When social bots attack: Modeling susceptibility of users in online social networks

Claudia Wagner
Institute for Information and Communication Technologies
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Silvia Mitter
Knowledge Management Institute
Graz University of Technology
Graz, Austria
smitter@student.tugraz.at

Christian Körner
Knowledge Management Institute
Graz University of Technology
Graz, Austria
christian.koerner@tugraz.at

Markus Strohmaier
Knowledge Management Institute and Know-Center
Graz University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Social bots are automatic or semi-automatic computer programs that mimic humans and/or human behavior in online social networks. Social bots can attack users (targets) in online social networks to pursue a variety of latent goals, such as to spread information or to influence targets. Without a deep understanding of the nature of such attacks or the susceptibility of users, the potential of social media as an instrument for facilitating discourse or democratic processes is in jeopardy. In this paper, we study data from the Social Bot Challenge 2011 - an experiment conducted by the WebEcologyProject during 2011 - in which three teams implemented a number of social bots that aimed to influence user behavior on Twitter. Using this data, we aim to develop models to (i) identify susceptible users among a set of targets and (ii) predict users' level of susceptibility. We explore the predictiveness of three different groups of features (network, behavioral and linguistic features) for these tasks. Our results suggest that susceptible users tend to use Twitter for a conversational purpose and tend to be more open and social since they communicate with many different users, use more social words and show more affection than non-susceptible users.

Keywords

social bots, infection, user models

1. INTRODUCTION

Online social networks (OSN) like Twitter or Facebook are powerful instruments since they allow reaching millions of users online. However, in the wrong hands they can also be used to spread misinformation and propaganda, as one could for example see during the US political elections [9]. Recently a new breed of computer programs so-called *social media robots* (short *social bots* or *bots*) emerged in OSN. Social bots are automatic or semi-automatic computer programs that mimic humans and/or human behavior in OSN. Social bots can be directed to attack users (targets) to pursue a variety of latent goals, such as to spread information or to influence users [7]. Recent research [1] highlights the danger of social bots and shows that Facebook can be infiltrated by social bots sending friend requests to users. The average reported acceptance rate of such friend requests was 59.1% which also depended on how many mutual friends the social bots had with the infiltrated users, and could be up to 80%. This study clearly demonstrates that modern security defenses, such as the Facebook Immune System, are not prepared for detecting or stopping a large-scale infiltration caused by social bots.

We believe that modern social media security defenses need to advance in order to be able to detect social bot attacks. While identifying social bots is crucial, identifying users who are susceptible to such attacks - and implementing means to protect against them - is important in order to protect the effectiveness and utility of social media. In this paper, we define a *target* to represent a user who has been singled out by a social bot attack, and a *susceptible user* as a user who has been infected by a social bot (i.e. the user has in some way cooperated with the agenda of a social bot). This work sets out to identify factors which help detecting users who are susceptible to social bot attacks. To gain insights into these factors, we use data from the Social Bot Challenge 2011 and introduce three different groups of features: network features, behavioral features and linguistic features. In total, we use 97 different features to first *predict infections* by training various classifiers and second aim to *predict users' level of susceptibility* by using regression models.

Thus, unlike previous research, our work *does not focus on detecting social bots in OSN, but on detecting users who are susceptible to their attacks.* To the best of our knowledge,

Published as part of the #MSM2012 Workshop proceedings, available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

this represents a novel task that has not been proposed or tackled previously. Our work is relevant for researchers interested in social engineering, trust and reputation in the context of OSN.

2. RELATED WORK

Social bots represent a rather new phenomenon that has received only little attention so far. For example, Chu et al. [3] use machine learning to identify three types of Twitter user accounts: users, bots and cyborgs (users assisted by bots). They show that features such as entropy of posts over time, external URL ratio and Twitter devices (usage of external Twitter applications) give good indications for differentiating between distinct types of user accounts [1]. Work by [6] describes how honeypots can be used to identify spam profiles in OSN. They present a long term study where 60 honeypots were able to harvest about 36.000 candidate content polluters over a period of 7 months. Based on the collected data they trained a classification model using features based on User Demographics, User Friendship Networks, User Content and User History. Their results and show that features which were most useful for differentiating between content polluters and legitimate users were User Friendship Network based features, like the standard deviation of followees and followers, the change rate of the number of followees and the number of followees. In the context of the goals of this paper, related work on spam detection in OSN is as well relevant. For example, Wang et al. [14] propose a general purpose framework for spam detection across multiple social networks. Unlike previous research, our work does not focus on detecting spammers or social bots in OSN, but on detecting users who are susceptible to their attacks.

Research about users' online behavior in general represents another field that is closely related to our research on user susceptibility. Predicting users' interaction behavior (i.e., who replies to whom, who friends whom) in online media has been previously studied in the context of email communications [12] and more recently in the context of social media applications. For example, Cheng et al. [2] consider the problem of reciprocity prediction and study this problem in a communication network extracted from Twitter. The authors aim to predict whether a user A will reply to a message of user B by exploring various features which characterize user pairs and show that features that approximate the relative status of two nodes are good indicators of reciprocity. Work described in [10] considers the task of predicting discussions on Twitter, and found that certain features were associated with increased discussion activity - i.e., the greater the broadcast spectrum of the user, characterized by in-degree and list-degree levels, the greater the discussion activity. The work of Hopcroft et al. [4] explores follow-back-behavior of Twitter users and find strong evidence for the existence of the structural balance among reciprocal relationships. In addition, their findings suggest that different types of users reveal interesting differences in their follow-back behavior: the likelihood of two elite users creating a reciprocal relationships is nearly 8 times higher than the likelihood of two ordinary users. Our work differs from the related work discussed above by focusing on modeling and predicting the behavior of users who are currently attacked by social bots.

3. THE SOCIAL BOT CHALLENGE

The Social Bot Challenge was a competition organized by Tim Hwang (and the WebEcologyProject). The competition took place between January and February 2011. The aim was to have a set of competing teams developing social bots that persuade targets to interact with them - i.e., reply to them, mention them in their tweets, retweet them or follow them. The group of targets consisted of 500 unsuspecting Twitter users which were selected semi-randomly: all users had an interest in or tweeted about *cats*. The majority of targets exhibited a high activity level, that means they tweeted more than once a day. We define a *susceptible user* as a target that interacted (i.e., replied, mentioned, retweeted or followed) at least once with a social bot.

3.1 Rules

Each team was allowed to create one lead bot (the only bot allowed to score points) and an arbitrary number of support bots. The participating teams got points for every successful interaction between their lead bot and any target. One point was awarded for any target who started following a lead bot and three points were awarded for any target who replied to, mentioned or retweeted a lead bot.

The following rules were announced for the game:

- No humans are allowed during the game. That means bots need to act in a completely automated way.
- Teams were not allowed to report other teams as spam or bots to Twitter, but other countermeasures and strategies to harm the opponents are allowed.
- The existence of the game needs to remain a secret. That means bots are not allowed to inform others about the game.
- The code needs to be published as open source under the MIT license.
- Teams are allowed to collaborate. That means they are allowed to talk to each other and exchange their code.

There was a period of 14 days during which teams were allowed to develop their social bots. Afterwards the game started on the Jan 23rd 2011 (day 1) and ended Feb 5th 2011 (day 14). During this period, bots were autonomously active for the first 7 days. At the 30th of January (day 8) the teams were allowed to update their codebase and change strategies. After this optional update, the bots continued to be autonomously active for the remaining time of the challenge

3.2 Participants and Challenge Outcome

The following three teams competed in the challenge.

- **Team A - @sarahbalham** The lead bot *sarahbalham* claims to be a young woman who grew up on the countryside and just moved to the city. This team didn't construct a bot-network,but only used one lead bot. This lead bot created 143 tweets, which is rather low

in comparison to the other teams, and used only a few @replies and hashtags. Despite low activity level this team could reach the highest number of mutual connections, which is 119 connections. Overall the team only collected 170 points, since only 17 interactions with targets were counted.

- **Team B - @ninjzz** The woman impersonated by this bot - *ninjzz* - doesn't provide much personal information, only that she is a bit shy and looking for friends on Twitter. Ninjzz was supported by 10 other bots, which also created some tweets. This bot was rather defensive in the first round of the challenge, but changed the strategy on day 8 and acted in a much more aggressive way in the second part of the challenge. Overall this team created 99 mutual connections and 28 interactions, and therefore collected 183 points.

- **Team C - @JamesMTitus** The bot *JamesMTitus* claims to be a 24 old guy from New Zealand, who is new on Twitter, and a real cat enthusiast. Team C with their bot *JamesMTitus* won the game by collecting 701 points, with 107 mutual connections and 198 interactions. This team had five support bots, who only created social connections but did not tweet at all. The team picked a very aggressive strategy, tweeted a lot and also made extensively use of @replies, retweets and hashtags.

4. DATASET

The authors of this paper were not involved in nor did they participate in the design, setup or execution of this challenge. The dataset used for this analysis was provided by the WebEcologyProject after the challenge took place. Table 1 provides a basic description of this dataset. Figure 1 shows infections over time - i.e., it depicts on which day of the challenge targets interacted with social bots for the first time. One can see from this figure that at the beginning of the challenge - on day 2 - already 87 users became infected. One possible explanation for this might be the usage of auto-following features which some of the targets might have used. One can see from Figure 2 that for the users who became infected at an early stage of the challenge, we do not have many tweets in our dataset. This is a limitation of the dataset we use, which includes only tweets authored between the 23th of January and the 5th of February and social relations which where existent at the this point in time or created during this time period. Since most of our features require a certain amount of tweets a user authored in order to contain meaningful information about the user, we decided to remove all users who became susceptible before day 7. While this means we loose 133 susceptible users as samples for our experiments, we believe (i) that the remaining 76 susceptible users and 298 non-susceptible users are sufficient to train and test our classifiers and regression models and (ii) that eliminating those users that might have used an auto-follow feature is a good since they are less interesting to study from a susceptibility viewpoint.

5. FEATURE ENGINEERING

We adopt a two-stage approach to modeling targets' susceptibility to social bot attacks: (i) We aim to identify infected

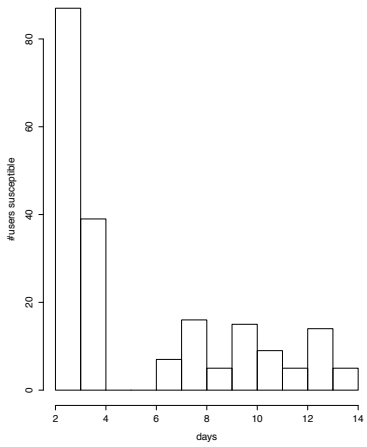

Figure 1: This figure shows for each day of the challenge the number of users who were infected - i.e., they interacted with a social bot for the first time.

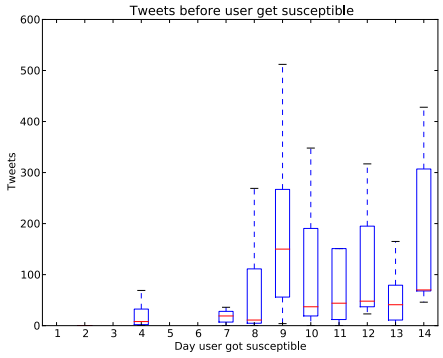

Figure 2: This figure shows when users were infected and how many tweets they have published before - i.e. between the start of the challenge and the day they were infected.

Table 1: Description of the Social Bot Challenge dataset

Susceptible	202
Non-Susceptible	298
Mean Tweets per User	146.49
Mean Nr of Follower/Followees per User	8.5

users via a binary classification task, and (ii) we aim to predict the level of susceptibility per infected user. To this end we explore three distinct feature sets that can be leveraged to describe the susceptibility of users: *linguistic features*, *behavioral features* and *network features*.

For all targets, we computed the features by taking all tweets they authored (up to the point in time where they become infected) and a snapshot of the targets' follow network which was as recorded at the 26th of January (day 4). Since we only study susceptible users who became infected on day 7 or later, this follow network snapshot does not contain any

future information (such as tweets or social relations which were created after a user became infected) which could bias our prediction results. Based on this aggregation of tweets, we constructed the interaction and retweet network of each user by analyzing their reply and retweet interactions.

5.1 Linguistic Features

Previous research has established that physical and psychological functioning are associated with the content of writing [8]. In order to analyze such content in an objective and quantifiable manner, Pennebaker and colleagues developed a computer based text-analysis program, known as the Linguistic Inquiry and Word Count (short LIWC) [11]. LIWC uses a word count strategy searching for over 2300 words or word stems within any given text. The search words have previously been categorized by independent judges into over 70 linguistic dimensions. These dimensions include standard language categories (e.g., articles, prepositions, pronouns including first person singular, first person plural, etc.), psychological processes (e.g., positive and negative emotion categories, cognitive processes such as use of causation words, self-discrepancies), relativity-related words (e.g., time, verb tense, motion, space), and traditional content dimensions (e.g., sex, death, home, occupation).

In this work we use those 70 linguistic dimensions[1] as linguistic features and compute them based on the aggregation of tweets authored by each target. Due to space limits we do not describe all 70 features in detail, but explain those which seem to be relevant for modeling the susceptibility of users in the result section.

5.2 Network Features

To study the predictiveness of network theoretic features we constructed the following three directed networks from the data. In each of the networks nodes correspond to targets, while edges are constructed differently.

- *User-Follower* - A network representing the target - follower structure in Twitter. There exists an directed edge from user A to user B if the user A is followed by B.
- *Retweet* - A network representing the retweet behavior of targets. In this network there exists an edge from A to B if user A retweeted a message from B.
- *Interaction* - The third network captures the general interaction behavior of targets. There exists an edge from user A to user B if user A either mentioned, replied, or retweeted user B.

For each point in time, we constructed a retweet and interaction network by analyzing all tweets users published before that timestamp. The follower-network is based on a snapshot which was as recorded at the 26th of January (day 4).

[1] `http://www.liwc.net/descriptiontable1.php`

5.2.1 Hub and Authority Score

Using Kleinberg's *HITS* algorithm [5], we calculated the authority as well as the hub score for all targets in our networks. A high authority-score indicates that a node (i.e., a user) has many incoming edges from nodes with a high hub score, while a high hub-score indicates that a node has many outgoing edges to nodes with high authority scores. For example, in the retweet network a high authority score indicates that a user is retweeted by many other users who retweeted many users, while a high hub score indicates that the user retweets many others who are as well retweeted by many others.

5.2.2 In- and Out-Degree

A high in-degree indicates that a node (i.e., a user) has many incoming edges, while a high out-degree indicates that a node has many outgoing edges. For example, in the interaction network a high in-degree means that a user is retweeted, replied, mentioned and/or followed by many other users, while a high out-degree indicates that the user retweets, replies, follows and/or mentions many other users.

5.2.3 Clustering Coefficient

The clustering coefficient is defined as the number of actual links between the neighbors of a node divided by the number of possible links between the neighbors of that node. A high clustering coefficient of a node indicates that a node has a central position in the network. For example, in the follow network a high clustering coefficient indicates that the users a user follows or is followed by, are also well connected via follow relations.

5.3 Behavioral Features

In our own previous work [13], we introduced a number of behavioral or structural measures that can be used to characterize user streams and reveal structural differences between them. In the following, we describe some of those measures and elaborate how we use them to gauge the susceptibility of targets.

5.3.1 Conversational Variety

The conversational variety per message $CVpm$ represents the mean number of different users mentioned in one message of a stream and is defined as follows:

$$CVpm = \frac{|U_m|}{|M|} \tag{1}$$

To measure the number of users being mentioned in a stream (e.g., via @replies or slashtags), we introduce $|U_m|$ for $u_m \in U_m$. A high conversational variety indicates that a user talks with many different users.

5.3.2 Conversational Balance

To quantify the conversational balance of a stream, we define an entropy-based measures, which indicates how evenly balanced the communication efforts of a user is distributed across his communication partners. We define the conversational balance of a stream as follows:

$$CB = -\sum_{u \in U_m} P(m|u) * log(P(m|u)) \tag{2}$$

A high conversational balance indicates that the user talks equally much with a large set of users, i.e. the distribution of conversational messages per user is even. Therefore a high score indicates that it is hard to predict with whom a user will talk next.

5.3.3 Conversational Coverage

From the number of conversational messages $|M_c|$ - i.e., messages which contain an @reply - and the total number of messages of a stream $|M|$, we can compute the conversational coverage of a user stream, which is defined as follows:

$$CC = \frac{|M_c|}{|M|} \tag{3}$$

A high conversational coverage indicates that a user is using Twitter mainly for a conversational purpose.

5.3.4 Lexical Variety

To measure the vocabulary size of a stream, we introduce $|R_k|$, which captures the number of unique keywords $r_k \in R_k$ in a stream. For normalization purposes, we include the stream size ($|M|$). The lexical variety per message $LVpm$ represents the mean vocabulary size per message and is defined as follows:

$$LVpm = \frac{|R_k|}{|M|} \tag{4}$$

5.3.5 Lexical Balance

The lexical balance LB of a stream can be defined, in the same way as the conversational balance, via an entropy-based measure which quantifies how predictable a keyword is on a certain stream.

5.3.6 Topical Variety

To compute the topical variety of a stream, we can use arbitrary surrogate measures for topics, such as the result of automatic topic detection or manual labeling methods. In the case of Twitter we use the number of unique hashtags $r_h \in R_h$ as surrogate measure for topics. The topical variety per message $TVpm$ represents the mean number of topics per message and is defined as follows:

$$TVpm = \frac{|R_h|}{|M|} \tag{5}$$

5.3.7 Topical Balance

The topical balance TB can, in the same way as the conversational balance, be defined as an entropy-based measure which quantifies how predictable a hashtag is on a certain stream. A high topical balance indicates that a user talks about many different topics to similar extents. That means the user has no topical focus and it is difficult to predict about which topic he/she will talk next.

5.3.8 Informational Variety

In the case of Twitter we define informational messages to contain one or more links. To measure the informational variety of a stream, we can compute the number of unique links in messages of a stream $|R_l|$ for $r_l \in R_l$. The informational variety per message $IVpm$ is defined as follows:

$$IVpm = \frac{|R_l|}{|M|} \tag{6}$$

5.3.9 Informational Balance

The informational balance IB can, in the same way as the conversational balance, be defined as an entropy-based measures which quantifies how predictable a link is on a certain stream. A high informational balance indicates that a user posts many different links as part of her tweeting behavior.

5.3.10 Informational Coverage

From the number of informational messages $|M_i|$ and the total number of messages of a stream $|M|$ we can compute the informational coverage of a stream which is defined as follows:

$$IC = \frac{|M_i|}{|M|} \tag{7}$$

A high informational coverage indicates that a user is using Twitter mainly to spread links.

5.3.11 Temporal Variety

The temporal variety per message $TPVpm$ of a stream is defined via the number of unique timestamps of messages $|TP|$ (where timestamps are defined to be unique on an hourly basis), and the number of messages $|M|$ in a stream. The temporal variety is defined as follows:

$$TPVpm = \frac{|TP|}{|M|} \tag{8}$$

5.3.12 Temporal Balance

The temporal balance TPB can, in the same way as the social balance, be defined as an entropy-based measure which quantifies how balanced messages are distributed across these message-publication-timestamps. A high temporal balance indicates that a user is tweeting regularly.

5.3.13 Question Coverage

From the number of questions $|Q|$ and the total number of messages of a stream $|M|$ per stream we can compute the question coverage of a stream which is defined as follows:

$$QRpm = \frac{|Q|}{|M|} \tag{9}$$

A high question coverage indicates that a user is using Twitter mainly for gathering information and asking questions.

6. EXPERIMENTS

In the following, we attempt to develop models that (i) identify susceptible users (whether a user becomes infected or not) and (ii) predict their level of susceptibility (the extent to which a user interacts with a social bot). We begin by explaining our experimental setup before discussing our findings.

6.1 Experimental Setup

For our experiments, we considered all targets of the Social Bot Challenge, and divided them into those who were not infected (*non-susceptible users*) and those who were infected, i.e. started interacting with a bot within day 7 or later (*susceptible users*). For each of those targets we constructed the features as described in section 5 and normalized them.

Identifying the most susceptible users in a given community is often hindered by including users that are not susceptible

at all. We alleviate this problem by first aiming to model the differences between susceptible and non-susceptible users in a binary classification task. Once susceptible users have been identified, we can then attempt to predict the level of susceptibility for each infected user. Therefore we performed the following two experiments.

1. *Predicting Infections* The first experiment sought to identify the factors that are associated with infections. To this end, we performed a binary classification task using 6 different classifier, partial least square regression (pls), generalized boosted regression (gbm), k-nearest neighbor (knn), elastic-net regularized generalized linear models (glmnet), random forest (rf) and regression trees (rpart). We divided our dataset into a balanced training and test set - i.e. in each training and test split we had the same number of susceptible and non-susceptible users. We performed a 10-cross-fold validation and selected the best classifier to further explore the most predictive features, and plotted ROC curves for each feature. The ROC curve is a method to visualize the prediction accuracy of ranking functions showing the number of true positives in the results plotted against the number of results returned. We use the area under the ROC curve (AUC) as the measure of feature importance.

2. *Predicting Levels of Susceptibility* After identifying susceptible users, it is interesting to rank them according to their probability of being susceptible for a bot attack, because one usually wants to identify the most susceptible users, i.e. those who are most in need for security measures and protection. In this experiment we aim to predict the susceptibility level of infected users and identify key features which are correlated with users' susceptibility levels. We define the susceptibility level of an infected user as the number of times a user followed, mentioned, retweeted or replied to a bot.

   We divided our dataset (consisting of infected users only) into a 75/25% split, fit a regression model using the former split and applied it to the latter. We used regression trees to model the susceptibility level of infected users, since they can handle strongly nonlinear relationships with high order interactions and different variable types. The resulting model can be interpreted as a tree structure providing a compact and intuitive representation.

7. RESULTS & EVALUATION

7.1 Predicting Infections

As a first step, we would like to compare the performance of different classifiers for this task and compare them with a random baseline classifier. We used all features and trained six different classifiers: partial least square regression (pls), generalized boosted regression (gbm), k-nearest neighbor (knn), elastic-net regularized generalized linear models (glmnet), random forests (rf) and regression trees (rpart). One can see from table 2 that generalized boosted regression models (gbm) perform best, since they have the highest accuracy.

Table 2: Comparison of classifiers' performance

	Susceptible			Non-Susceptible			
Model	**F1**	**Rec**	**Prec**	**F1**	**Rec**	**Prec**	**Overall**
random	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*
gbm	**0.71**	**0.70**	**0.74**	**0.70**	**0.74**	**0.68**	**0.71**
glmnet	0.69	0.75	0.67	0.73	0.72	0.77	0.71
rpart	0.64	0.56	0.78	0.44	0.60	0.36	0.54
pls	0.67	0.69	0.68	0.68	0.71	0.70	0.68
knn	0.70	0.71	0.71	0.72	0.75	0.71	0.71
rf	0.68	0.72	0.66	0.70	0.70	0.74	0.69

To understand which features are most predictive, we explore the importance of different features by using our best performing model. Table 2 shows the importance ranking of features using the area under the ROC curve as a ranking criterion.

One can see from Table 3 that the most important features for differentiating susceptible and non-susceptible is the out-degree of a user node in the interaction network. Figure 3 shows that susceptible users tend to actively interact (i.e., retweet, mention, follow or reply to a user) with more users than non-susceptible users do on average. That means, susceptible users tend to have a larger social network and/or communication network. One possible explanation for that is that susceptible users tend to be more active and open and therefore easily create new relations with users. Our results also show that susceptible users also tend to have a high in-degree in the interaction network, which indicates that most of their interaction efforts are successful (i.e., they are followed back by users they follow and/or get replies/mentions/retweets from users they reply/mention/retweet).

Further, susceptible users tend to use more verbs (especially present tense verbs, but also past tense verbs and auxiliary verbs) and use more personal pronouns (especially first person singular but also third person singular in their tweets. This suggest that susceptible users tend to use Twitter to report about what they are currently doing.

Interestingly, our results also show that susceptible users have a higher conversational variety and coverage than non-susceptible users, which means that susceptible users tend to talk to many different users on Twitter and that most of their messages have a conversational purpose. This indicates that susceptible users tend to use Twitter mainly for a conversational purpose rather than an informational purpose. Further, susceptible users also have a higher conversational balance which indicates that they do not focus on few conversation partners (i.e., heavily communicate with a small circle of friends) but spend an equal amount of time in communicating with a large variety of users. Its suggests again that susceptible users are more open to communicate with others, also if they are not in their closed circle of friends.

Our results further suggest that susceptible users show more affection - i.e. they use more affection words (e.g., *happy*, *cry*), especially words which expose positive emotions (e.g., *love*, *nice*) - and use more social words (e.g., *mate*, *friend*) than non-susceptible users, which might explain why they are more open to interact with social bots. Susceptible users also tend to use more motion words (e.g., *go*, *car*), adverbs

(e.g., *really*, *very*), exclusive words (e.g., *but*, *without*) and negation words (e.g., *no*, *not*, *never*) in their tweets than non-susceptible users. It indicates again that susceptible users tend to use Twitter to talk about their activities and emotionally communicate.

To summarize, our results suggest that susceptible users tend to use Twitter mainly for a conversational purpose (high conversational coverage) and tend to be more open and social since they communicate with many different users (high out-degree and in-degree in the interaction network and high conversational balance and variety), use more social words and show more affection (especially positive emotions) than non-susceptible users.

Table 3: Importance ranking of the top features using the area under the ROC curve (AUC) is used as ranking criterion. The importance value is proportional to the most important feature which has an importance value of 100%.

Feature	Importance
out-degree (interaction network)	100.00
verb	98.01
conversational variety	96.93
conversational coverage	96.65
present	94.66
affect	90.15
personal pronoun	89.71
first person singular	89.27
conversational balance	87.28
motion	87.28
past	86.56
adverb	86.20
pronoun	84.41
negate	84.33
positive emotions	83.25
third person singular	82.38
social	82.02
exclusive	81.86
auxiliary verb	81.70
in-degree (interaction network)	81.66

7.2 Predicting Levels of Susceptibility

To model the susceptibility level of users, we use regression trees and aim to identify features which correlate with users' susceptibility levels. To gain insights into the factors which correlate with high or low susceptibility levels of a user, we inspect the regression tree model which was trained on 75% of our data. One can see from Figure 4 that users who use more negation words (e.g. not, never, no) tend to interact more often with bots, which means they have a higher susceptibility level. Further, users who tweet more regularly (i.e. have a high temporal balance) and users who use more words related with the topic death (e.g. bury, coffin, kill) tend to interact more often with bots than other susceptible users.

One can see from Figure 4 that the structure of the learned tree is very simple which means that our features only allow differentiating between rather lower and rather high susceptibility scores. For a more finer-grained susceptibility level prediction our approach is of limited utility. Also the rank correlation of users given their real susceptibility level and their predicted susceptibility level and the goodness of fit of the model is rather low. One potential reason for that is that our dataset is too small for fitting the model (we only have 76 samples and 97 features). Another potential reason is that our features do not correlate with susceptibility scores of users. We leave the task of elaborating on this problem to future work.

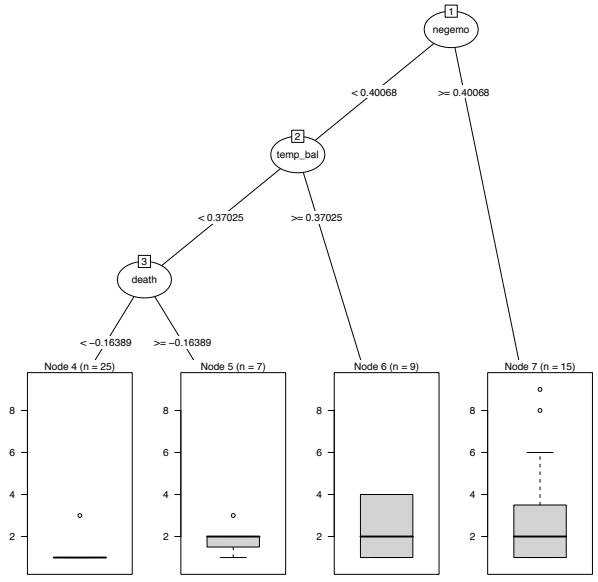

Figure 4: Regression tree model fitted to the susceptibility scores of our training split users. The tree-structure shows based on which features and thresholds the model selects branches and the box plots indicate the distribution of the susceptibility scores of users in each branch of the tree.

8. CONCLUSIONS AND OUTLOOK

In this work, we studied susceptibility of users who are under attack from social bots. To this end, we used data collected by the Social Bots Challenge 2011 organized by the WebEcologyProject. Our analysis aimed at (i) identifying susceptible users and (ii) predicting the level of susceptibility of infected users. We implemented and compared a number of classification approaches that demonstrated the capability of a classifier to outperform a random baseline.

Our analysis revealed that susceptible users tend to use Twitter mainly for a conversational purpose (high conversational coverage) and tend to be more open and social since they communicate with many different users (high out- and in-degree in the interaction network and high conversational balance), use more social words and show more affection (especially positive emotions) than non-susceptible users. Although finding that active users are also more susceptible for social bot attacks does not seem to be too surprising, it is an intriguing finding in itself as one would assume that users who are more active socially would develop some kind of social skills or capabilities to distinguish human users from social bots. This is obviously not the case and suggests that attacks of social bots can be effective even in cases where users have experience with social media and are highly active.

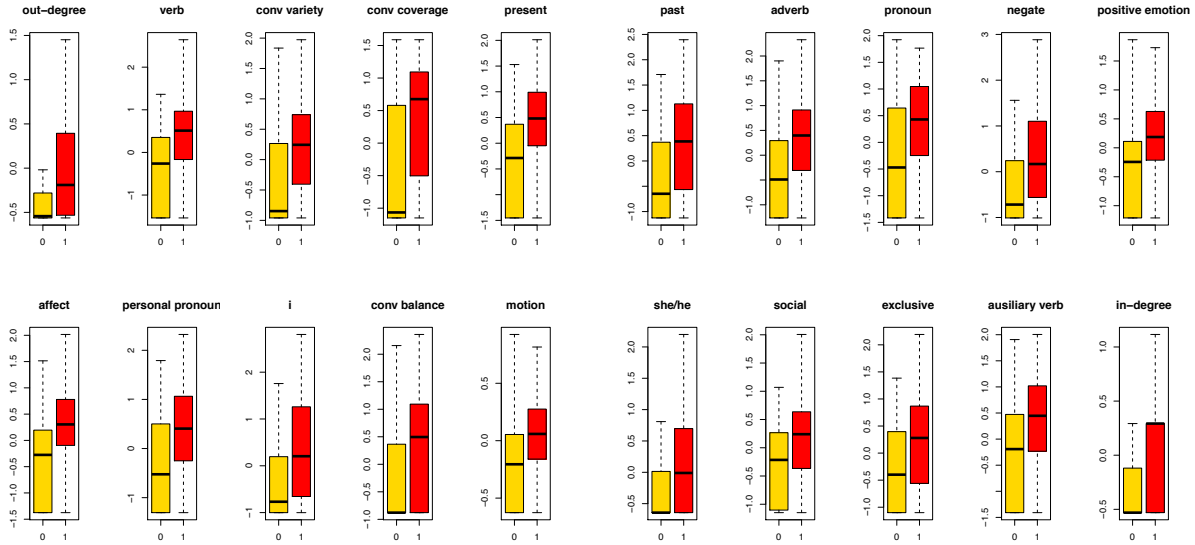

Figure 3: Box plots for the top 20 features according to the area under the ROC curve (AUC). Yellow boxes (class 0, left) represent non-susceptible users, red boxes (class 1, right) represent susceptible users. Differences between susceptible and non-susceptible users can be observed.

While our work presents promising results with regard to the identification of susceptible users, identifying the level of susceptibility is a harder task that warrants more research in the future. In general, the results reported in this work are limited to one specific domain (cats). In addition, all our features are corpus-based and therefore the size and structure of our dataset can have an influence on our results. In conclusion, our work represents a first important step towards modeling susceptibility of users in OSN. We hope that our work contributes to the development of tools that help protect users of OSN from social bot attacks, and that our exploratory work stimulates more research in this direction.

Acknowledgments

We want to thank members of the WebEcology project, especially Tim Hwang for sharing the dataset and Ian Pierce for technical support. Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science. This research is partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760.

9. REFERENCES

[1] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network. In *Proceedings of the 27th Annual Computer Security Applications Conference*, page 93. ACM Press, Dec 2011.

[2] J. Cheng, D. Romero, B. Meeder, and J. Kleinberg. Predicting reciprocity in social networks. In *he Third IEEE International Conference on Social Computing (SocialCom2011)*, 2011.

[3] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter. In *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC10*, page 21. ACM Press, Dec 2010.

[4] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: reciprocal relationship prediction. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1137–1146, New York, NY, USA, 2011. ACM.

[5] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In H. J. Karloff, editor, *SODA*, pages 668–677. ACM/SIAM, 1998.

[6] K. Lee, J. Caverlee, and S. Webb. *Uncovering Social Spammers : Social Honeypots + Machine Learning*, pages 435–442. Number i. ACM, 2010.

[7] D. Misener. Rise of the socialbots: They could be inuencing you online. web, March 2011.

[8] J. Pennebaker, M. Mehl, and K. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[9] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *CoRR*, abs/1011.3768, 2010.

[10] M. Rowe, S. Angeletou, and H. Alani. Predicting discussions on the social semantic web. In *Extended Semantic Web Conference*, Heraklion, Crete, 2011.

[11] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. 2010.

[12] J. R. Tyler and J. C. Tang. When can i expect an email response? a study of rhythms in email usage. In *Proceedings of the eighth conference on European Conference on Computer Supported Cooperative Work*, pages 239–258, Norwell, MA, USA, 2003. Kluwer Academic Publishers.

[13] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.

[14] D. Wang, D. Irani, and C. Pu. A social-spam detection framework. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference on*, pages 46–54. ACM Press, Sep 2011.