention Trump more often, although the sentiment scores of all their political tweets tend to vary more.

- **Mean Sentiment Analysis:** Trump's Advocates and Hillary's Advocates used to post positive tweets about the candidate they support and negative tweets about the other candidate. We also noted that Regular Users used to post positive tweets in the non-political context and that Trump's Advocates and Hillary's Advocates publish negative tweets in the non-political context.

- **Language Patterns:** Trump's Advocates are more engaged posting tweets with hashtags related to the candidate's campaign they support than Hillary's Advocates. We observed that in the four groups, the most used word was "trump" pointing that the person of Donald Trump and his presidential campaign were frequent topics of debate in all groups.

- **Popular users of each group:** Four of the five most popular users of Hillary's Advocates were official US media users. The most popular Trump's Advocates were users who on their own Twitter profile described themselves as Trump's supporters. The five most popular Political Bots were users with recently created accounts, and Twitter suspended all of them. Four of the most popular Regular Users were from known personalities.

- **Mood Variation Analysis:** Both Hillary's advocates and regular users tended to experience a positive mood variation more often than a negative one after retweeting a tweet from Hillary. Trump's advocates, on the other hand, tend to experience a negative mood variation more often in the same scenario. On the other hand, the mood variation around tweets from candidate

Donald Trump had a different, somewhat unexpected, effect. For advocates of the candidate, positive mood variations occurred only slightly more often than negative ones. However, for both Hillary's advocates and regular users, the mood variations are much more positive.

- **Homophily Analysis:** There was homophily among Trump's Advocates and among Hillary's Advocates on all types of connections analyzed, whether for unidirectional or reciprocal connections. This shows that advocates are more connected and interact more frequently with other users who support the same candidate. Political Bots presented heterophily close to the baseline (zero) for all types of connections, except when analyzing homophily in reciprocal retweet connections demonstrating that Political Bots connect with more people regardless of whether they are advocates or whether they are Regular Users. However, Political Bots retweeted themselves more often, which demonstrates that they do not frequently establish retweet reciprocity with non-bot users. The Regular Users had heterophily in all unidirectional connections and homophily in the retweet and mention reciprocal connections. We understand that this change can be explained by non-political characteristics, such as company, location, personal preferences, etc. Although Regular Users mention and retweet more advocates and Political Bots, the reciprocity of mention and retweet may indicate a proximity between them on Twitter, even though they are not directly engaged in a political campaign.

As future work, we intend to characterize our dataset using other analysis, such as, network metrics, characterize the users by their gender, age, and ethnicity using Twitter user profiles photos; and perform a temporal characterization correlating mood variation with external events that may have influenced the users sentiments. We also intend to refine the groups of Political Bots and Regular Users to identify new groups, such as Trump's Bots, Hillary's Bots, and groups related to social movements. Furthermore, we intend to analyze homophily considering multiplex connections and similar language patterns as one type of connection.

REFERENCES

BARBERA, P. et al. Tweeting from left to right: Is online political communication more than an echo chamber? PSYCHOLOGICAL SCIENCE, v. 26, n. 10, p. 1531–1542, 2015. PMID: 26297377. Disponível em: <https://doi.org/10.1177/0956797615594620>.

BHATTACHARYYA, P.; GARG, A.; WU, S. F. Analysis of user keyword similarity in online social networks. SOCIAL NETWORK ANALYSIS AND MINING, Springer, v. 1, n. 3, p. 143–158, 2011.

BISHOP, C. M. PATTERN RECOGNITION AND MACHINE LEARNING (INFORMATION SCIENCE AND STATISTICS). Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738.

BOLLEN, J. et al. Happiness is assortative in online social networks. CORR, abs/1103.0784, 2011. Disponível em: <http://arxiv.org/abs/1103.0784>.

BRADY, W. J. et al. Emotion shapes the diffusion of moralized content in social networks. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, Proceedings of the National Academy of Sciences, v. 114, n. 28, p. 7313–7318, Jun 2017. ISSN 1091-6490. Disponível em: <http://dx.doi.org/10.1073/pnas.1618923114>.

BRAND, M. Pattern discovery via entropy minimization. In: . [S.l.]: Morgan Kaufmann, 1998.

CAETANO, J. A. et al. Utilizando análise de sentimentos para definição da homofilia política dos usuários do twitter durante a eleição presidencial americana de 2016. In: VI BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING. Porto Alegre - RS, Brazil: SBC, 2017. (BraSNAM 2017).

CHOUDHURY, M. D.; COUNTS, S.; GAMON, M. Not all moods are created equal! exploring human emotional states in social media. In: . Association for the Advancement of Artificial Intelligence, 2012. Disponível em: <https://www.microsoft.com/en-us/research/publication/not-all-moods-are-created-equal-exploring-human-emotional-states-in-social-media/>.

COLEMAN, J. Relational analysis: the study of social organizations with survey methods. HUMAN ORGANIZATION, Society for Applied Anthropology, v. 17, n. 4, p. 28–36, 1958.

COLLEONI, E.; ROZZA, A.; ARVIDSSON, A. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. JOURNAL OF COMMUNICATION, Wiley Online Library, v. 64, n. 2, p. 317–332, 2014.

CURRARINI, S.; JACKSON, M. O.; PIN, P. An economic model of friendship: Homophily, minorities, and segregation. ECONOMETRICA, Wiley Online Library, v. 77, n. 4, p. 1003–1045, 2009.

DEGROOT, M.; SCHERVISH, M. PROBABILITY AND STATISTICS. Addison-Wesley, 2012. ISBN 9780321500465. Disponível em: <https://books.google.com.br/books?id=4TlEPgAACAAJ>.

DEMASI, O.; MASON, D.; MA, J. Understanding communities via hashtag engagement: A clustering based approach. INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA, 2016. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13115>.

DICKERSON, J. P.; KAGAN, V.; SUBRAHMANIAN, V. S. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING (ASONAM), 2014 IEEE/ACM INTERNATIONAL CONFERENCE ON. [S.l.: s.n.], 2014. p. 620–627.

DIENER, E. Subjective well-being. the science of happiness and a proposal for a national index. AM PSYCHOL, 2000. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/11392863>.

EASLEY, D.; KLEINBERG, J. NETWORKS, CROWDS, AND MARKETS: REASONING ABOUT A HIGHLY CONNECTED WORLD. [S.l.]: Cambridge University Press, 2010.

FERRARA, E. et al. The rise of social bots. COMMUN. ACM, ACM, New York, NY, USA, v. 59, n. 7, p. 96–104, jun. 2016. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2818717>.

FERRARA, E. et al. Predicting online extremism, content adopters, and interaction reciprocity. In: ____. SOCIAL INFORMATICS: 8TH INTERNATIONAL CONFERENCE, SOCINFO 2016, BELLEVUE, WA, USA, NOVEMBER 11-14, 2016, PROCEEDINGS, PART II. Cham: Springer International Publishing, 2016. p. 22–39. ISBN 978-3-319-47874-6. Disponível em: <http://dx.doi.org/10.1007/978-3-319-47874-6$_3$¿.

GIACHANOU, A.; CRESTANI, F. Like it or not: A survey of twitter sentiment analysis methods. ACM COMPUT. SURV., ACM, New York, NY, USA, v. 49, n. 2, p. 28:1–28:41, jun. 2016. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/2938640>.

HALBERSTAM, Y.; KNIGHT, B. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. JOURNAL OF PUBLIC ECONOMICS, Elsevier, v. 143, p. 73–88, 2016.

HERNANDEZ-SUAREZ, A. et al. Predicting political mood tendencies based on twitter data. In: 2017 5TH INTERNATIONAL WORKSHOP ON BIOMETRICS AND FORENSICS (IWBF). [S.l.: s.n.], 2017. p. 1–6.

HUBER, G. A.; MALHOTRA, N. Political homophily in social relationships: Evidence from online dating behavior. THE JOURNAL OF POLITICS, v. 79, n. 1, p. 269–283, 2017. Disponível em: <https://doi.org/10.1086/687533>.

JIN, S.; ZAFARANI, R. Emotions in social networks: Distributions, patterns, and models. In: PROCEEDINGS OF THE 2017 ACM ON CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT. New York, NY, USA: ACM, 2017. (CIKM '17), p. 1907–1916. ISBN 978-1-4503-4918-5. Disponível em: <http://doi.acm.org/10.1145/3132847.3132932>.

KLEIN, D.; MANNING, C. D. et al. Fast exact inference with a factored model for natural language parsing. ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, MIT; 1998, p. 3–10, 2003.

KWAK, H. et al. What is twitter, a social network or a news media? In: PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. New York, NY, USA: ACM, 2010. (WWW '10), p. 591–600. ISBN 978-1-60558-799-8. Disponível em: <http://doi.acm.org/10.1145/1772690.1772751>.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: PROCEEDINGS OF THE FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, VOLUME 1: STATISTICS. Berkeley, Calif.: University of California Press, 1967. p. 281–297. Disponível em: <http://projecteuclid.org/euclid.bsmsp/1200512992>.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. ANNUAL REVIEW OF SOCIOLOGY, JSTOR, p. 415–444, 2001.

MINGOTI, S. ANÁLISE DE DADOS ATRAVÉS DE MÉTODOS DE ESTATÍSTICA MULTIVARIADA: UMA ABORDAGEM APLICADA. Editora UFMG, 2005. ISBN 9788570414519. Disponível em: <https://books.google.com.br/books?id=W7sZllHmmGIC>.

MISLOVE, A. et al. You are who you know: inferring user profiles in online social networks. In: ACM. PROCEEDINGS OF THE THIRD ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING. [S.l.], 2010. p. 251–260.

MITRA, T.; COUNTS, S.; PENNEBAKER, J. Understanding anti-vaccination attitudes in social media. INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA, 2016. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13073>.

RANGANATH, S. et al. Understanding and identifying advocates for political campaigns on social media. In: PROCEEDINGS OF THE NINTH ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING. New York, NY, USA: ACM, 2016. (WSDM '16), p. 43–52. ISBN 978-1-4503-3716-8. Disponível em: <http://doi.acm.org/10.1145/2835776.2835807>.

RIBEIRO, F. N. et al. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ DATA SCIENCE, v. 5, n. 1, p. 23, 2016. ISSN 2193-1127. Disponível em: <http://dx.doi.org/10.1140/epjds/s13688-016-0085-1>.

ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. COMPUT. APPL. MATH., Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 20, n. 1, p. 53–65, nov. 1987. ISSN 0377-0427. Disponível em: <http://dx.doi.org/10.1016/0377-0427(87)90125-7>.

SEBASTIANI, F. Machine learning in automated text categorization. ACM COMPUTING SURVEYS, v. 34, p. 1–47, 2002.

SILVA, N. F. F. D.; COLETTA, L. F. S.; HRUSCHKA, E. R. A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM COMPUT. SURV., ACM, New York, NY, USA, v. 49, n. 1, p. 15:1–15:26, jun. 2016. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/2932708>.

SUBRAHMANIAN, V. S. et al. The darpa twitter bot challenge. COMPUTER, v. 49, n. 6, p. 38–46, June 2016. ISSN 0018-9162.

THELWALL, M. et al. Sentiment in short strength detection informal text. J. AM. SOC. INF. SCI. TECHNOL., John Wiley & Sons, Inc., New York, NY, USA, v. 61, n. 12, p. 2544–2558, dez. 2010. ISSN 1532-2882. Disponível em: <http://dx.doi.org/10.1002/asi.v61:12>.

VAROL, O. et al. Online human-bot interactions: Detection, estimation, and characterization. INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA, 2017. Disponível em: <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587>.

WANG, H. et al. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In: PROCEEDINGS OF THE ACL 2012 SYSTEM DEMONSTRATIONS. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (ACL '12), p. 115–120. Disponível em: <http://dl.acm.org/citation.cfm?id=2390470.2390490>.

WONG, F. M. F. et al. Quantifying political leaning from tweets, retweets, and retweeters. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, v. 28, n. 8, p. 2158–2172, Aug 2016. ISSN 1041-4347.

YUAN, G. et al. Exploiting sentiment homophily for link prediction. In: PROCEEDINGS OF THE 8TH ACM CONFERENCE ON RECOMMENDER SYSTEMS. New York, NY, USA: ACM, 2014. (RecSys '14), p. 17–24. ISBN 978-1-4503-2668-1. Disponível em: <http://doi.acm.org/10.1145/2645710.2645734>.

ZAKI, M. J.; JR, W. M. DATA MINING AND ANALYSIS: FUNDAMENTAL CONCEPTS AND ALGORITHMS. New York, NY, USA: Cambridge University Press, 2014. ISBN 0521766338, 9780521766333.