User Behavior Modelling for Fake Information Mitigation on Social Web

Zahra Rajabi, Amarda Shehu, and Hemant Purohit

Volgenau School of Engineering, George Mason University, Fairfax VA 22030, USA

Abstract. The propagation of fake information on social networks is now a societal problem. Design of mitigation and intervention strategies for fake information has received less attention in social media research, mainly due to the challenge of designing relevant user behavior models. In this paper we lay the groundwork towards such models and present a novel, data-driven approach for user behavior analysis and characterization. We leverage unsupervised learning to define user behavioral categories over key behavior dimensions. We then relate these categories to content-based, user-based, and network-based features that can be extracted in near-real time and identify the most discriminative features. Finally, we build predictive models via supervised learning that leverage these features to determine a user's behavior category. Rigorous evaluation indicates that the constructed models can be valuable in predicting user behavior from recent activity. These models can be employed to rapidly identify users for intervention in mitigation strategies, crisis communication, and brand management.

Keywords: Disinformation, Fake News Mitigation, User Behavioral Model, Social Media Mining, Unsupervised Learning.

1 Introduction

In 2017, 67% of Americans reported that they obtained at least some part of their news on social media[1]. This massive burst of data is naturally accompanied with the threat of disinformation, such as spam and fake news spread by malicious intent users, affecting different aspects of democracy, journalism, and freedom of expression [11]. Fake news dissemination was best highlighted during the 2016 presidential election, in which the spread of fake stories favoring each party gravely threatened trust in government [2]. The propagation of fake information on social networks is now a recognized societal problem [14]. Despite many efforts, social networking platforms have yet to effectively address this challenge In particular, mitigating the dissemination of fake content is now a critical challenge for researchers across academia leading to emerging research areas of social cyber-security [4] and social cyber forensics [1,3].

In this paper, we present a novel, data-driven approach for user behavioral analysis and characterization that enables us to identify vulnerable users for

[1] http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

fake news mitigation. Our main contributions are: a.) Identification of key user behavior dimensions in reactions to the exposure of fake vs. fact information, specifically *initiation*, *propagation*, and *reception* behavior types. These dimensions allow us to organize users in behavior categories via unsupervised learning. b.) Validation of the hypothesis that behavior categories for users can be predicted by features extracted from shared content, user profile and activity, as well as the structural characteristics in the corresponding user interaction network. We also employ feature selection approaches to analyze the significance of our content-based, user-based, and network-based set of features to identify the most representative features of user behavior categories. These features are then used to build classification models to predict such categories. c.) Extensive experiments to evaluate state-of-the-art multiclass classification algorithms for user behavioral pattern prediction using a dataset collected from *Hoaxy* [8] platform. Our evaluations show that the predictive models demonstrate promising performance in categorizing users based on their reactions in response to fake/fact exposures, which consequently gives us an oversight to develop a solid baseline for designing an effective fake content mitigation strategy. In the remainder of this paper, we briefly present in Section 2 related works. Section 3 describes our proposed approach, followed by its evaluation and results in Section 4.

2 Related Work

In recent years, with increasing consumption of news over social media, the extreme consequences of fake information dissemination, from misleading election campaigns to inciting violence during crises, have led many researchers to focus on the problem of fake news detection [7,8,15]. Comprehensive reviews of this area of research in [14,17] show that existing studies mostly rely on static datasets to develop models based on supervised learning methods rather than online learning settings due to potential concept drifts. These approaches involve exploitation of user, content, and network-based information which inspired us in the feature design of our user behavioral modeling.

Users play a critical role as the creators and spreaders of fake content in social web. Therefore, assessing the credibility of users and modeling their behavior types could provide a valuable approach to design intervention strategies [12,6], which can optimize the dissemination of real news. For instance, [6] proposed an intervention framework using multivariate point process, however, authors did not consider the types of users and their behaviors. Given the uncertainty of user intent and activities, it is essential, although very challenging, to discriminate between malicious and naive users who unintentionally engage in fake content propagation. Therefore, modeling user behavior for identifying candidate vulnerable users for intervention strategies is an emerging research need.

While there exist extensive research on social media on user modeling and user credibility [10,9,16,17], the main goal of these studies has centered around content filtering for spam, improving user interest profiling for content and link recommendation systems, personalization in search, as well as influencer ranking.

Our research instead complements such user modeling research by investigating user behavior types to inform the mitigation strategies for the propagation of malicious, fake content.

3 Methods

We first represent the key behavior dimensions relevant to the mitigation task and then describe the unsupervised learning setup that allows elucidating the organization of users in different behavior categories. Once such categories are identified, information theoretic measures expose characteristics/features that best relate with the identified categories. Supervised learning methods then yield predictive models of behavior categories from such features.

3.1 Key Dimensions of User Behavior

We have identified three major behavioral dimensions to capture user reactions to fake over fact cascade exposures, namely initiating, propagating, and receiving (but no further action) fake content. We define $\texttt{FoF}_{\text{prop}}$, $\texttt{FoF}_{\text{received}}$, and $\texttt{FoF}_{\text{init}}$ to be log-ratio of fake over factual information respectively propagated, received or initiated by a user. As described in Section 4, these dimensions allow visualization of a three-dimensional semantic space as a baseline representation of user engagements within the network in response to different information cascade exposures. More importantly, they facilitate the application of unsupervised learning methods to identify behavioral categories based on user engagement in spread of fake versus fact cascades.

3.2 Identifying Behavior Categories via Unsupervised Learning

Clustering algorithms can group and categorize users within the three-dimensional space defined above. We consider several clustering algorithms such as kmeans, Agglomerative clustering, DBScan, and spectral clustering (*c.f.* survey on algorithms in [13]). In Section 4, we evaluate the performance of clustering algorithms (over different parameter values) along popular metrics, such as the Silhouette coefficient, the Calinski-Harabaz score, and the Davies-Bouldin score. These metrics do not rely on ground truth availability, which is our case.

Clustering Performance Evaluation. The Silhouette coefficient is calculated as $(b-a)/max(a,b)$, where a is the mean intra-cluster distance, and b is the mean nearest-cluster distance for each sample (user). This metric is computed as the mean Silhouette Coefficient of all samples ranging from -1 (worst) to 1 (best). The Calinski-Harabaz score, also known as the variance ratio criterion, is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. A higher Calinski-Harabaz score indicates a model with better-defined clusters. Unlike the Silhouette score, the Calinski-Harabaz score is unbounded;

the higher the score the better the cluster separation. The Davies-Bouldin is defined as the ratio of within-cluster distances to between-cluster distances and is bounded in $[0, 1]$. A lower score is better.

3.3 User Behavior Categories Representation

The above evaluation measures highlight the most effective clustering method and corresponding behavior categories, which can be used to label users. Our dataset has only ground truth for fake/fact content and no user labels for our analysis is provided. Therefore, our proposed approach of automatically labeling users with behavioral categories opens a way to supervised learning models that relate features to the discovered behavior categories/classes for users. Information-theoretic measures, such as Mutual Information (MI) measure is employed to evaluate characteristics/features of user nodes in the diffusion cascades that best relate with the identified behavior classes. A feature selection algorithm is employed to identify the most important features, and supervised learning methods are then utilized to build predictive models of behavior categories from the extracted features. In this section, we propose our data mining framework for user behavior category representation and prediction.

Features Extraction The features representing each user are extracted using the propagated retweets' content, user profiles, and network structure.

1. ***Content-based Features***:
   - `sentiment`: the average of sentiment intensity score of tweet texts shared by user is computed using Sentiment Analyzer tool in *nltk* (Natural Language Toolkit) Sentiment package.
   - `Tweet text length`: the average of length of tweet texts shared by user.
2. ***User-based Features***:
   - `followers_count`: the number of users following a user; it shows a Twitter accounts popularity;
   - `friends_count`: the number of users a user is following; it informs the users interest-driven participation;
   - `influence_score`: the social reputation of each user based on follower and following counts that is computed by $\log((1 + \text{followers\_count})^2 + \log(\text{statuses\_count}) - \log(\text{friends\_count}))$;
   - `listed_count`: the number of public lists of which a user is a member;
   - `statuses_count` the number of tweets and retweets shared by a user;
   - `has_url`: a boolean feature showing if a user has a url or not;
   - `sociability`: the ratio of the number of friends_count to followers_count: $\log(1 + \frac{1+\text{friends\_count}}{1+\text{followers\_count}})$;
   - `favorability`: ratio of the number of favorites to the total number of tweets; it informs higher engagement in contrast to just posting tweets. $\log(1 + \frac{(1+\text{favourites\_count})}{(1+\text{statuses\_count})})$;
   - `survivability`: the potential active existence on the platform over time, and it is measured as the difference between current timestamp and the timestamp at which a tweet is created;

- `activeness`: the number of tweet statuses to the period of time since account creation; it determines the likelihood of a user to be active over a period of time on average: $\log(1 + \frac{(1+\text{statuses\_count})}{(1+\text{survivability}))}$;
- `favourites_count`: the number of tweets a user has favored over time as a measure of user engagement level.

3. ***Network-based Features***:
   - `betweenness`: the normalized sum of the fraction of all-pairs' shortest paths that pass through a node/user. Betweenness values are normalized by $b = b\frac{(n-1)}{(n-2)}$ where n is the number of nodes in graph G.
   - `degree_centrality`: the fraction of nodes/users to which a particular user is connected.
   - `load_centrality`: the normalized fraction of all shortest paths that pass through a node.

Feature Ranking Feature Selection is the process of identifying relevant features from a feature set that contribute most to the prediction variable and removing the irrelevant ones, in order to improve performance of predictive model. Features can be ranked according to different metrics, e.g. F-test statistic, Mutual Information (MI) measure, and p-value.

- F`-value`: ANOVA F-test statistic captures linear dependency of two random variables and computes F-value for each feature. This test measures the ratio of between-groups to within-groups variances; we note that the groups here are user behavior categories. When F-values are near 1, the null hypothesis is true (establishing independence).
- p`-value`: This allows determining whether the null hypothesis can be rejected with 95% confidence level (corresponding to p-values of 0.05). The smaller the p-value, the stronger the evidence to reject the null hypothesis.
- `MI measure`: This measure captures mutual dependency between random variables. MI quantifies the amount of information obtained about one random variable through observing the other random variable, with zero value showing two random variables are independent, whereas higher values meaning higher dependency. In Section 4 we provide the F-value, p-value, and MI measure for each of the features.

F-value along with the p-value enables us in deciding whether results are significant enough to reject the null hypothesis. We investigated both univariate feature selection and recursive feature elimination (RFE) and in both methods, the top 1/3 of the features (5 of the 16 initial features) are very similar. In particular, RFE selects smaller sets of features recursively with the least important features pruned at each iteration. We employed an SVM classifier with linear kernel in the RFE estimator for this purpose.

User Behaviors Estimation: In previous sections, we described how we group similar users based on their associated behaviors into three user behavioral clusters (classes). In this section, we focus on building a behavioral model that predicts a user behavior class by incorporating extracted features into supervised classification models. We consider ten different classifiers such as Nearest Neighbors (k-NN), SVM with RBF kernel, Random Forest, a multilayer perceptron classifier (MLP), AdaBoost, XGBoost, Naive Bayes, Decision Tree, and Quadratic Discriminant Analysis (QDA). Each classifier is used with recommended default parameter settings and not optimized for performance. Specifically, the number of neighbors in k-NN is 3, for SVM with RBF kernel γ is chosen automatically, the maximum depth of decision tree is set to 5 both in the Decision Tree and the Random Forest classifier (`max_depth = 5`). In the latter, the number of estimators is set to 10 (`n_estimators = 10`), and the maximum number of features is set to 1 (`max_features = 1`). In the MLP classifier, settings include L2 penalty (regularization term) parameter $\alpha = 1$. In XGBoost classifier, the parameters are set as follows: `n_estimators = 100, learning_rate = 1.0, max_depth = 1, random_state = 0`.

Each classifier is trained on a balanced version of the training dataset to effectively compare performance while addressing the class imbalance. Two options are considered for this: Balanced bagging versus SMOTE. We note that the balanced bagging effectively provides an ensemble method with each of the ten classifiers acting as the base classifier. While balanced bagging undersamples, SMOTE (Synthetic Minority Over-sampling Technique) oversamples [5].

The classification performance is evaluated using accuracy and F1 score. Accuracy evaluates the number of correct predictions over the total number of predictions, whereas the F1 score = 2·(precision·recall)/(precision+recall), where $precision = \frac{TP}{TP+FP}$ and recall/sensitivity = $\frac{TP}{TP+FN}$; TP, FP, and FN refer to the number of true positives, false positives, and false negatives, respectively. We note that in this multi-class setting, the F1 score is a micro-average; that is, contributions of all classes are aggregated to compute an average metric.

4 Experiments and Results

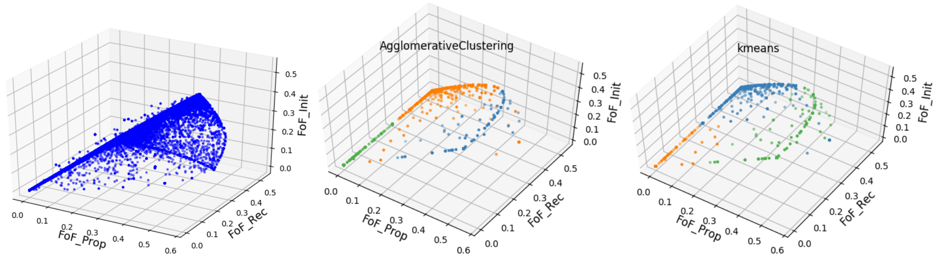
4.1 Experimental Setup

We describe three sets of experiments. First, we evaluate the performance of various clustering algorithms that allow us to learn user behavior groups (categories) in an unsupervised manner. Comparative analysis shows the most effective approach that we employ for getting the associated behavior category user labels for training. Second, we conduct a detailed selective analysis on user feature sets to choose the most relevant set for the identified behavior classes. Third, we show multi-class classification models that learn the relationship between features and the behavior classes. A principled comparison of these models along several performance metrics is presented.

Dataset: Our dataset contains records of retweets between May 16th 2016 and Dec 31st 2017 provided by [8]. Each record is a retweet of a tweet that contains at least one link to an article, which can be either a claim or a fact-checking source. The dataset consists of $20,987,210$ retweets, with $19,917,712$ (95%) linking to claim articles (fake) and $1,069,498$ (5%) to fact-checking articles (fact). We randomly sample $5,000$ users participating in retweet cascades over which we identify behavior categories via clustering to consider as the labels (inferred ground truth) for users.

4.2 Visualization of User Behavior Categories

The three key behavior dimensions introduced in Section 3 are used to visualize the user behavior space shown in Figure 1(a). We observe groups of users who receive but block major fake over fact cascades, users that propagate more fakes than facts cascades, and users that act somewhere in between. These observations can be quantified via automatic groupings of users, as done via clustering algorithms. Figure 1(b)-(c) shows the three user behavior clusters detected by Agglomerative and kmeans clustering algorithms with the best performances and table 1 shows their clustering performance metrics as described in Section 3).



(a) (b) (c)

Fig. 1. (a) Visualization of 3D user behavior space. Each dimension represents the log of ratio of user reactions (initiation, propagation, or reception) to fake over fact cascade exposures. The results of the Agglomerative clustering and kmeans clustering are shown in (b) and (c), respectively. Different colors show the emergent user clusters.

We call the three user categories detected in the semantic space of user behaviors as `malicious`, `good`, and `vulnerable/naive users`, based on their locality: `good user` has insignificant participation in the spread of fakes, `malicious user` who participate significantly in spreading or initiating fake over facts; and vulnerable user who mostly have lower rate of fakes to facts propagation and higher fake to fact reception. These are users who have volatile reactions in terms of behavior of reception, initiation, and propagation of fakes through the network.

Table 1. Comparing the performance of clustering algorithms using evaluation metrics.

Clustering algorithm	**Silhouette**	**Calinski-Harabaz**	**Davies-Bouldin**
kmeans	0.88	22913.55	0.30
Agglomerative Clustering	0.87	20951.39	0.32

4.3 Feature Ranking and Selection

In this section, we evaluate the significance of each feature using F-value, p-value, and MI measure calculated as described in Section 3 for the user behavior categories obtained via clustering. Table 2 shows results computed over user behavior categories obtained via Agglomerative clustering. Features with MI measure above 0.5 (important ones) are highlighted in bold. (Results computed over categories obtained via kmeans are similar and omitted due to space limit).

Table 2. The F-value, p-value, and MI measure for features computed over user behavior categories obtained via Agglomerative clustering.

Feature	**F-value**	**MI**	**p-value**
influence score	0.13	**0.68**	0
betweenness	0.00	0.04	0.45
deg centrality	0.07	0.06	0
clustering coefficient	0.00	0.00	0.1
load centrality	0.00	0.00	0.45
followers_count	0.01	**0.55**	0
friends_count	0.02	0.38	0
listed_count	0.00	0.39	0.02
statuses_count	0.01	**0.60**	0
has_url	0.03	0.16	0
tweet character length	0.16	**1.00**	0
sentiment	0.01	0.44	0
sociability	0.08	**0.65**	0
favorability	0.04	**0.66**	0
survivability	1.00	**0.67**	0
activeness	0.01	**0.62**	0

Feature selection by recursive feature elimination (RFE algorithm) ranks the following top five features as the most significant within each user behavior category obtained by Agglomerative clustering: *followers_count, friends_count, statuses_count, survivability* and *tweet_character_length*; And kmeans clustering: *followers_count, friends_count, listed_count, statuses_count, survivability.* We note great agreement between these two sets and the features with MI measure higher than 0.5 shown in Table 2. Further, besides belonging to all feature types - content, user profile, and interaction network, visual comparisons of feature distributions within each group can be found here: `https://drive.google.com/drive/folders/1B5-xVFMK9y6yW2OGmpCY8BvcltFGMGMV`.

4.4 Comparative Analysis of Predictive User Behavior Models

We examined the performance of various multi-class prediction models learned using aforementioned top features for classification of three user behavior categories obtained via both kmeans and agglomerative clusterings. A representative dataset of $5,000$ sampled users over which clustering is performed to obtain labels is subjected to both 10-fold cross validation (CV) and a $60-40$ split strategy for train-test sets. Table 3 shows the performance of 10 different classifiers on the test set (and average over 10 folds); the classifiers are trained over a balanced version of the training set, where we address class imbalance using balanced bagging and SMOTE sampling methods. As Table 3 shows, the majority of the classifiers saturate in performance around 0.80 in both accuracy and F1 score, highlighting the scope of further improvement in predicting user behavior.

Overall, the good performance belongs to MLP and SVM with balanced bagging and also k-NN classifier with balanced training set using SMOTE (n-nearest = 3). We also had additional experiments (results omitted due to space limit) for imbalanced settings and found both accuracy and F1 score reaching up to 0.80 for 10-fold CV. These results provide a preliminary evidence that it is possible to build user behavior predictive models that can be further improved with larger datasets or more features, by exploiting information available in near real-time for mitigation strategies.

Table 3. Comparison of performance of classification approaches in terms of Accuracy and F1 score for kmeans clustering labels for both 10-fold CV and split-strategy (in brackets). Notations: *BB-Acc(10-CV (Split)): Balanced Bagging Accuracy, BB-F1: Balanced Bagging F1, SMOTE-Acc: Synthetic Minority Oversampling Technique (SMOTE) Accuracy, SMOTE-F1: SMOTE F1-score.*

Classifier	**BB-Acc**	**BB-F1**	**SMOTE Acc**	**SMOTE F1**
3-NN	0.52 (0.52)	0.75 (0.51)	**0.76** (0.74)	0.76 (0.74)
RBF SVM	**0.80** (0.52)	0.81 (0.79)	0.53 (0.54)	0.53 (0.54)
Decision Tree	0.50 (0.55)	0.79 (0.52)	0.59 (0.58)	0.59 (0.58)
Random Forest	0.60 (0.54)	0.81 (0.61)	0.57 (0.56)	0.58 (0.56)
MLP	**0.80** (0.52)	0.81 (0.79)	0.53 (0.51)	0.53 (0.51)
AdaBoost	0.50 (0.36)	0.79 (0.49)	0.58 (0.57)	0.58 (0.57)
XGBoost	0.53 (0.54)	0.80 (0.54)	0.60 (0.59)	0.60 (0.59)
Gaussian NB	0.77 (0.38)	0.79 (0.76)	0.53 (0.54)	0.53 (0.54)
QDA	0.53 (0.42)	0.61 (0.62)	0.45 (0.35)	0.45 (0.35)

5 Conclusion and Future Work

In this paper we have presented a novel, data-driven approach for user behavior analysis on social web for assisting fake content mitigation strategies. The identification of key behavior dimensions allows leveraging unsupervised learning to organize users along behavior categories. We identified diverse features from user information that is available in near realtime to validate predictability of user behavior categories. Supervised learning models show that user behavior categories can be predicted from such features. Although, we acknowledge the limitation of the experiments, in particular, the approach to data sampling and extracted features. Given this preliminary foundation work for user modeling to serve user intervention strategies, we will address these limitations in our future work. Furthermore, behavioral psychologists can contribute detailed models of user behavior that can inform or refine the presented data-driven modeling approach. This research provides the groundwork for advanced user modeling toward mitigation-focused social cyber-security research.

References

1. Al-khateeb, S., Hussain, M.N., Agarwal, N.: Social cyber forensics approach to study twitters and blogs influence on propaganda campaigns. In: SBP-BRiMS. pp. 108–113. Springer (2017)
2. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J Economic Perspectives **3**(2), 211–236 (2017)
3. Bock, K., Shannon, S., Movahedi, Y., Cukier, M.: Application of routine activity theory to cyber intrusion location and time. In: 2017 13th European Dependable Computing Conference (EDCC). pp. 139–146 (2017)
4. Carley, K.M., Cervone, G., Agarwal, N., Liu, H.: Social cyber-security. In: SBP-BRiMS. pp. 389–394. Springer (2018)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J Artificial Intel Res **16**, 321–357 (2002)
6. Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., Zha, H.: Fake news mitigation via point process based intervention. In: 34th Int'l Conf. on Machine Learning,. pp. 1097–1106. JMLR.org (2017)
7. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM **59**(7), 96–104 (2016)
8. Hui, P.M., Shao, C., Flammini, A., Menczer, F., Ciampaglia, G.L.: The hoaxy misinformation and fact-checking diffusion network. In: ICWSM (2018)
9. Jiang, M., Cui, P., Faloutsos, C.: Suspicious behavior detection: Current trends and future directions. IEEE Intelligent Systems **31**(1), 31–39 (2016)
10. Jin, L., Chen, Y., Wang, T., Hui, P., Vasilakos, A.V.: Understanding user behavior in online social networks: A survey. IEEE Communications Magazine **51**(9), 144–150 (2013)
11. Purohit, H., Pandey, R.: Intent mining for the good, bad, and ugly use of social web: Concepts, methods, and challenges. In: Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining, pp. 3–18. Springer (2019)
12. Sameki, M., Zhang, T., Ding, L., Betke, M., Gurari, D.: Crowd-o-meter: Predicting if a person is vulnerable to believe political claims. In: HCOMP. pp. 157–166 (2017)
13. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Joo, E.M., Weiping, D., Chin-Teng, L.: A review of clustering techniques and developments. Neurocomputing **267**, 664–681 (2017)
14. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter **19**(1), 22–36 (2017)
15. Starbird, K.: Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In: ICWSM. pp. 230–239 (2017)
16. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: Detection, estimation, and characterization. In: ICWSM. pp. 280–289 (2017)
17. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health informationa survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(5), e1209 (2017)