Caetano *et al. Journal of Internet Services and Applications* (2018) 9:18
https://doi.org/10.1186/s13174-018-0089-0

Journal of Internet Services and Applications

RESEARCH Open Access

Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election

Josemar A. Caetano*, Hélder S. Lima, Mateus F. Santos and Humberto T. Marques-Neto

Abstract

This paper proposes an analysis of political homophily among Twitter users during the 2016 American Presidential Election. We collected 4.9 million tweets of 18,450 users and their contact network from August 2016 to November 2016. We defined six user classes regarding their sentiment towards Donald Trump and Hillary Clinton: *whatever, Trump supporter, Hillary supporter, positive, neutral,* and *negative*. Next, we analyzed their political homophily in three scenarios. Firstly, we analyzed the Twitter *follow, mention* and *retweet* connections either unidirectional and reciprocal. In the second scenario, we analyzed multiplex connections, and in the third one, we analyzed friendships with similar speeches. Our results showed that *negative* users, users supporting Trump, and users supporting Hillary had homophily in all analyzed scenarios. We also found out that the homophily level increase when there are reciprocal connections, similar speeches, or multiplex connections.

Keywords: Internet, Online social networks, Sentiment analysis, Homophily

1 Introduction

The 2016 American Presidential Election was characterized by an intense competition, especially after the primaries of political parties that resulted in the dispute between Donald Trump, representing the Republican Party, and Hillary Clinton, representing the Democratic Party [34]. The political conflict between these two candidates was reflected in the discussions among users on online social networks like Twitter [27]. Although candidates' tweets can reach a large number of users, disputed debates in Twitter and in other social networks shows that not all users have the same sentiment regarding the candidates messages [38].

In this paper, we defined six Twitter user classes during the 2016 American Presidential Election that include sentiment variations that a user can have towards Donald Trump and Hillary Clinton. Figure 1 shows the division of Twitter users into politically engaged users and not politically engaged users. We named the first class as Whatever since its users did not engage in politics. A politically engaged user can be either a supporter or a non-supporter. Thus, we named the second class as Trump Supporter and the third one as Hillary Supporter. We defined the three non-supporter classes Positive, Neutral, and Negative as classes that have users expressing positive, neutral, and negative sentiment towards both candidates, respectively. We collected 4.9 million tweets published by 18,450 users, their profiles and their relationships with other users.

After classifying the Twitter users, we analyzed how connected each user is with his/hers peers using 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 [12]. The characteristics of peers (e.g., friends) can be immutable, such as ethnicity, or even mutable such as beliefs, professions [22], and sentiments towards a topic [40] or towards political candidates [6].

We analyzed homophily in three scenarios: (i) uniplex connections, (ii) multiplex connections, and (iii) friendship with a similar speech. A uniplex network is a network where there is only one type of connection between nodes; a multiplex network is a network with

*Correspondence: josemar.caetano@sga.pucminas.br
Department of Computer Science, Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, MG, Brazil

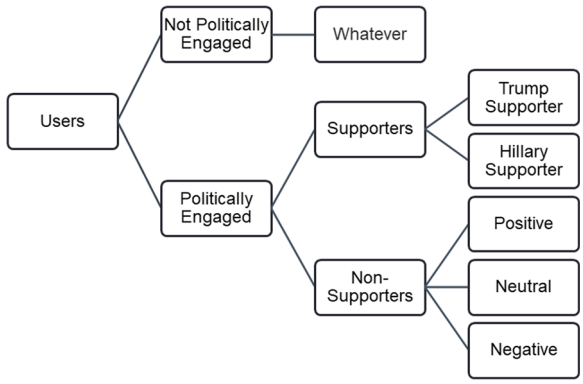

Fig. 1 Six classes of Twitter users defined using sentiment polarity towards Donald Trump and Hillary Clinton

more than one type of connection between nodes; friendship with a similar speech is a network where users follow each other and use hashtags and words in common. In this work, we considered the following Twitter user connections: follow (user A follows user B), mention (user A mentions user B), and retweet (user A retweets user B). These connections can be either unidirectional or reciprocal.

Our results revealed that, in all analyzed scenarios, Negative, Trump Supporter, and Hillary Supporter users had homophily. We also found out that the homophily level tends to increase when there are only reciprocal connections among users of those three classes. The level of homophily also increases when we analyzed the multiplex connections. Another circumstance that also increases homophily is when the users of those classes present similar speeches.

We hope that this work contributes to a better understanding of the political engagement of users in online social networks during the 2016 American Presidential Election. We also believe that the methodology proposed here is replicable on next elections in the US, and also in another country's political elections.

The paper is organized as follows. We first discuss the related work. Next we present our methodology divided into five: (i) collecting Twitter data; (ii) identifying political and non-political tweets; (iii) performing sentiment analysis in political and non-political tweets; (iv) identifying each user class; and (v) analyzing political homophily in Twitter. We then discuss the results for each scenario analysed, and then present our main conclusions and possible directions for future work.

2 Related work

Some papers have already proposed political homophily analysis on Twitter. In this section, we present some previous work related to the present paper. Colleoni, E., Rozza, A., and Arvidsson, A [8] investigated the political homophily in an American Democratic and Republican voters database. They used machine learning and social network analysis approaches to classify users' party preference. The authors noticed that, the Democrats show a higher level of political homophily when comparing with the Republicans. They also found out that homophily levels are higher when reciprocal connections are considered. 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, G. A., and Malhotra, N [18] investigated the political homophily on online dating sites. Unlike the previous works, 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 welcoming 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, Y., and Knight, B. [17] 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, they are exposed to more information than users connected to minority political groups and they receive the information faster than users connected to minority political groups.

Brady, W. J. et al.[4] researched how the propagation of polemic content happens over twitter social network. The authors did not focus on homophily itself although their findings are related to the present work. They found that polemic discourses are much more likely to be retweeted when analyzing the intra-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 Barberá, P. et al. [1].

Barberá, P. et al. [1] proposed a statistical model for political ideology estimation based on ideological positions. 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

to share information with 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 those groups, although this finding can also be related to a higher homophily once the groups are composed of more similar individuals than expected by chance.

Monti et al. [24] modeled the political disaffection on twitter, they randomly selected 50,000 Italian Twitter users, and collected their followers. The dataset analyzed contained 261,313 users and more than 35 million tweets from those users. The authors classified tweets as political (related to politics), negative (has negative sentiment expressed) and general (tweet do not mention any candidate). They considered as politically disaffected only tweets, political, negative, and general. They applied different classifiers to automatic identify political disaffection in tweets. Their results showed that Random Forest presented the best result on classification. They validated and compared their results with public opinion surveys regarding vote intentions and political topics. They found out a strong relationship between their classifier and the public opinion surveys. In the present work, we analyzed not only the disaffection but also the affection for one or both candidates. We considered only political tweets to classify users. Additionally, we investigated the homophily of each group to understand how much are they connected, and each user in those groups were classified based on the sentiment he/she expressed towards a candidate.

In a preliminary paper [6], we performed the political homophily analysis classifying users by the average sentiment expressed towards the candidates considering only follow connections. We analysed two homophily scenarios. In the first scenario, we defined four sentiment classes related to the candidates Donald Trump and Hillary Clinton and in the second scenario, we defined six sentiment classes. Our results showed that the existence of homophily among users that expressed negative sentiment towards Donald Trump and Hilary Clinton. The homophily was higher among users that had average negative sentiment towards Donald Trump. There was heterophylly among users that didn't publish tweets about the candidates and among users with neutral average sentiment toward them. The main differences between the present work and [6] is the way we classify the users and the use of multiplex connections, we analyzed the retweet and mention features as possible connections in the uniplex connections, and we analyzed the similar speech homophily in Twitter using hashtags and the most important words given by the LDA algorithm [2].

These papers demonstrated that political homophily is a phenomenon present on Twitter, therefore in this paper, we perform a more comprehensive homophily analysis considering uniplex, multiplex connections, and similar speeches (as connections) among groups. Thus, we considered mentions, retweets, and similar speeches in Twitter as connections in addition to the usual following connections investigated in related works. Moreover, we perform the homophily analysis among six sentiment classes: Whatever, Neutral, Negative, Trump Supporter, Hillary Supporter, and Positive.

3 Methodology

Figure 2 shows the steps of the methodology followed in this work. In the first step, we have collected Twitter data. Section 3.1 describes the process of collecting timelines, connections among users, and Twitter user profiles. In the second step, we performed an analysis of each users' timeline to define which tweets are about politics and which tweets are not about politics. The details of this identification process are described in Section 3.2. In the third step of the proposed methodology, we performed the tweets' texts sentiment analysis which has two distinct approaches: one for political tweets that were about both candidates at the same time and another one for non-political tweets and political tweets that were about only one candidate.

The political users' timelines and sentiment analysis allowed us to classify each user. Thus, in the fourth step of the methodology, we arranged users into six different classes: *Whatever, Trump Supporter, Hillary Supporter, Positive, Neutral,* and *Negative*. The process used for classifying the users is explained in Section 3.4. Finally, in the fifth and last step of the proposed methodology, we analyzed the users' political homophily regarding their friendships, retweets, and the text features (hashtags and

Fig. 2 Methodology steps

important words) of their tweets. We present the political homophily analysis in Section 3.5.

3.1 Data collection

In this section we begin presenting the Twitter social network, its features and nomeclatures and information about Donald Trump and Hillary Clinton use of Twitter. Finally, we discuss how we collected data from it.

3.1.1 Twitter and the 2016 american presidential election

Twitter is a social network that allows the publication of tweets - short messages, which, in 2016, were limited to 140 characters [16]. Released in 2006, it had about 328 million active users in 2017 and it was a very popular online social networks during the 2016 American Presidential Election [31].

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 necesarily this one will follow his/her follower.

A Twitter user profile is composed of the following attributes: name, profile description, photo, and location. A timeline is the set of tweets that a Twitter user 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, identified by the username @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. 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.

3.1.2 Collecting data from twitter

We collected data on Twitter social network from August 1, 2016 to November 30, 2016 using the official Twitter API [36] to get tweets, user profiles, and their contact networks. It is noteworthy that the American Presidential Election occurred on November 8, 2016 and that televised debates between Donald Trump and Hillary Clinton happened on September 26, 2016, October 9, 2016, and October 19, 2016.

This API provides up to 200 tweets (per request) published by a user. Additionally, the API has a limit of 300 requests per 15 minutes time window. We started the data collection identifying *seed users*, that is, people who were publishing content about the American election on Twitter. We obtained this identification through the API's streaming method, which enables real-time collection of tweets. The objective of the real-time collection was to identify users who published tweets in English and did a retweet of a candidate's tweet. We based on the hypothesis that if a user retweeted a candidate tweet, then they read that tweet and made a reference to what the candidate said. Therefore, this user is probably using Twitter to discuss, or promote political discussions.

For each seed user, we collected their profile, timeline, and their contact network (followers and friends), that is, the network first hop. We also collected the same information from each user of seed user's contacts network (network second hop). In total, we collected data from 18,450 users (5284 seed users and 13,166 users from the first hop and second hop). Figure 3 shows the steps followed in the data collection.

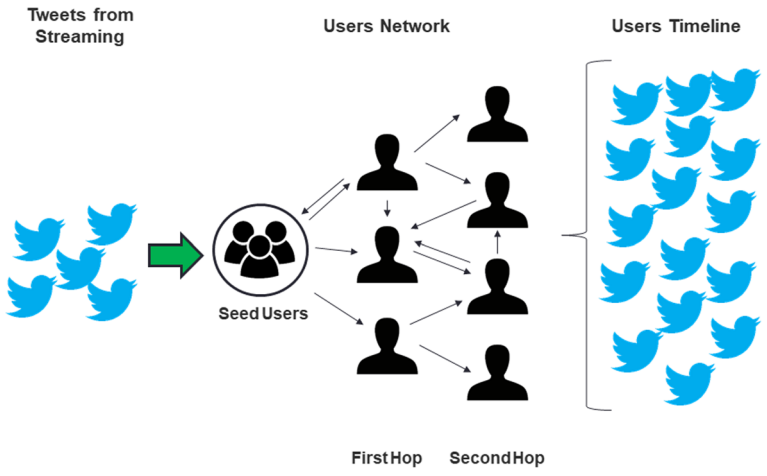

Fig. 3 Data collection steps

3.1.3 Dataset

Table 1 shows the total of tweets, user profiles, and relationships collected. We collected 4,935,128 tweets published by 18,450 users and 437,515 connections (follow relationship) among them. From the 437,515 relationships, 192,933 were unidirectional and 122,291 were reciprocal.

3.2 Political analysis

A prerequisite for analyzing tweets is to estimate the political bias of the users involved [39]. One way to extract a tweet political bias is through its syntactic and semantic features. Several papers have proposed this two analysis to find user communities and different tweet categories [11, 13, 23, 25, 32].

We performed a semantic analysis to identify which tweets had political content.Therefore considering mentions, targets and URLs occurrence in them. We defined the following conditions for the identification of political tweets:

- The tweet is a candidate's retweet;
- The tweet targets at least one candidate;
- The tweet mentions at least one candidate;
- The tweet has a candidate's proper name;

Whether one of these conditions was satisfied for one candidate, then that tweet was considered as a political tweet for her/him. If any of these conditions were satisfied for both candidates, then that tweet was considered as political for both of them. We considered as indicators of candidate targeting or mentioning, the Twitter username, the name, and the name abbreviation. Thus, for the candidate Donald Trump, we considered the words: "@realDonaldTrump","Trump" and "DT". For the candidate Hillary Clinton, we considered the words: "@HillaryClinton", "Hillary","HC". We performed the syntactic and semantic analysis with all tweet words converted to lower case to prevent words written in a different case from being ignored.

3.3 Sentiment analysis

After identifying the political and non-political tweets in each user timeline, we performed the sentiment analysis on each tweet text. We considered that if a political tweet is about a candidate A, then the tweet's sentiment is towards candidate A. Thus, a political tweet's sentiment is an opinion about a candidate A.

However, we observed in our dataset that political tweets that were about both candidates at the same time were a challenge for analyzing their political sentiment. Because a whole text sentiment analysis only takes into account what was the user's sentiment when he/she published a tweet, not his/her opinion about each subject in the tweet text.

Thus, whether a user expresses a very positive opinion about his/her favorite candidate and at the same time expresses a very negative opinion about the other candidate, then the sentiment analysis algorithm would consider the tweet them neutral since the very negative sentiment cancels the very positive sentiment. Furthermore, in these tweets, we were not able to correctly identify the user's opinion about each candidate either if his/her sentiment was positive or negative.

To address this problem, we divided the sentiment analysis into two approaches. In the first approach, we performed a sentiment analysis considering the whole tweet text. Thus, the text sentiment was assigned to a candidate if and only if the tweet was about that candidate. We used the same approach for nonpolitical tweets. In the second approach, we identified the words related to each candidate if and only if the tweet was about both candidates. Thus, we were able to perform a sentiment analysis considering just the related words to them.

Figure 4 presents a diagram with the two strategies followed in the tweet sentiment analysis.

3.3.1 Identifying words associated with both candidates

We identified the subjects and words associated with Hillary Clinton and Donald Trump in the same tweet using the Stanford Parser tool [20]. It allows the identification of subjects in a text and their associated words. Then we identified what subjects were related to each candidate considering the definitions presented in Section 3.2.

Furthermore, we considered words associated with pronouns possibly related to candidates in the analysis. That means, we defined the pronouns "he" and "him" as related to the candidate Donald Trump and the pronouns "she" and "her" as related to the candidate Hillary Clinton.

Table 1 Dataset

Collection	Total of documents
tweets	4935128
users	18450
relationships	437515

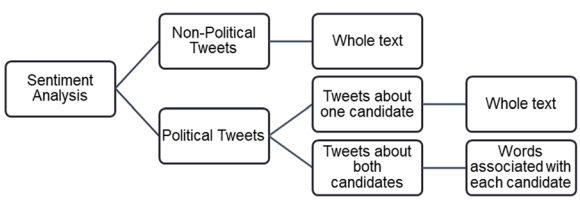

Fig. 4 The two sentiment analysis approaches followed

The words associated with the candidates' subjects and pronouns were used to build each candidate's bag of words. We performed a sentiment analysis in each candidate's bag of words to define the user's opinion towards the candidate.

The Stanford Parser library takes as its parameter a natural language parser model. The English-language parser template used in this work is called englishPCFG [19]. This template receives text as input and returns a list of tuples. Each tuple is a grammatical relation present in the text. The first position of a tuple is called *governor* and the second position of a tuple is called *dependency*. The tuple indicates that the word *governor* grammatically modifies the word *dependency*.

A grammatical relation can be a relation between subject and adjective, adjective and adverb, negation and adjective, etc. The library recognizes 50 grammatical relations. We selected the most relevant grammatical relations that would enable the identification of opinion words associated with each subject. We based the selection on the morphological classes covered by the SentiStrength lexeme dictionary [33].

3.3.2 Applying sentiment analysis in tweets

Some papers have proposed a sentiment analysis state of the art and practice strategies for different contexts and databases [16, 28, 30]. In [28], the authors presented a benchmark testing sentiment analysis tools in tweet databases. The results showed that SentiStrength tool [33] presented the best results.

Therefore, we performed the tweets' sentiment analysis using the SentiStrength tool. This tool uses a dictionary lexicon-based approach [16], that means, there is a dictionary of positive and negative words that are used to perform a text sentiment analysis.

SentiStrength analyzes the sentiment of each sentence in the text separately. At the beginning of the analysis, each sentence has a positive value α equal to 1 and a negative value β equal to -1. If the sentence contains words that belong to the word dictionary, then the values associated with those words are compared with α and β. If the word w sentiment value is greater than α, then α is updated to the w sentiment value. Similarly, if the w sentiment value is less than β, Then β is updated to the w sentiment value. A negative word has a value between -1 (slightly negative) and -5 (very negative). A positive word has a value between 1 (slightly positive) and 5 (very positive).

After analyzing each sentence, SentiStrength identifies the highest positive value α (max) and the smallest negative β value (min) of all sentences (whole text). The final sentiment analysis result is the difference (scale) between the values max and min. Therefore, the text sentiment can be between -4 (very negative) and 4 (very positive). The value 0 indicates a neutral sentiment.

SentiStrength dictionary consists of 700 words with a sentiment range from -5 (very negative) to 5 (very positive) [33]. In addition to the word dictionary, SentiStrength uses lists of emojis and boosting words (i.e., very, most, worst, best, etc.) to improve the sentiment analysis results [16].

3.4 Users sentiment classification

We calculated the average sentiment of each of the 18,450 users towards the candidates considering each user's political timeline (Section 3.2).

The process of averaging user sentiment towards candidates resulted in two scales, one referring to Donald Trump and the another to Hillary Clinton. The higher the value, the more positive the sentiment is, and the lower the value, more negative is the sentiment. Table 2 shows the average and standard deviation of the sentiment of the 18,450 users towards the candidates Donald Trump and Hillary Clinton.

In this paper, we assumed that a user has a positive sentiment towards a candidate when the average of the sentiment is greater than 0 and that a user has a negative sentiment towards a candidate when the average is less than 0. When the sentiment average has a 0 value, we considered that there is neutral sentiment towards the candidate. If a tweet is not political, then we considered as concealed sentiment (null value).

We assumed that there were at least six sentiment user classes in the 2016 American Presidential Election context acting on online social networks. Table 3 shows the name, description, and total of users of each one of the six classes defined. We assumed that each user from each class has more relationship with their peers and therefore, we wanted to analyze how similar a user is to his/her peers.

Figure 5 shows the total users of each class. The Whatever class contains 53.18% of users showing that more than half of the users did not publish any political tweets related to the candidates. Trump Supporter, Hillary Supporter, Positive, Neutral, and Negative classes have 5.43%, 19.00%, 0.55%, 12.35% and 9.49% of the total users, respectively.

3.5 Homophily analysis

To mathematically represent the homophily level of the kind i for each individual, Colleoni, Rozza and Arvidsson

Table 2 Average and standard deviation sentiments toward Donald Trump and Hillary Clinton observed in the dataset

	Donald Trump	*Hillary Clinton*
Mean	-0.130	-0.055
Standard Deviation	0.296	0.334

Table 3 Six classes of Twitter users definition

Class	Description
Whatever	Concealed sentiment towards both Donald Trump and Hillary Clinton
Trump Supporter	Positive sentiment towards Donald Trump and non positive sentiment towards Hillary Clinton; or negative sentiment towards Hillary Clinton and non negative sentiment towards Donald Trump
Hillary Supporter	Positive sentiment towards Hillary Clinton and non positive sentiment towards Donald Trump; or negative sentiment towards Donald Trump and non negative sentiment towards Hillary Clinton
Positive	Positive sentiment towards both Donald Trump and Hillary Clinton
Neutral	Neutral sentiment towards both Donald Trump and Hillary Clinton; or Neutral sentiment towards Donald Trump and concealed sentiment towards Hillary Clinton; or Concealed sentiment towards Donald Trump and neutral sentiment towards Hillary Clinton
Negative	Negative sentiment towards both Donald Trump and Hillary Clinton

[8] applied the following equation:

$$H_i = \frac{s_i}{s_i + d_i} \tag{1}$$

Where H_i is the homophily index, s_i represents the number of connections between i individuals (homogeneous connections), d_i represents the number of connections that bind individuals of kind i with individuals of other kinds (heterogeneous connections).

In this way, Currarini, Jackson and Pin [9] recommended to use inbreeding homophily index, developed by Coleman [7] to normalize the H_i. This measure is given by:

$$IH_i = \frac{H_i - w_i}{1 - w_i} \tag{2}$$

Where IH_i is the homophily index defined in Eq. 2 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, the stronger is the homophily occurrence.

The opposite of homophily is Heterophily since there is 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:

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

3.5.1 Multiplexity

McPherson, Smith-Lovin, and Cook [22] describe that people's social relationships are not entirely equal because there are different levels of closeness between individuals in society. For example, relationships between two people close to each other as marriage and friendship, or less close to each other as co-workers, schoolmates, neighbors, and acquaintances.

Multiplexity is the number of connection types linking two people [37]. Homophily patterns tend to be stronger when there are more types of relationships between people, that means, the higher the edges multiplexity, the higher the homophily level [22].

Fischer [15] was one of the pioneers in analyzing social networks with more than one kind of relationship. In that study, the authors investigated the differences in social relations between residents of small and large cities. They analyzed the impact of kinship ties, co-workers and neighborhood on people's social networks. An interesting finding was that, nonrelatives friends had a average

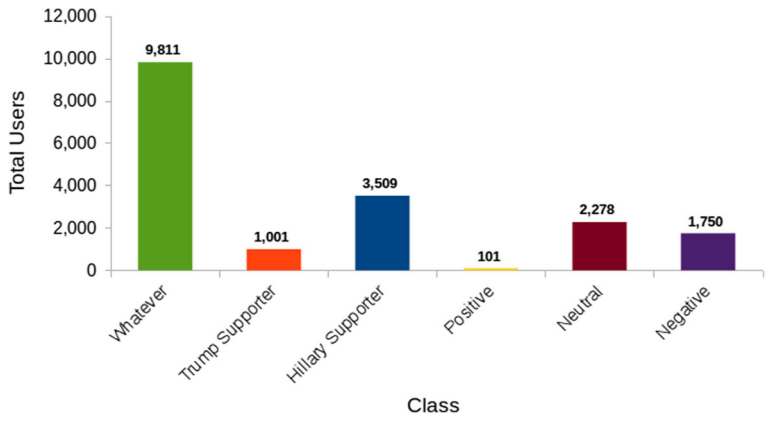

Fig. 5 Total users

age difference of 6 years, whereas among relatives, except siblings, the age difference was increased to 24 years. This example demonstrates that analyzing homophily by considering multiplexity assists in obtaining more refined information.

The term multiplex tie is commonly used to refer to occasions where there is an overlap of relationship types between two nodes in a network [14, 22, 37]. When there is only one kind of relationship between two nodes, one uses the term uniplex tie [14, 37] or simplex tie [22].

Figure 6 illustrates a Twitter user network containing multiplex ties. In this graph, the nodes correspond to the users and the edges are related to the interaction type among Twitter users (follow, retweet and mention). The relationships between node A and nodes B, C, and D are a uniplex tie since there is only one relationship among node A and other nodes. On the other hand, the relationship between nodes B and C is a multiplex tie, because there is more than one connection between them.

4 Experimental results

In this section, we first describe how we calculated the political homophily, next we present the user network considered for homophily analysis, then we present how we analyzed three different scenarios of political homophily on Twitter and their results.

We considered the following Twitter user connections in all scenarios: follow (user A follows user B), mention (user A mentions user B), and retweet (user A retweets user B). These connections can be either unidirectional or reciprocal. In each scenario, we measured homophily in two contexts: (i) considering only unidirectional edges and (ii) considering only reciprocal edges. The results obtained by [8] motivated our choice of using these two contexts. We present these scenarios results in Section 4.3.

In the first scenario, we calculated homophily considering only uniplex connections. In the second scenario, we analyzed homophily in a network only considering multiplex connections. Our goal was to verify if the multiplexity interferes in the classes' homophily level. We have a hypothesis that if a user has multiple connections with another user, then the homophily level is higher than with only one type of connection (e.g., connections between users B and C in Fig. 6). We describe the results obtained in this scenario in Section 4.4.

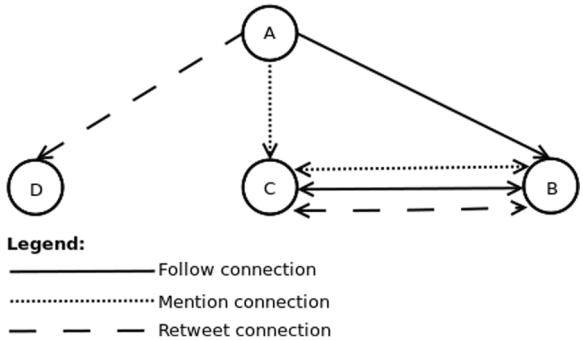

Fig. 6 The connection types between Twitter users

We were also interested in understanding whether Twitter users tend to post similar content and whether homophily intensifies in this scenario. Thus, we decided to analyze the hashtags used in their tweets, and also the most important words in their timelines given by the LDA algorithm [2].

Thus, in one homophily analysis scenario, we considered only reciprocal follow connections among users who had hashtags co-occurrence in their timeline. We present the results obtained in this scenario in Section 4.6.

In the third scenario, we considered only the reciprocal follow connections among users who had at least three most important words in common (according to the LDA algorithm). The results obtained are presented in Section 4.5.

4.1 Performing homophily analysis

We calculated the homophily indexes for these classes considering the scenarios previously presented. We calculated the H_i (Eq. 1) for each one of the Twitter user classes and we also calculated the index IH_i (Eq. 2). The index IH_i is useful for comparing the homophily level among different classes when their number of users is different. For example, to compare whether homophily between Hillary Supporter users is greater than the homophily between Trump Supporter users.

Note that we considered 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. Thus, w_i values were 0.5318, 0.1235, 0.0949, 0.0543, 0.1902, and 0.0055 in all scenarios that we analyzed for Whatever, Neutral, Negative, Trump Supporter, Hillary Supporter, and Positive, respectively.

4.2 Network description

We used a multiplex network, as presented in Section 3.5.1 to perform a more detailed homophily analysis considering the different interaction types among Twitter users. Thus, we defined the connection as: (i) follow, (ii) mention, and (iii) retweet. These connections represent the interaction among Twitter users. Figure 6 exemplifies the three connection types.

To perform the homophily analysis, we defined a network where the nodes represent users, and the edges represent the connections among them. This network contained 18,450 nodes and 795,986 edges. There is a

unidirectional follow connection between two users when only one of them follows the other. There is unidirectional retweet connection between two users when only one of them retweets the other, and there is a unidirectional mention connection with two users when only one of them mentions the other. Therefore, there is a reciprocal follow connection with two users when both follow each other. There is reciprocal retweet connection between two users when both retweet each other, and there is reciprocal mention connection between two users when both mention each other.

The multiplex network contains 795,986 connections: 437,515 of them were follow connections, and 302,588 of them were retweet connections, and 55,883 were mention type connections. We did not consider auto loop connections, that means, we only considered connections that were not own retweets and were not own mentions. Figure 7 shows the three types of relationships among the user classes (inner and outer edges pattern).

The Whatever users are followed, retweeted, and mentioned more than they follow, retweet, and mention other users. They are among the users that most mentioned their peers. We noted that although the Whatever class had the highest number of nodes in the network, there were few follow connections and retweet connections among their peers. Neutral, Trump Supporter, and Positive users had few connections with other users. In all connection types, Negative users had the highest number of connections with Hillary Supporter users. These totals were even higher than among their peers, despite their higher number of connections among peers. Hillary Supporter users had the highest number of follow connections, retweet connections, and mention connections among peers. Although the Trump Supporter users had few connections with Hillary Supporter users, the Hillary Supporter users retweeted and mentioned many users from Trump Supporter class.

The reciprocity in a relationship is a factor that indicates a higher level of proximity between users [8]. We carefully analyzed homophily in reciprocal relationship scenarios and noted that in our database there are 122,291 reciprocal follow connections, 4,030 reciprocal retweet connections, and 468 reciprocal mention connections. Therefore, the proportion of reciprocal follow connections is more meaningful than in other types of connections among users since there is 38.79% reciprocity and 1.35% retweet reciprocity and 0.84% mention reciprocity.

4.3 Homophily in uniplex connections scenarios

In this section, we present the homophily analysis considering the three connection types. Figure 8 shows the results obtained. Each chart refers to a specific connection type and the class' IH_i. For each class, we calculated 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 obtained referring to follow connections, retweet connections, and mention connections. In Section 4.3.4 we analyze the homophily results considering the three different connection types.

4.3.1 Follow connections

In this scenario, we noted that there was homophily in Negative, Trump Supporter and in Hillary Supporter classes. The homophily level in these classes is even higher when analyzing only reciprocal connections that occur among users. We noted that Whatever class had strong heterophily, indicating that unlike the Negative users, Trump Supporter users, and Hillary Supporter users, they tend to follow users of others classes more frequently.

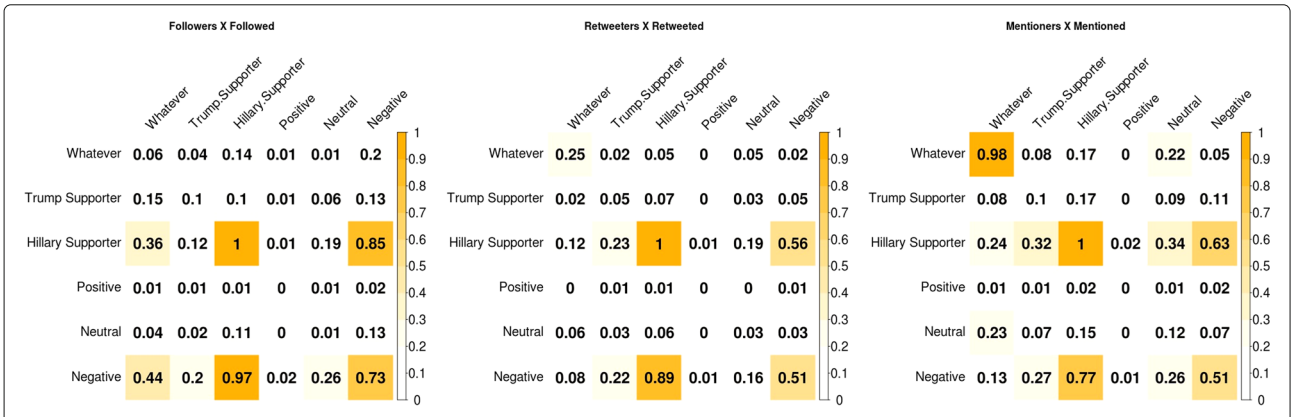

Fig. 7 Follow (left), retweet (center), and mention (right) connections among user groups. In the matrix on the left, the horizontal lines represent the followers, and the vertical lines represent the followed ones. In the matrix on the middle, the horizontal lines represent the retweeters, and the vertical lines represent the retweeted. In the matrix on the right, the horizontal lines represent the mentioners, and the vertical lines represent the mentioned ones. In all matrices, the cells contain the normalized total connections among classes

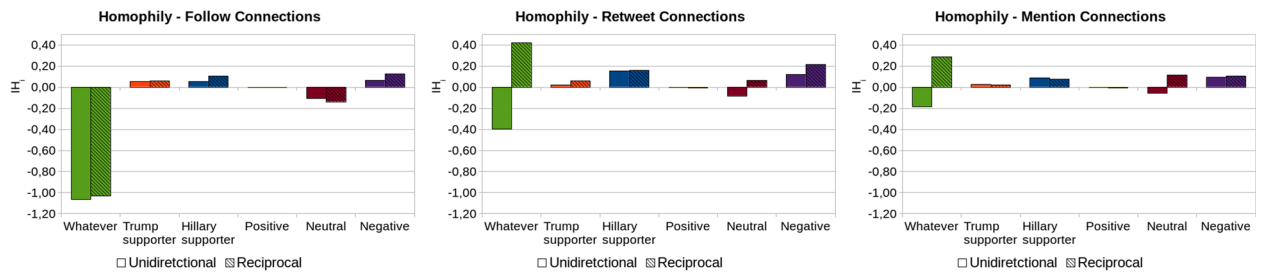

Fig. 8 Homophily in uniplex scenarios. Users classes' IH_i in three uniplex connection scenarios: follow connections (*left*), retweet connections (*center*) and mention connections (*right*). We considered the following analysis for each network: (i) only unidirectional connections and (ii) only reciprocal connections

The results also revealed that the Neutral and Positive classes had heterophily. However, the IH_i of Positive class is very close to zero (baseline that determines homophily), which demonstrates that Positive users tend to follow users from other classes as would be expected by chance in the network [22].

The highest IH_i occurred among Negative users. We also noted that 27.58% of the reciprocal Hillary Supporter users connections occurred among their peers (Table 4). Their IH_i is higher than the reciprocal follow connections of Negative class since 20.72% of the connections occurred among them. We noted that even though the Hillary Supporter class had the highest H_i value, the Negative class IH_i is higher than the Hillary Supporter class' IH_i. It occurs because Negative users total is almost half of the Hillary Supporter users total.

4.3.2 Retweet connections

In this scenario, all classes had homophily when analyzing only the reciprocal connections, except the Positive class. Negative, Trump Supporter, and Hillary Supporter classes had homophily either using unidirectional connections or reciprocal connections. The homophily was higher in the reciprocal connections among Negative and Trump supporter classes.

We noted that occurred a change from heterophily in unidirectional connections to homophily in reciprocal connections in Whatever class and Neutral class. Curiously the Whatever class had the lowest IH_i in unidirectional connections and the highest IH_i in reciprocal connections.

4.3.3 Mention connections

In this scenario, we analyzed homophily considering mention connection among users. We noticed that when analyzing the unidirectional links, there was homophily among the Negative users, Trump Supporter users, and Hillary Supporter users; heterophily close to the baseline among the Positive users and heterophily among the Whatever users and Neutral users.

Figure 8 shows unidirectional connections had a significant difference compared to the reciprocal connections of Whatever and Neutral classes. The results showed that the Whatever class had the lowest IH_i in unidirectional mentions. When considering just the reciprocal mentions, the IH_i reaches the highest value when compared with other classes. We also observed a similar phenomenon in the Neutral users. The Negative, Trump Supporter, and Hillary Supporter classes also had homophily in reciprocal connections. However, the homophily level increased only for Negative class when compared with the results in reciprocal connections. The Positive had homophily level close to the baseline in reciprocal connections.

Table 4 H_i of user classes represented like percentages

User class	Follow connections		Mention connections		Retweet connections	
	Unidirectional	Reciprocal	Unidirectional	Reciprocal	Unidirectional	Reciprocal
Whatever	3.25%	4.92%	44.41%	66.51%	34.59%	72.86%
Trump Supporter	10.37%	10.78%	7.62%	7.50%	7.39%	11.11%
Hillary Supporter	23.36%	27.58%	26.19%	25.13%	31.30%	31.77%
Positive	0.40%	0.37%	0.28%	0.00%	0.38%	0.00%
Neutral	2.79%	0.22%	7.10%	22.43%	4.93%	17.98%
Negative	15.30%	20.72%	18.05%	19.08%	20.14%	28.94%

The values represent the percentage of connections among users from the same class considering unidirectional connections and reciprocal connections referring to follow, mention and retweet types

4.3.4 Discussing uniplex connections results

After analyzing the homophily in scenarios involving follow, retweet, and mention connections, we found some general and specific characteristics that describe the different interaction types among Twitter users. A phenomenon that became evident is that for most connection types the homophily becomes stronger when we analyzed only the reciprocal connections, corroborating the work of Colleoni et al. [8] who also identified political homophily on Twitter. However, Colleoni et al. [8] considered only the follow connection type. Therefore, we demonstrated that mention and retweet interactions also exhibit the same behavior.

Another recurrent characteristic is that, in all analyzed scenarios, Negative, Trump Supporter, and Hillary Supporter classes had homophily. Among these classes, Trump Supporter is the class with lowest homophily level. Trump Supporter class had IH_i values ranging from 0.02 to 0.06 in the reciprocal connections, while Negative and Hillary Supporter classes had IH_i values ranging from 0.08 to 0.21 in the reciprocal connections.

The Whatever users and Neutral users connect in a diversified way when analyzing only unidirectional connections (Fig. 7). When analyzing both mention and retweet connection types of these two classes, we noted that heterophily occurred in unidirectional connections and homophily in reciprocal connections indicating that Whatever users and Neutral users have unidirectional interactions with users of different classes and more interactions among their peers (Table 4). We understand that the change of heterophily in the unidirectional connections for homophily in the reciprocal connections can be explained by non-political features, for example, company, city, personal preferences, and others. Although Whatever users and Neutral users mention and retweet more users with political engagement, the reciprocity of mention and retweet may indicate closer proximity among their peers.

The Positive class had heterophily close to baseline for all connection types. This feature can be related to the class definition since users that are positive towards both Donald Trump and Hillary Clinton are not common [35]. Thus, we believe Positive users do not characterize an organized community, which reflects the levels of homophily presented by them.

4.4 Multiplex homophily scenarios

After analyzing the results obtained in the uniplex connections scenarios, we noticed that the levels of homophily vary according to the connection type. This finding motivated us to investigate the pattern of homophily in a scenario considering multiplex connections. We considered the hypothesis that homophily is higher among friends who have more than one type of interaction with each other.

In Twitter, there is not friendship type interaction explicitly, as it exists on Facebook and in other online social networks. Kwak et al. [21] describes that Twitter users do not follow other users just to establish a friendship, but instead to track news channels and get information of their interest. This user behavior can explain why we detected only 39% reciprocal follow connections.

We assumed that the reciprocal follow connections represent friendship relationships among users [5, 10, 29]. Therefore, we verified that in situations where exists an overlap of a friendship connection with a reciprocal mention or retweet connections, the homophily is higher than the homophily in cases where there is only a friendship connection.

Figure 9 shows the values of IH_i in uniplex friendship connections, which represents the baseline of our analysis. We also presented IH_i in multiplex friendship connections, which corresponds to the homophily level in cases where there are a friendship connection and some other type of connection among users. The values of IH_i increase when analyzing only the multiplex connections of Negative users, Trump Supporter users, and Hillary Supporter users validating the hypothesis that homophily increases when multiplex ties are analyzed. The Positive class presents a heterophily near the baseline.

There was heterophily among Whatever users and among Neutral users in both uniplex and multiplex connections. This result differs from what was presented in Section 4.3.2 and Section 4.3.3, where high homophily occurred in cases of reciprocity in the retweet connection and mention connection scenarios, respectively. We have the hypothesis that many users who do not have political engagement usually mutually mention or retweet without necessarily being friends.

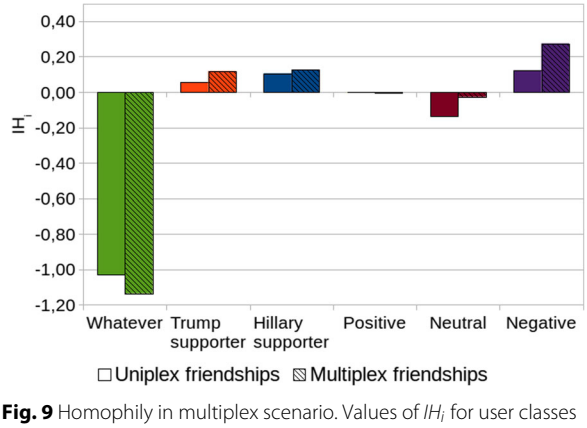

Fig. 9 Homophily in multiplex scenario. Values of IH_i for user classes in a scenario that compares homophily between a uniplex network and a multiplex network

The results of the homophily analysis in the uniplex and multiplex connections also validate the higher homophily among Negative users. It is important to note that the multiplex friendship connections represented only 525 cases, which corresponds to about 0.43% of the total number of friendships.

4.5 Homophily among users with similar speeches

In this section, we present the results obtained for scenarios that check homophily among users with similar speeches. In Section 4.6 we discuss results regarding friends who only used hashtags in common and in Section 4.7 we analyze homophily among friends who only used common important words (given by the LDA algorithm) in their tweets.

4.6 Enhancing friendship with hashtags usage

We found that 48,848 out of 122,291 friendships involve users who used some common hashtag in their timeline corresponding to about 40% of friendships. This result shows hashtags are indeed a prevalent feature on Twitter [11]. Figure 10 presents IH_i values of the friendships that have users who published some common hashtags and compares them with the friendships of users who did not have hashtags in common.

We found out that values of IH_i increase among Negative, Trump Supporter, and Hillary Supporter users when we analyzed only the friendships among users who used hashtags in common reinforcing the homogeneity observed in these groups.

The Positive IH_i values demonstrate that they had heterophily close to the baseline. The Whatever and Neutral classes had heterophily independently of having friends with hashtags in common or not.

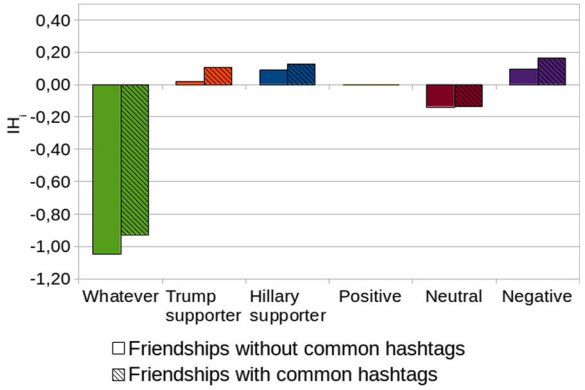

Fig. 10 Homophily among friends with common hashtags. IH_i values for user classes in a scenario that compares: (i) homophily among users who had friendships and did not publish hashtags in common to (ii) users who have friendships and published hashtags in common

4.7 Enhancing friendship with most important words in common

We applied the Latent Dirichlet Allocation (LDA) algorithm [3] in all user timelines to find out the main topic of each user and then we aggregated this words to obtain the most frequent words (according to the LDA algoritm) of each class. The LDA is a probabilistic model for topic detection that uses the Bayesian approach to learn the latent structure of topics that comprise a given text. We removed stop-words and lowercase letters, and then tokenized all tweets of each user. We ran the LDA algorithm using a Python library for topic modeling called `gensim` [26]. We choose $k = 1$ as the number of topics for the LDA algorithm, since we were interested to obtain the main topic of each user timeline. Each topic returned by the LDA algorithm contains words ordered by importance. Table 5 shows the five most frequent important words of each class.

The Whatever users did not have any most frequent words directly related to politics. The two Political Bots most frequent words were about politics. The word "trump" was the most frequent word in all politically engaged classes indicating that Donald Trump was frequently subject of discussion in these classes. The word "hillary" was the second most frequent word in Trump Supporter, Hillary Supporter, and Positive classes indicating that Hillary was also subject of discussions between those users but not as frequently discussed as Donald Trump.

We identified the 15 most important words of each user's timeline. Then, we defined that there is a connection between two users when both used at least three most important words in common. We considered the words in common occurrence as a homophily scenario because we have a hypothesis that if a user has a friendship with another user, then they probably use words in common and they are more related to each other. The same most important words can also indicate that they were talking about the same topic.

We found out that 34,791 out of 122,291 friends connections involve users who have most important words in common. This corresponds to about 28.5% of friendships between users. Figure 11 shows the IH_i values

Table 5 Most frequent words of each class

Whatever	Neutral	Negative	Trump Supporter	Hillary Supporter	Positive
today	trump	trump	trump	trump	trump
day	today	today	hillary	hillary	hillary
love	vote	vote	president	election	vote
time	good	good	obama	vote	president
life	love	love	america	president	america

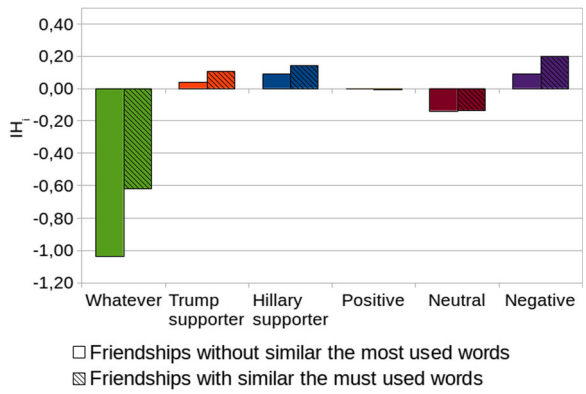

Fig. 11 Homophily among friends with most important words in common. The IH_i values for user classes in a scenario that considers the homophily among users that had at least three most important words in common

for friendships among users who used most important words in common and a comparison with friendships among users who did not use most important words in common.

We noted that the IH_i values increased among Negative users, Trump Supporter users, and among Hillary Supporter users when we analyzed only the friendships among users who had most important words in common corroborating the homogeneity observed in those groups.

In the Positive class, the IH_i values indicate an invariant homophily near the baseline. The Whatever and Neutral classes also demonstrated heterophily independently of their friends published tweets with most important words in common or not.

After analyzing the homophily among friends who published hashtags and among friends that had the most important words in common, we found that the homophily among the Negative, Trump Supporter, and Hillary Supporter reached the highest homophily levels observed in this work. Therefore, users with similar tweet text features (hashtags and important words) have more political homophily.

5 Threats to validity

We are aware one limitation of this work is the risk to have many tweets that contain sarcastic political content in our dataset since the SentiStrength does not ensure sarcasm recognition. Thus, some users that had a high average sentiment toward the candidates, in fact, had a low average sentiment toward them. However, detecting sarcasm is an open problem in the sentiment analysis research area [30].

6 Conclusion

In this paper, we used sentiment analysis to perform an analysis of the political homophily phenomenon on Twitter during the 2016 US presidential election. We collected tweets, user profiles and contact networks over 122 days (08/01/2016 to 11/30/2016). We used sentiment analysis to identify six user classes on Twitter: Whatever, Neutral, Negative, Trump Supporter, Hillary Supporter, and Positive. Then, we analyzed the political homophily in these classes. We defined two types of networks to analyze homophily: uniplex and multiplex. A uniplex network is a network where there is only one type of connection among nodes, and a multiplex network is a network with more than one type of connection among nodes. We defined as Twitter user connections: follow (user A follows user B), mention (user A mentions user B), and retweet (user A retweets user B). These connections can be either unidirectional or reciprocal. We used the metric IH_i [7] to calculate the homophily in the identified groups. We analyzed homophily in three scenarios: (i) uniplex connections, (ii) multiplex connections and (iii) friendship with a similar speech.

In the first scenario, we analyzed the homophily for each type of connection in uniplex networks. We identified that there is homophily among user classes with political engagement (Negative, Trump Supporter, and Hillary Supporter) in all types of connections analyzed, whether for unidirectional or reciprocal connections. Therefore, users with political engagement are more connected and interact more frequently with other users from the same class. The Positive class presented heterophily close to the baseline for all types of connections. This demonstrates that Positive users, despite presenting political speeches, are not so connected with one another and do not build communities. The Whatever users and Neutral users had heterophily in all unidirectional connections and homophily in the retweet reciprocal connections and mention reciprocal connections. We understand that this change can be explained by non-political characteristics, such as company, location, personal preferences, etc.

In the second scenario, we compared homophily in uniplex connections with homophily in multiplex connections. Our goal was to check whether homophily among users who have more than one type of interaction is enhanced. To perform the analysis in these scenarios, we considered that two Twitter users have a friendly relationship when they follow each other. Thus, we considered that multiplex connection exists when in addition to the friendship relationship there is also a retweet or mention reciprocal relationship; and exists a uniplex connection when there is only friendship relationship between two users. The results showed that the IH_i in multiplex connections increased for Neutral, Negative, Trump Supporter, and Hillary Supporter classes when compared to the level of homophily obtained in uniplex connections. Positive users presented heterophily near the baseline for both uniplex connections and multiplex connections. The Whatever users had heterophily in both network types.

Therefore, we showed that homophily is enhanced for most classes when multiplex connections exist.

In the third scenario, we analyzed homophily among friends who had a common speech. We considered that the use of hashtags and the most important timeline words to determine their speech. We observed that there is homophily in Negative, Trump Supporter, and Hillary Supporter classes. We noted that, in both analyses, the heterophily among Whatever and Neutral users occurred when we analyzed the friends who had similar speeches. Positive users, as well as in most of the analyzed scenarios, maintained heterophily close to the baseline.

In most of the analyzed scenarios, Negative users had the highest homophily level. We concluded Negative users form a more homogeneous community, they are more engaged in the use of common hashtags, and they are mentioned and retweeted more often among peers.

As future work, we intend to characterize and analyze homophily involving other characteristics, such as day and time of the week with more frequency of messages. We also intend to analyze homophily by gender, age, and ethnicity through user profiles photos; perform a temporal political homophily analysis correlating it with external events that may have influenced the users sentiments. We also expect to enhance the user classification through data ming techniques to identify candidates' advocates, political bots, and other user classes.

Abbreviations
API: Application programming interface; LDA: Latent dirichlet allocation; URL: Uniform resource locator

Acknowledgements
We acknowledge colleagues of the Human Behavior and Software Engineering (HUB-SE) laboratory for their incentives and suggestions.

Funding
This work is supported by MASWeb (grant FAPEMIG/PRONEX APQ-01400-14), FAPEMIG (grant APQ-02924-16), PUC-Minas, CNPq, CAPES and STIC AmSud 18-STIC-07.

Availability of data and materials
Readers can contact the corresponding author to request the dataset used in this work. We would be pleased to provide it.

Authors' contributions
This work is the outcome of a research project at the Postgraduate Program of the PUC Minas developed by Josemar A. Caetano (JC), Hélder S. Lima (HL), and Mateus F. Santos (MS), and supervised by Humberto T. Marques-Neto (HM). JC and MS were responsible for collecting the Twitter data, for performing the political and sentiment analysis, and for classifying the Twitter users. HL and JC were responsible for performing the political homophily analysis. HM provided direction and guidance during all steps of the research. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 5 January 2018 Accepted: 20 June 2018
Published online: 03 September 2018

References
1. Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R. Tweeting from left to right: Is online political communication more than an echo chamber? Psychol Sci. 2015;26(10):1531–42. https://doi.org/10.1177/0956797615594620. https://doi.org/10.1177/0956797615594620.
2. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
3. Blei DM, Ng AY, Jordan MI, Lafferty J. Latent dirichlet allocation. J Mach Learn Res. 2003;3:2003.
4. Brady WJ, Wills JA, Jost JT, Tucker JA, Van Bavel JJ. Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci. 2017;114(28):7313–8. https://doi.org/10.1073/pnas.1618923114. http://dx.doi.org/10.1073/pnas.1618923114.
5. Brandt C, Leskovec J. Status and friendship: Mechanisms of social network evolution. In: Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion. New York: ACM; 2014. p. 229–230. https://doi.org/10.1145/2567948.2577327. http://doi.acm.org/10.1145/2567948.2577327.
6. Caetano JA, Lima HS, dos Santos MF, Marques-Neto HT. 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, BraSNAM 2017. Brazil: SBC, Porto Alegre - RS; 2017.
7. Coleman J. Relational analysis: the study of social organizations with survey methods. Hum Organ. 1958;17(4):28–36.
8. Colleoni E, Rozza A, Arvidsson A. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. J Commun. 2014;64(2):317–32.
9. Currarini S, Jackson MO, Pin P. An economic model of friendship: Homophily, minorities, and segregation. Econometrica. 2009;77(4):1003–45.
10. Davis Jr CA, Pappa GL, de Oliveira DRR, de L Arcanjo F. Inferring the location of twitter messages based on user relationships. Trans GIS. 2011;15(6):735–51.
11. DeMasi O, Mason D, Ma J. Understanding communities via hashtag engagement: A clustering based approach. In: ICWSM. Palo Alto: AAAI; 2016. p. 102–111.
12. Easley D, Kleinberg J. Networks, crowds, and markets: Reasoning about a highly connected world. New York: Cambridge University Press; 2010.
13. Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The rise of social bots. Commun ACM. 2016;59(7):96–104. https://doi.org/10.1145/2818717. http://doi.acm.org/10.1145/2818717.
14. Ferriani S, Fonti F, Corrado R. The social and economic bases of network multiplexity: Exploring the emergence of multiplex ties. Strateg Organ. 2013;11(1):7–34.
15. Fischer CS. To dwell among friends: Personal networks in town and city. Chicago: University of chicago Press; 1982.
16. Giachanou A, Crestani F. Like it or not: A survey of twitter sentiment analysis methods. ACM Comput Surv. 2016;28(2):28:1–41. https://doi.org/10.1145/2938640. http://doi.acm.org/10.1145/2938640.
17. Halberstam Y, Knight B. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. J Public Econ. 2016;143:73–88.
18. Huber GA, Malhotra N. Political homophily in social relationships: Evidence from online dating behavior. J Polit. 2017;79(1):269–83. https://doi.org/10.1086/687533. https://doi.org/10.1086/687533.
19. Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pp. 423–430. Stroudsburg: Association for Computational Linguistics; 2003. https://doi.org/10.3115/1075096.1075150. https://doi.org/10.3115/1075096.1075150.
20. Klein D, Manning CD, et al. Fast exact inference with a factored model for natural language parsing. Advances in neural information processing systems. Marylebone: MIT Press; 2003, pp. 3–10.
21. Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, pp. 591–600. New York: ACM; 2010. https://doi.org/10.1145/1772690.1772751. http://doi.acm.org/10.1145/1772690.1772751.

22. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. Annu Rev Sociol. 2001;27:415–44.
23. Mitra T, Counts S, Pennebaker JW. Understanding anti-vaccination attitudes in social media. In: ICWSM. Palo Alto: AAAI; 2016. p. 269–278.
24. Monti C, Rozza A, Zappella G, Zignani M, Arvidsson A, Colleoni E. Modelling political disaffection from twitter data. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '13, pp. 3:1–3:9. New York: ACM; 2013. https://doi.org/10.1145/2502069.2502072. http://doi.acm.org/10.1145/2502069.2502072.
25. Ranganath S, Hu X, Tang J, Liu H. Understanding and identifying advocates for political campaigns on social media. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM '16, pp. 43–52. New York: ACM; 2016. https://doi.org/10.1145/2835776.2835807. http://doi.acm.org/10.1145/2835776.2835807.
26. Řehůřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta: ELRA; 2010. p. 45–50. http://is.muni.cz/publication/884893/en.
27. Reuters. In breathless u.s. election, twitter generates buzz not cash. 2016. https://www.reuters.com/article/us-usa-election-twitter/in-breathless-u-s-election-twitter-generates-buzz-not-cash-idUSKCN12R2OV. Accessed 15 Dec 2017.
28. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci. 2016;5(1):23. https://doi.org/10.1140/epjds/s13688-016-0085-1. http://dx.doi.org/10.1140/epjds/s13688-016-0085-1.
29. Shin WY, Singh BC, Cho J, Everett AM. A new understanding of friendships in space: Complex networks meet twitter. J Inf Sci. 2015;41(6):751–64.
30. Silva NFFD, Coletta LFS, Hruschka ER. A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Comput Surv. 2016;49(1):15:1–26. https://doi.org/10.1145/2932708. http://doi.acm.org/10.1145/2932708.
31. statista. Most famous social network sites worldwide as of september 2017, ranked by number of active users (in millions). 2017. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed 15 Dec 2017.
32. Subrahmanian VS, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Zhu L, Ferrara E, Flammini A, Menczer F. The darpa twitter bot challenge. Computer. 2016;49(6):38–46. https://doi.org/10.1109/MC.2016.183.
33. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment in short strength detection informal text. J Am Soc Inf Sci Technol. 2010;61(12):2544–58. https://doi.org/10.1002/asi.v61:12. http://dx.doi.org/10.1002/asi.v61:12.
34. Times TNW. Some donald trump voters warn of revolution if hillary clinton wins. 2016. https://www.nytimes.com/2016/10/28/us/politics/donald-trump-voters.html. Accessed 15 Dec 2017.
35. Twitter. Clinton: Half of trump supporters 'basket of deplorables'. 2016. http://www.bbc.com/news/av/election-us-2016-37329812/clinton-half-of-trump-supporters-basket-of-deplorables. Accessed 15 Dec 2017.
36. Twitter. Twitter api docs. 2017. https://dev.twitter.com/overview/api. Accessed 15 Dec 2017.
37. Verbrugge LM. Multiplexity in adult friendships. Soc Forces. 1979;57(4):1286–309.
38. Vilares D, Thelwall M, Alonso MA. The megaphone of the people? spanish sentistrength for real-time analysis of political tweets. J Inf Sci. 2015;41(6):799–813. https://doi.org/10.1177/0165551515598926.
39. Wong FMF, Tan CW, Sen S, Chiang M. Quantifying political leaning from tweets, retweets, and retweeters. IEEE Trans Knowl Data Eng. 2016;28(8):2158–72. https://doi.org/10.1109/TKDE.2016.2553667.
40. Yuan G, Murukannaiah PK, Zhang Z, Singh MP. Exploiting sentiment homophily for link prediction. In: Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14, pp. 17–24. New York: ACM; 2014. https://doi.org/10.1145/2645710.2645734. http://doi.acm.org/10.1145/2645710.2645734.