AAFA: Associative Affinity Factor Analysis for Bot Detection and Stance Classification in Twitter

Saad Sadiq\textsuperscript{1}, Yilin Yan\textsuperscript{1}, Asia Taylor\textsuperscript{2}, Mei-Ling Shyu\textsuperscript{1}, Shu-Ching Chen\textsuperscript{3}, Daniel Feaster\textsuperscript{4}

\textsuperscript{1}Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL, USA
saadsadiq@miami.edu, y.yan4@umiami.edu, shyu@miami.edu

\textsuperscript{2}Department of Computer Science
University of Miami, Coral Gables, FL, USA
axt634@miami.edu

\textsuperscript{3}School of Computing and Information Sciences
Florida International University, Miami, FL, USA
chens@cs.fiu.edu

\textsuperscript{4}Department of Public Health Sciences, Miller School of Medicine
University of Miami, Miami, FL, USA
dfeaster@biostat.med.miami.edu

Abstract—The rise in popularity of social interacting websites such as Facebook, Twitter, and Snapchat has been challenged by the upsurge of unwelcomed and troubling bodies on these systems. This includes spam senders, malware systems, and other content contaminators. It is noted that highly automated accounts with 450 tweets per day produced almost 18\% of entire Twitter circulation in the 2016 U.S. Presidential election. It is also observed that those disruptive systems called bots are inclined more towards circulating negative news than positive information. This paper introduces a novel framework named Associative Affinity Factor Analysis (AAFA) designed for stance detection and bot identification. Using AAFA, the proposed framework identifies real people from bots and detects the stance in bipolar affinities. The 2016 U.S. Presidential election campaign was used as a test use case because of its significant and unique counter-factual properties. The results show that our proposed AAFA framework achieves high accuracy when compared to several existing state-of-the-art methods.

Keywords—Bot detection; stance classification; association affinity; factor analysis

I. INTRODUCTION

In recent years, there has been a major growth in the use of microblogging platforms. Microblogs allow the users to exchange small contents such as short videos, sentences, and links. Some previous research efforts were paid on these kinds of multimedia data \cite{1}--\cite{10}. Twitter is one of the most widely used microblog platforms. Users range from regular users to politicians, celebrities, and company representatives. Therefore, it is possible to collect posts of users from different social and interested groups. On the flip side, Twitter is littered with automated agents called chat bots. Chat bots are rudimentary software systems with minimal automation and basic conversation abilities. They direct their scripts on social media outlets to tirade, obscure the facts, or merely make the conversations cloudy. It is estimated that as many as 48 million Twitter accounts are bots and from the 19.4 million tweets during elections, 1300 tweets per day were produced by bots \cite{11}. As an example, these anonymous chat machines were an integral part of a prearranged effort to disturb the 2016 U.S. Presidential election. To visualize the overall picture, Figure 1 is used to show the daily tweet count between Hillary Clinton and Donald Trump for the election time period in 2016. It is worthy to note that 33\% of pro-Trump traffic was driven by bots and highly automated accounts, compared to 22\% for Clinton. The popularity measure between the two candidates is shown in Figure 2 by mapping the retweet and favorite counts of the two candidates. The objective of this paper is to build an automated system for bot detection in Twitter accounts by the proposed Associative Affinity Factor Analysis (AAFA) framework.

Another important research direction is stance analysis which implies the political tendency of the public. In this paper, “stance classification” is defined as automatically determining whether a Twitter user tends to endorse the candidate of Democratic or Republican Party. By tweets from the Twitter accounts, researchers can deduce whether a user is either for or against the target. Therefore, another objective of this paper is to automatically infer the stances of Twitter users to see whether a user is likely a Hillary Clinton or Donald Trump supporter. While most election predictions reply on polls, automated stance classification can be applied to a much larger number of samples and bring complementary information to predict the election results.

The remaining of the paper is organized as follows. In Section 2, some previous work on bot detection and stance analysis are briefly presented. Then, some domain
knowledge about the Twitter data and how to clean and extract the election dataset are introduced in Section 3. Section 4 shows the proposed framework in details. The experimental results are provided in Section 5, which proves the efficiency of AAFA. Finally, Section 6 concludes this paper with several future research directions.

II. PREVIOUS WORK

Based on our best knowledge, though bot detection and stance analysis have not been used for election prediction, some earlier work ran experiments that used Twitter hashtags and emoticons such as #bestfeeling, #epicfail, and #news to identify positive, negative, and neutral tweets to train and analyze the sentiment of a tweet [12]. The sentiments were identified as a powerful predictor in differentiating the behaviors of various accounts. Agarwal et al. [13] proposed a 3-way task of separating tweets into positive, negative, and neutral, and then used 3 models: unigram, feature-based, and tree kernel-based models to split the data. It was proposed in [14] to use a psychometric instrument to classify six mood states including tension, depression, anger, vigor, fatigue, and confusion. The authors used aggregated Twitter content to compute a six-dimensional mood vector for each day in the timeline. One challenge in Twitter analysis is to identify and collect the right corpus that corresponds well to the domain and context of the tweets. This was attempted in [15] to focus and improve the corpus by an automatic collection and by using TreeTagger for POS-tagging. The wide scale effects of socioeconomic events on the overall general mood of tweets were explored by [14] over the longer periods of time. This provides a useful yardstick to track the sentiments but this method does not solve the problem of context invariance. A significant impact was made by [16] by creating a 60-“honeypot” trap for 7 months to send gibberish tweets and consequently attracted 36000 fake Twitter accounts. They follow each other to avoid Twitter filters, resulting in thousands of followers among themselves.

A hypothesis was proposed in [17] that every non-hyperbolic tweet was from Donald Trump’s staff while every hyperbolic tweet was from Donald Trump himself. The researchers collected Donald Trump’s tweets from Donald Trump’s account including the “source” information and found out that most tweets are from either iPhones or Android phones. Their analysis showed that the iPhone and Android tweets are clearly from different people since tweets from them used different hashtags, retweeted in distinct ways, and were posted during different times. They also found that the iPhone tweets were less angry and more positive with benign announcements, while the Android tweets tended to be more negative with angry words. In [18], machine learning techniques were utilized to do sentiment analysis on candidates’ Twitter mentions. They collected millions of tweets posted by users who discussed U.S. politics for Americans and non-Americans worldwide, and classified them based on their sentiment. Each posted tweet related to Hillary Clinton or Donald Trump was labeled with either positive, neutral, or negative. The authors concluded that there were much more negative tweets about both candidates than positive tweets, while there were fewer tweets that mentioned Hillary Clinton than Donald Trump. In [19], two groups of hashtags were defined arbitrarily, where each group was assumed to support Hillary Clinton or Donald Trump, respectively. After that, the author used descriptive statistics methods and concluded that Donald Trump’s campaign knew more about how to use Twitter chat bots than the Hillary Clinton’s side.

III. TWITTER DATASET

A. Data Collection and Pre-processing

In order to do stance analysis for the 2016 U.S. election test use case, a dataset that includes the supporters of both sides is necessary. However, due to privacy issues, it is nearly impossible to get the account names of the supporters. Luckily, Wikipedia provides the lists of Hillary Clinton and Donald Trump presidential campaign endorsements [20]. These lists include “big names” who have publicly claimed their endorsements for the office of the president to Hillary Clinton or Donald Trump as their presidential nominees. Since these supporters are notable individuals, the information was reliable and did not change much in the campaign. After data cleaning, 310 supporters of Hillary Clinton and 412 supporters of Donald Trump were included to build the experimental dataset.

In addition, the Twitter API was used to collect 3240 tweets from each supporter with time, resource, retweet,
etc. After the data collection, we extracted the details of the supporters’ accounts, cleaned the text data from all tweets, and mapped the truncated words to get the hashtag information.

B. Counterfactual Bipartisanship

Identifying bots in a binary stance topic such as Hillary Clinton vs Donald Trump is a huge challenge. It is observed that the election dataset is a unique domain where people who support one candidate or party have counterfactual negative sentiments for the other party, i.e., the supporters of Hillary Clinton would predominantly be the haters of Donald Trump. Figure 3 shows the retweet counts of the month by month data between the two poles of our stance detection. The information contains followers’ count, favorites, retweet count, and other account metadata of their accounts. Figure 4 ranks the most important features from the Twitter metadata.

C. Generating Personal Bot Army

It is currently an active research problem to correctly identify a bot from a real person. Some of the infamous twitter bots participating in the 2016 U.S. elections went under the names like @keksecorg, @NeilTurner, @WhiteGenocideTM, etc. However, the list of verified bots is very small and it was very difficult to get a dataset that has the ground truth information. To solve this problem, we purchased 3000 fake bots from the social freelance marketplace called Fiverr [21]. Moreover, the metadata of their accounts was extracted and around 50 tweets were obtained for each account. The bots were advertised to perform the following tasks:

1) retweet specific hashtags
2) plagiarize posts from specific accounts
3) tweets about specified topics
4) retweet specified accounts
5) post curated links every 30 minutes
6) tag targeted user account in retweets

To compare these bots with real accounts, the friends’ counts were compared among Bots, Hillary Clinton’s supporters, and Donald Trump’s supporters in Figure 5. A stark separation was observed between human supporters and bot followers from the dataset. This shows that we have a strong case of detecting bots from humans in our dataset.

IV. THE PROPOSED FRAMEWORK

In this paper, an Association Affinity-based Factor Analysis (AAFA) framework [22]–[26] is proposed for the iden-
The proposed AAFA framework uses Multiple Factor Analysis (MFA) [39], [40], generally a combination of Principal Component Analysis (PCA) [41] and Multiple Correspondence Analysis (MCA) [42], for mixed-variable Twitter election datasets. MFA is implemented in two stages. Initially, a PCA is executed on a subset of feature space $j$, as shown in Figure 7. This is further standardized by dividing the weights of the features by $\lambda_j^1$, i.e., the first eigenvalue of set $j$. These normalized and unit variance principal components form the basis function of MFA of a dataset with mixed variables. Second, the categorical variables are

Figure 5: Density plot of the friend counts among Bots, Donald Trump’s and Hillary Clinton’s supporters

Figure 6: Proposed AAFA Framework

Figure 7: Combined dataset where quantitative and nominal/ordinal variables are juxtaposed
calculated as follows. To merge these two types of variables, (i.e., continuous and categorical variables in the analysis such as stratifying the variable categories. We then apply multiple correspondence analysis (MCA) to scale the variables and yield the eigenvalues.

Let us formulate this process by defining the dataset as \( I \) samples having \( J \) sets of data with mixed variables such that \( J = J_q + J_c \), where \( J_q \) and \( J_c \) refer to the sets of quantitative and categorical variables, respectively. Moreover, \( i \) refers to an arbitrary sample, \( k \) indicates a particular feature column, and \( j \) represents a group of features, such that at the crossing of row \( i \) and column \( k \), belonging to set \( j \), we have:

1) if \( j \) is a quantitative set, the value \( x_{ikj} \) of the variable \( k \) for the unit \( i \);
2) if \( j \) is a categorical set, \( z_{ikj} = 1 \) if \( i \) belongs to the category \( k \) and 0 if it doesn’t.

These standardized datasets are then combined together to build a distinct matrix. This balances the influence of both continuous and categorical variables in the analysis such that both variables can equally determine the dimensions of variability. To merge these two types of variables, (i.e., sets \( J_q \) and \( J_c \)), the equivalence between MCA and PCA is calculated as follows.

1) Applying Global PCA to the table with the general term \((z_{ikj} - w_{kj})/w_{kj}\);  
2) Assigning the weight \( w_{kj}/Q_j \) to column \( k \) of set \( j \);  
3) Assigning the weight \( p_i \) to row \( i \).

Here, \( z_{ikj} = 1 \) if \( i \) belongs to the category \( k \), and 0 otherwise. \( w_{kj} = \sum_{i \in I} p_i \cdot z_{ikj} \) with \( p_i \) being the uniformly distributed weight allocated to each sample \( i \), with a default value of 1. Furthermore, a distance is generated among units \( i \), and \( I \) refer to the total numbers of samples in cluster \( c \). The hierarchical clusters are depicted in the form of a dendogram ranked by the increase in the inertia. To explain the hierarchical divisions, we measure the correlation between each division (i.e., a categorical variable) and

1) each quantitative feature by the square correlation ratio \( \eta^2 \);
2) each categorical feature by the Cramer’s coefficient \( V \).

Cramer’s coefficient normalizes the \( \chi^2 \) statistic by using the maximum value of the \( \chi^2 \) statistic to divide it as given in Equation (3).

\[
V = \sqrt{\frac{\chi^2}{T \cdot \min(I-1,K-1)}} \tag{3}
\]

where \( \chi^2 \) is the chi-square statistic, \( T \) is the grand total of the table, and \( I \) and \( K \) refer to the total numbers of samples and features of the table.

C. Affinity-based Stance Detection

Consider each hashtag in a tweet as a concept and find the recurring itemset in Donald Trump’s retweets or comment feed. If we are able to find multiple instances of people continuously together based on the Association Affinity Network (AAN) [6], [46], [47], then they are bots. The confidence score is replaced by the average sentiment score. It was observed that bots usually have consistently positive or negative sentiments in their tweets. In addition, for real human supporters, individuals who endorsed Hillary Clinton tended to use different hashtags compared to those supported Donald Trump.

One attempt in the literature [19] built two groups of arbitrary hashtags, from their domain knowledge, to find the Hillary Clinton and Donald Trump supporters. However, this approach lacks reproducibility and domain invariance. To overcome this challenge and find distinct hashtags, we apply the log odds ratio approach [48]. For a hashtag \( n \),
we calculate $C^H_n$ and $C^T_n$, which represent the numbers of times $n$ was used by the Hillary Clinton supporters and Donald Trump supporters. Similarly, $U^H_n$ and $U^T_n$ represent the numbers of distinct accounts of the Hillary Clinton and Donald Trump supporters that used hashtag $n$. Next, the scores $S^C_n$ and $S^U_n$ are calculated to measure the likelihood values of a hashtag being associated with either of the candidates (as shown in Equations (4) and (5)).

$$S^C_n = \log_2 \left( \frac{C^H_n + 1}{\sum_{i=1}^{N} C^H_i + 1} \right) \left( \frac{C^T_n + 1}{\sum_{i=1}^{N} C^T_i + 1} \right);$$  

(4)

$$S^U_n = \log_2 \left( \frac{U^H_n + 1}{\sum_{i=1}^{N} U^H_i + 1} \right) \left( \frac{U^T_n + 1}{\sum_{i=1}^{N} U^T_i + 1} \right).$$  

(5)

Here, $N$ refers to the total number of supporters. The scores and the ranked hashtags are given in Table II. For comparison, the hashtag lists are shown by the domain knowledge [19] and the tweets from the candidates (i.e., Hillary Clinton and Donald Trump) [18] in Table I. It is clear that some unique hashtags can only be automatically found using the proposed framework.

V. EXPERIMENT AND RESULTS

Using the dataset extracted and the cleaned information, the experiments are conducted and three-fold cross validation is applied for the comparisons.

A. Results of Bot Detection

The clustering is based on the inertia gained when we go from one cluster to two clusters. We also get further significant inertia gain while going from two clusters to three clusters, which indicates that there are subgroups of human followers within the Hillary Clinton and Donald Trump camps. Hierarchical clustering performed on the principal components gives three kinds of insights, namely

1) the principal components
2) the projections of variables on these principal components
3) the variable associations and clusters

Figure 8 shows the graph of the quantitative variables after applying the PCA. The coordinate axes here represent the first two principal components and the arrows depict the cosine angles between the variables and the principal components. The percentage on each coordinate axis represents the proportion of variances retained by the each principal component. Variables close to the circle illustrate a high correlation with other variables and the direction of the variable vector indicates the correlation polarity between any two given variables.

Figure 8: Correlation circle for the continuous variables after performing PCA

Figure 9 illustrates the square correlation ratio (variable associations) between both types of variables, continuous and categorical, along the coordinate axes of the first and second principal components. The squared correlation ratio quantifies the correlation between the continuous and categorical variables. It was used in the framework to calculate one way ANOVA test of data separation. The proposed AAFA framework was performed on the Twitter dataset to
Table I: Ranked hashtags based on domain knowledge and tweets from the candidates (case insensitive)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Domain knowledge</th>
<th>Candidate tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>votehillary2016</td>
<td>DemsInPhilly</td>
</tr>
<tr>
<td>2</td>
<td>VoteHillary</td>
<td>MakeAmericaGreatAgain</td>
</tr>
<tr>
<td>3</td>
<td>NeverTrump</td>
<td>DebateNight</td>
</tr>
<tr>
<td>4</td>
<td>IAmWithHer</td>
<td>AmericaFirst</td>
</tr>
<tr>
<td>5</td>
<td>WeAreWithHer</td>
<td>MAGA</td>
</tr>
<tr>
<td>6</td>
<td>NeverHillary</td>
<td>V oteTrump</td>
</tr>
<tr>
<td>7</td>
<td>TrumpLikes</td>
<td>DemConvention</td>
</tr>
<tr>
<td>8</td>
<td>StopTrump</td>
<td>W omanCard</td>
</tr>
<tr>
<td>9</td>
<td>DumpTrump</td>
<td>EstoyConElla</td>
</tr>
<tr>
<td>10</td>
<td>TrumpUnfit</td>
<td>GOPDebate</td>
</tr>
</tbody>
</table>

Table II: Ranked hashtags based on the proposed framework (case insensitive)

<table>
<thead>
<tr>
<th>Rank</th>
<th># of hashtags used</th>
<th># of distinct accounts that use a hashtag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CIR Dobbs</td>
<td>TrumpPence16</td>
</tr>
<tr>
<td>2</td>
<td>RenewUs TrumpPence16</td>
<td>HoldTheFloor CrookedHillary</td>
</tr>
<tr>
<td>3</td>
<td>RaiseTheWage PJNET</td>
<td>RestoreTheVRA WakeUpAmerica</td>
</tr>
<tr>
<td>4</td>
<td>ActOnClimate WakeUpAmerica</td>
<td>VAWA PJNET</td>
</tr>
<tr>
<td>5</td>
<td>WomenSucceed TrumpTrain MarriageEquality V oteTrump</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>DoYourJob AmericaFirst WorldAidsDay Jesus</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>RestoreTheVRA ProLife</td>
<td>GunViolence TrumpRally</td>
</tr>
<tr>
<td>8</td>
<td>DisarmHate TeaParty ProtectOurCare Hannity</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>TimeIsNow MakeAmericaGreatAgain StopGunViolence Trump45</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>GetCovered ConfirmGorsuch LoveIsLove TrumpPence2016</td>
<td></td>
</tr>
</tbody>
</table>

Table III: Comparative evaluation of Multiple Factor Analysis with several leading models

<table>
<thead>
<tr>
<th>Classifier</th>
<th>F1-score</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>UnSupRF</td>
<td>0.73</td>
<td>0.58</td>
<td>0.95</td>
<td>0.56</td>
</tr>
<tr>
<td>BotOrNot</td>
<td>0.71</td>
<td>0.66</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>K-Means</td>
<td>0.75</td>
<td>0.61</td>
<td>1.00</td>
<td>0.60</td>
</tr>
<tr>
<td>ExpMax</td>
<td>0.93</td>
<td>0.92</td>
<td>0.88</td>
<td>0.99</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table IV: Accuracy and F-score comparisons in stance analysis

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.8089</td>
<td>0.7128</td>
</tr>
<tr>
<td>RF</td>
<td>0.8492</td>
<td>0.8156</td>
</tr>
<tr>
<td>DAC</td>
<td>0.7493</td>
<td>0.6190</td>
</tr>
<tr>
<td>Linear</td>
<td>0.7812</td>
<td>0.6853</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.7867</td>
<td>0.7061</td>
</tr>
</tbody>
</table>

identify bots vs humans, and compared with the following state-of-the-art unsupervised models in this domain:

1) Unsupervised Random Forest (UnSupRF) [49]
2) Truthy’s BotOrNot [38]
3) K-Means Clustering [50]
4) Fuzzy C-Means Clustering [51]
5) Expectation Maximization (ExpMax) [52]

As can be seen from Table III, the proposed framework achieves a significant improvement over those models in the comparison. For example, our proposed framework achieves 0.96 in accuracy in comparison to 0.66 from the industry’s goto platform Truthy’s BotOrNot.

B. Results of Stance Classification

To apply the affinity-based stance detection method, the hashtags of people supporting Hillary Clinton and Donald Trump were extracted. The associative affinity was evaluated for one-itemset and two-itemset hashtags occurring in their tweets. The final ranking of the predictive hashtags was ranked according to an empirically selected threshold. Each hashtag itemset has a dynamic threshold but the hashtags with the highest affinities were selected. Table II illustrates these case insensitive ranked hashtags for the two candidates. The final hashtag lists are selected based on both the “number of a hashtag being used” and the “number of distinct accounts that use a hashtag”. The overlapped hashtags are cleaned and finally a list of 128 hashtags is created to generate the feature vectors for the Hillary Clinton and Donald Trump supporters. Based on the number of a hashtag used, a feature vector is generated and normalized for each account.

For comparison, our stance classification model is evaluated against several popular classifiers including Support Vector Machine (SVM) [53], Random Forest (RF) [54], discriminant analysis classifier (DAC), Linear Regression, as well as Logistic Regression (LR) [55]. As shown in Table IV, an average accuracy of 80 percent is obtained without...
any domain knowledge and polls. Random Forest performs the best for this task due to the nature of our feature vectors (i.e., different weights for the hashtags).

An important insight is to observe whether the accuracy of a clustering method would be affected if we use real accounts with unknown predilections. This would help us also identify and evaluate the undecided voters. Currently, it is out of the scope of this paper to assert the ground truth for accounts having relative unknowns, but an extension of this framework will be to collect and extend the dataset with hand-labeled real accounts, and re-evaluate the stance detection and bot identification.

VI. CONCLUSIONS

The power of propaganda is reinforced when a limited number of individuals believe that it is prevalent. The part played by false news and fabricated information in the 2016 U.S. elections proved to be a painful experience for the information technology industry. This paper proposes a novel framework to detect the stance between the followers of the two dominant presidential candidates, Hillary Clinton and Donald Trump, and to separate real vs bot accounts. For our best knowledge, we are the first group that uses machine learning algorithms for stance analysis in election predictions. We are able to accurately identify the truth behind the number of Twitter followers and social media popularity by dissecting the real followers from paid bots. Our results show that the proposed framework is more accurate than the industry’s most popular tool, BotOrNot. In the future, other information in tweets including the resources, retweets, favorites, etc. would be also considered for better stance detection and bot classification.

REFERENCES


