Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 52-19 (2019) 175–180

Analysis of the Impact of Poisoned Data within Twitter Classification Models

Kristopher R. Price * **Jaan Priisalu** ** **Sven Nomm** ***

* *Tallinn University of Technology, Tallinn, Estonia (Tel: +372 5878-1236; e-mail: pricekr1221@gmail.com).*
** *Tallinn University of Technology, Tallinn, Estonia (E-mail: jaan.priisalu@taltech.ee).*
*** *Tallinn University of Technology, Tallinn, Estonia (E-mail: sven.nomm@taltech.ee).*

Abstract: Many social-networks today face growing problems of group polarization, radicalization, and fake news. These issues are being exacerbated by the phenomenon of bots, which are becoming better at mimicking real people and are able to spread fake news faster within social-networks. Methods exist for detecting these social-media bots, but they may be vulnerable to manipulation. One way this might be done is through what is called a poisoning attack, where the data used to train a model is altered with the goal of reducing the models accuracy. The goal of this research is to study how poisoning attacks may be applied to models for detecting bots on Twitter. The results show that by introducing mislabeled data- points into a such a models training data, attackers can reduce its accuracy by up to twenty percent. The possibility of more effective poisoning techniques exists, and remains a topic for future research.

Keywords: automation, social-media, social-engineering, data-science, adversarial machine-learning.

1. INTRODUCTION

Today, many people communicate through social media. This medium has become a ubiquitous part of society that is used by many to read the news, contact friends, and plan events. Social networks such as Facebook tend to recommend users to connect with people already within their social circle rather than with complete strangers. The social structure resulting from this feature, where everyone knows one another and shares similar beliefs, can create a homogeneous social network. In such a setting, people are less willing to accept information or alternate views that contradict their own. According to David M.J. Lazer, it is within this context that fake news has been able to find a mass audience (Lazer et al., 2018).

Inside these homogeneous social-networks, fake news not only confirms but also reinforces and strengthens peoples beliefs. Political researcher Cass Sunstein refers to this phenomenon, where existing biases in a group are shifted in an extreme direction, as group polarization (Sunstein, 2002). Research has shown that there is a strong link between the spread of fake news stories and the radicalization of terrorists (Johnson, 2018).

Worryingly, research suggests that polarizing fake news can be spread faster by social bots. Researchers found that because of how often social bots post content, they may be viewed as more trustworthy and be better at influencing groups of people (Aiello et al., 2014). Researchers also found that bots played a large role in amplifying the presence of fake news stories in social-networks shortly before it went viral (Shao et al., 2018).

A large part of this problem is the increasing difficulty online users have in distinguishing bots and people. Many data-scientists have dealt with this by using machine-learning to detect social-media bots. However, there is a dearth of research on the opposite: how bots might avoid detection. Many methods for detecting bots have become less effective as bot-programmers catch on and refactor them (Chen et al., 2017). Rather than study how the behavior of bots may change over time, the goal of this research is to ascertain how effective a poisoning attack may be on reducing the effectiveness of existing methodologies in detecting bots on Twitter.

1.1 Using Machine Learning to Identify Bots on Twitter

When classifying social media accounts into two classes – genuine or automated, many data scientists use supervised machine learning algorithms. These algorithms are trained on data about social media accounts that have already been classified. These algorithms generate predictive models that can identify automated accounts, based on with features in the training data are most associated with bots.

Researchers Efthimion et. al. created just such an algorithm. Their algorithm was trained using the Cresci-2017 dataset, as well as an archive of 200,000 tweets and 400 user profiles confirmed to have engaged in malicious activity to influence the 2016 US Election (Popken, 2018). One way that bots may avoid detection is by posting the same messages repeatably, but slightly reworded. Efthimion et. al. measured the Levenshtein distance between the text of different tweets made by the same account to determine how similar they were. The resulting algorithm had a

10.1016/j.ifacol.2019.12.170

success rate of over 96% at correctly identifying bots from both data-sets, with the most important factors being whether or not an account was geo- enabled, didnt have a profile picture, and followed a high number of people. (Efthimion et al., 2018).

1.2 Adversarial Machine Learning

When classifying social media accounts as genuine or automated, many data scientists use supervised machine learning algorithms. These algorithms are trained on data about social media accounts that have already been classified. These algorithms then create a predictive model based on this training data, which scientists run on a test data-set. The predictive accuracy of these models is given by the ratio of test data that it correctly labels as genuine or automated to the data it misclassifies.

Adversarial Machine Learning is the idea that machine-learning algorithms that are tested and trained in a controlled environment may not perform as well in the wild and may in fact be more liable to deception and manipulation. In this research we are concentrating on effects of training data manipulation.

Researchers Xiao et. al. developed a framework for describing this type of setting. One of the attacks described in this framework is called a poisoning attack. In a poisoning attack, adversaries are aware of what kind of data is being collected and are able to introduce maliciously crafted samples. Xiao et. al. simulated a poisoning attack on algorithms meant to detect malware in PDFs, and found that by introducing only a few malicious data-points, they were able to increase the algorithms misclassification rate by ten times (Xiao et al., 2015a).

2. METHODOLOGY

This research is focused on experimenting with and determining the effectiveness of several poisoning attack methods. This research mainly makes use of the Cresci-2017 dataset, consisting of genuine accounts as well as social- bots and spambots, and fake-followers (Cresci et al., 2017a,b). This data was previously used to study the evolution of bots on Twitter over time, and is publicly available in the Bot Repository, an online data repository maintained by the Network Science Institute of Indiana University (Varol, 2018). The analysis of the Cresci-2017 data as well as the training and testing of models based on this data is performed using R-Studio.

Initial analysis of the data consists of establishing a ground- truth for the accuracy, precision, recall, and F1 Score of models trained on the Cresci-2017 data. These same metrics will be measured after several different poisoning attacks, in order to gauge their effectiveness. This process is repeated for models trained by several different algorithms, including K-Nearest-Neighbor, Generalized Linear Models, and Support Vector Machines. The most optimal attack method is determined based on the results of these methods, and is analyzed to determine how it would be implemented in a real-life scenario.

2.1 Initial Data Analysis

Before experimenting with poisoning the data, an initial data analysis must be performed. This means establishing baselines for the effectiveness of three binary classifiers that detect spambots, social-bots, and fake-followers when trained on non-poisoned data. To do this, the Cresci-2017 dataset was imported. 90% of the rows were randomly assigned to a training subset, and the remaining 10% were assigned to a test subset.

These were used to train and test three different models most frequently used for social network account classification one based on a Generalized Linear Model, one based on K-Nearest-Neighbor, and one based on a Support Vector Machine. After training these models, a confusion matrix was used to check their accuracy, recall, precision, and F1 Score.

This process was repeated several times to obtain the metrics for the models in distinguishing genuine human operated Twitter accounts from social-bots, fake- followers, and spambots, respectively. The effectiveness of the experimental poisoning attacks is measured against these baselines.

When designing an optimal poisoning attack, the features with the most impact on the classification of an account must be considered. Several methods were used to determine which features most heavily influenced the classification of an account. The first method was simply obtaining the Fisher scores for every feature (table B.2). Features with a higher Fisher score tend to more heavily predict the classification outcome of a model. The second method was measuring the impact on the F1-Score when a feature was excluded from a models training-data (table B.1). Features were ranked from most to least important based on which exclusions lowered the F1-Score the most. The average rank for each feature was calculated based on its order placement in both tables (table B.3).

2.2 Data Poisoning

After successfully establishing baselines to measure against, the next step is to measure the effectiveness of a poisoning attack. Several different methods of poisoning the data were used, based on previous research.

Label Flipping The first and simplest method used was a label-flipping attack based on research by (Xiao et al., 2015b), (Biggio et al., 2012). In the label-flipping attack, the data was poisoned by changing the class of a random row from bot to genuine or vice versa. Random samples of N% of rows were taken from the Cresci-2017 data-set, from N=1 to N=20. These rows were copied and had their classification labels flipped before being inserted back into the data-set. The Support Vector Machine (SVM), Generalized Linear Model (GLM), and K-Nearest Neighbor (KNN) algorithms were used to fit this poisoned data to several models, using 10-fold cross-validation. After generating these models, the accuracy, recall, precision, and F1-score were measured. In order to avoid bias toward a specific set of random samples, this process was repeated three times with different random samples taken to train and poison the data.

Feature Poisoning While a simple classification label flipping attack may be effective in raising the classification error of the model, but it is interesting to explore effects of poisoning the rest of the data. We experiment by altering the values of the features. The effectiveness of feature poisoning will be tested by selecting one feature estimated to have the most impact on the classification of Twitter accounts as genuine or as automated.

Two experiment control variables will be used to measure the effectiveness of feature poisoning. The first variable will be the percentage of points that must be poisoned, ranging from 1 to 30%. The second variable will be how much the value targeted feature will be altered, ranging from 1 to 2500. The accuracy metric will be used to measure each combination of percentage of poisoned rows and how much they are altered by. Experiments are run for each combination of control variables.

3. RESULTS

3.1 Initial Data Analysis

In Table 1, three different models have been trained using the GLM, SVM, and KNN algorithms to classify Twitter accounts from the Cresci-2017 data-set as either Genuine human users or Social-bots mimicking human users. The metrics for accuracy, recall, precision and the F1-Score are given. This process was repeated in Table 2 to classify accounts as Genuine human users or Fake-followers, and in Table 3 to classify accounts as Genuine or as Spambots.

Table 1. Results for Genuine/Social-bot Classifier

Model	Accuracy	Recall	Precision	F1 Score
GLM	93%	91%	97%	94%
KNN	98%	99%	98%	98%
SVM	96%	96%	98%	97%

Table 2. Results for Genuine/Social-bot Classifier

Model	Accuracy	Recall	Precision	F1 Score
GLM	98%	96%	99%	98%
KNN	95%	95%	95%	95%
SVM	95%	95%	95%	95%

Table 3. Results for Genuine/Spambot Classifier

Model	Accuracy	Recall	Precision	F1 Score
GLM	95%	91%	99%	95%
KNN	98%	97%	98%	97%
SVM	93%	94%	91%	92%

Based on the feature-selection process, the *favourites_count* feature was determined to be the most influential in determining how a Twitter account was classified. This feature represents how many tweets an account has liked. Bots tended to overall like a very low number of tweets. The tables used to determine this can be found in Appendix B.

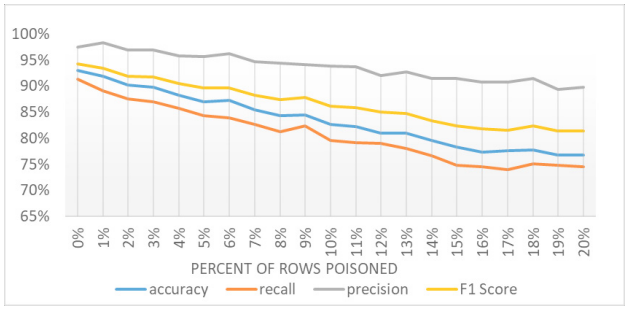

Fig. 1. Results of Label Flipping Attack against GLM-trained Model for Detecting Social-Bots.

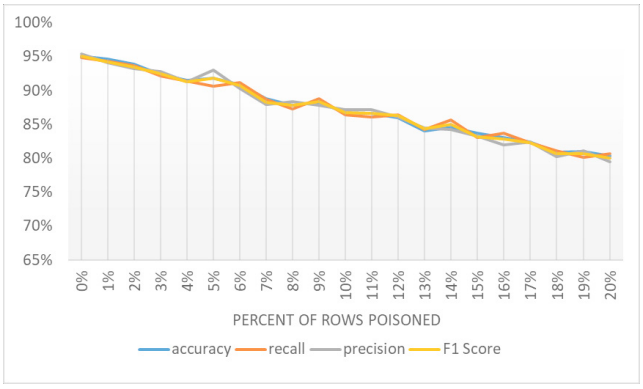

Fig. 2. Results of Label Flipping Attack Against KNN-Trained Model for Detecting Fake-Followers.

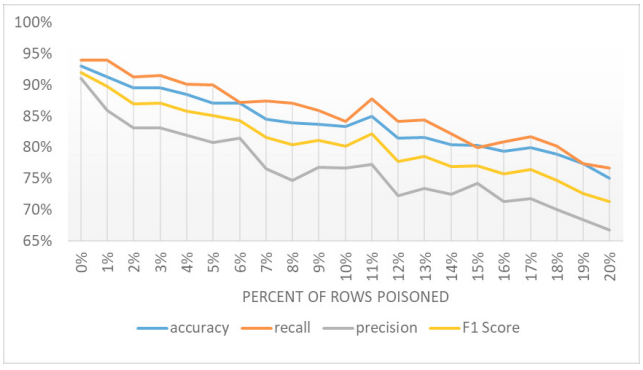

Fig. 3. Results of Label Flipping Attack Against SVM-Trained Model for Detecting Spambots.

3.2 Label Flipping

Fig. 1, Fig. 2, and Fig. 3 represent the results of a Label-Flipping attack on models trained to detect Social-bots, Fake-followers, and Spambots. These models are trained using the GLM, KNN, and SVM algorithms, respectively. As these figures illustrate, the resulting negative impact of poisoned data is consistent between different models.

The results of other models poisoned using the Label Flipping attack can be found in Appendix A.

3.3 Feature Poisoning

Fig. 4 represents the average results of 10 Feature Poisoning attacks on a GLM-trained model for detecting Spambots. The first variable measured is how many data-points are poisoned, from 1 to 30%. Each individual Feature

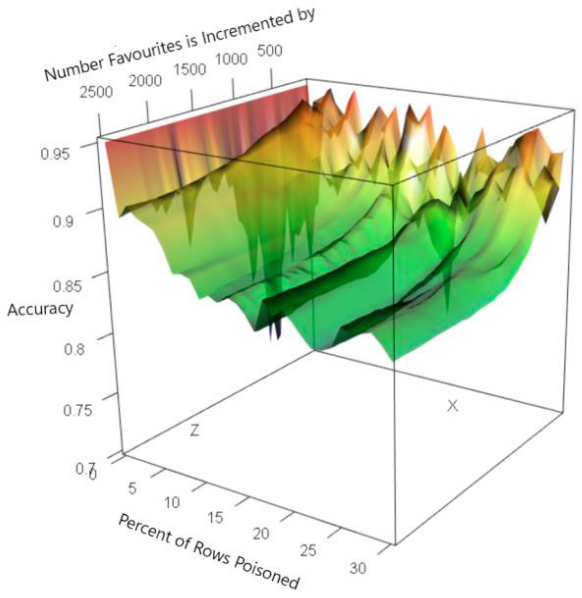

Fig. 4. Results of Feature Poisoning Attack Against GLM-Trained Model for Detecting Spambots.

Poisoning attack simulation targets a different set of randomly selected accounts. The favourites_count feature, or the number of tweets an account has liked, was determined to be the most influential in how an account was classified. Because of this, the second variable measured is how much the favourites_count of each account is incremented by, ranging from 1 to 2500.

In Fig. 4, the Y-axis represents the accuracy metric, the X- axis represents the percent of rows poisoned, and the Z-axis represents the value that favourites_count is incremented by.

Several other models were trained using KNN and SVM to detect Social-bots and Fake-followers. However, for these models, no significant drop in accuracy was recorded as a result of this Feature Poisoning attack. Focus was thus placed on the GLM-trained model for detecting Spambots.

There is a noticeable sudden drop in accuracy when only a small percentage of accounts are poisoned. We do not have proven explanation of this phenomenon but we assume the drop being caused by properties of data. Accounts that were poisoned were randomly selected, so it is assumed that this spike is the result of a specific set of accounts that, if poisoned, may cause a drastic drop in the accuracy of the model. Determining what this set of accounts is was not possible with the resources available for this study, and remains a topic for future research.

4. ANALYSIS

After analyzing the results, the next step is to determine an optimal attack strategy based on them. The results of the experiments show that a Feature Poisoning attack can be very effective if it targets specific data-points determined to have a high impact on the model. However, such an attack would be very computationally intensive in order to determine what the optimal data-points to target

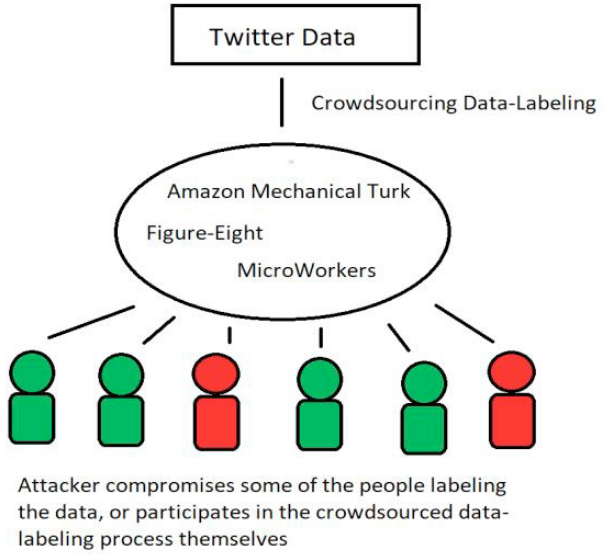

Fig. 5. Illustration of Proposed Attack Method for Targeting Crowdsourcing Process.

are. Exploring this more in depth is beyond the scope of this research. Therefore, based on the current results, the optimal attack method that is the least computer-intensive and takes the least time is the Label Flipping attack.

In order to implement this method, an attacker must have knowledge of what training data is being collected, when it is being collected, and by whom. Since this attack method explicitly targets supervised-learning algorithms, the attacker must also know who is labeling the data. Many researchers use crowdsourcing from sites such as Amazon Mechanical Turk or Figure-Eight to create a ground truth and efficiently label large data-sets (Cresci et al., 2017b). By targeting and compromising this process, an attacker can potentially reduce the ultimate accuracy of a Twitter-bot detector to 75%, similar to the results seen in Fig. 3. This attack is illustrated in Fig. 5.

5. CONCLUSION

The possibility of a poisoning attack against training data for Twitter Bot detection models deserves attention. Right now, these models are regularly retrained on new data as bots change their behavior to avoid detection. This training data is already difficult to collect. While obtaining Twitter account data using the Twitter API is relatively simple, finding pre- labeled data-sets of accounts confirmed to be human or bots is very difficult. The Cresci-2017 data-set used in this study contained a substantial set of Twitter accounts, but many of the automated accounts in the data-set were very old, some having been created as early as 2009 (Cresci et al., 2017b). If someone influenced or altered this training data maliciously, it would make it even harder to train models to accurately spot automated accounts.

Analyzing the results of this research shows that bot-detection models are very vulnerable to poisoning attacks. By introducing mis-labeled data-points into the Cresci-2017 data-set, the accuracy, recall and precision of the

models were lowered by as much as 20%. This label-flipping attack was determined to be the most optimal attack, since the impact was mostly consistent regardless of what data-points were mis-labeled or what algorithm was being trained on them. One way an attacker could practically implement this method is by participating directly in the crowdsourcing process that researchers use to label Twitter accounts as humans or bots.

Depending on what tools they have, attackers may be able to have an even greater negative impact on the accuracy of these models than the aforementioned label-flipping attack. Evidence for one such method was found in this study. By targeting a specific set of data-points in a poisoning attack instead of randomly-selected ones, an attacker could maximize their negative impact on the accuracy of a model in detecting bots. Hardware limitations prevented further investigation, but it is likely that data scientists with more resources may be able to drastically lower the accuracy of a model for detecting bots by altering only a few data-points, similar to Biggio et. al.s initial research (Biggio et al., 2012).

By better understanding this form of adversarial-machine-learning, data scientists can make their own algorithms for detecting bots more resilient. In doing so, they can reduce the impact of these poisoning attacks, and ultimately mitigate the negative impact that malicious bots have on social-media.

REFERENCES

Aiello, L.M., Deplano, M., Schifanella, R., and Ruffo, G. (2014). People are strange when you're a stranger: Impact and influence of bots on social networks. *arXiv preprint arXiv:1407.8134*.

Biggio, B., Nelson, B., and Laskov, P. (2012). Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 1467–1474. Omnipress.

Chen, Z., Tanash, R.S., Stoll, R., and Subramanian, D. (2017). Hunting malicious bots on twitter: An unsupervised approach. In *International Conference on Social Informatics*, 501–510. Springer.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017a). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 963–972. International World Wide Web Conferences Steering Committee.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017b). Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15(4), 561–576.

Efthimion, P.G., Payne, S., and Proferes, N. (2018). Supervised machine learning bot detection techniques to identify social twitter bots. *SMU Data Science Review*, 1(2), 5.

Johnson, J. (2018). The self-radicalization of white men:fake news and the affective networking of paranoia. *Communication Culture & Critique*, 11(1), 100–115.

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science (New York, NY)*, 359(6380), 1094.

Popken, B. (2018). Twitter deleted 200,000 russian troll tweets. read them here. *NBC News*, 14.

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1), 4787.

Sunstein, C.R. (2002). The law of group polarization. *Journal of political philosophy*, 10(2), 175–195.

Varol, O. (2018). Bot repository. URL `https://botometer.iuni.iu.edu/bot-repository/datasets.html`.

Xiao, H., Biggio, B., Brown, G., Fumera, G., Eckert, C., and Roli, F. (2015a). Is feature selection secure against training data poisoning? In *International Conference on Machine Learning*, 1689–1698.

Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., and Roli, F. (2015b). Support vector machines under adversarial label contamination. *Neurocomputing*, 160, 53–62.

Appendix A. LABEL FLIPPING ATTACK RESULTS

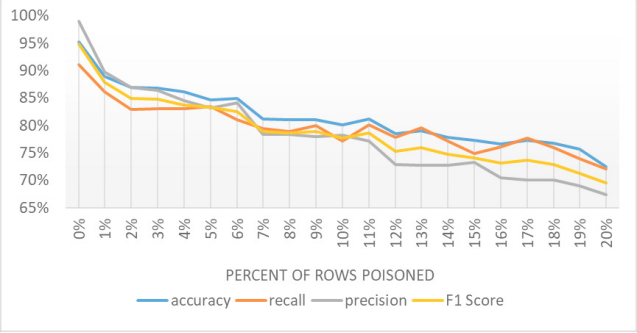

Fig. A.1. Results of Label-Flipping Attack Against GLM-Trained Model for Detecting Spambots.

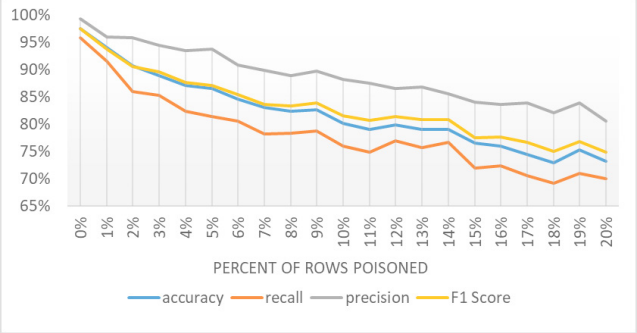

Fig. A.2. Results of Label Flipping Attack Against GLM-Trained Model for Detecting Fake-Followers.

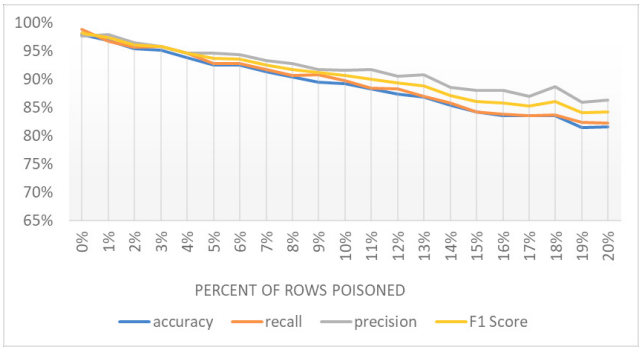

Fig. A.3. Results of Label Flipping Attack Against KNN-Trained Model for Detecting Social-Bots.

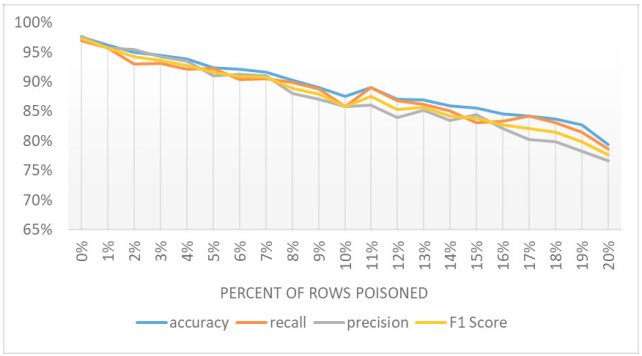

Fig. A.4. Results of Label Flipping Attack Against KNN-Trained Model for Detecting Spambots.

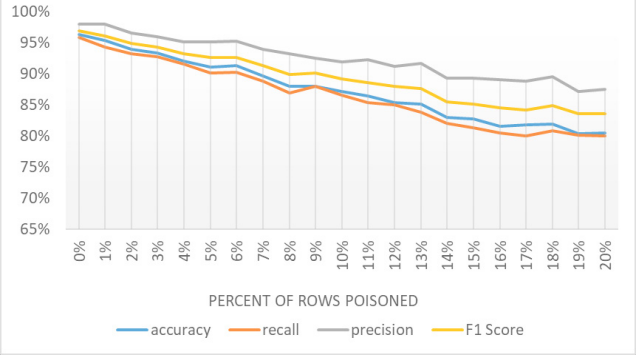

Fig. A.5. Results of Label Flipping Attack Against SVM-Trained Model for Detecting Social-Bots.

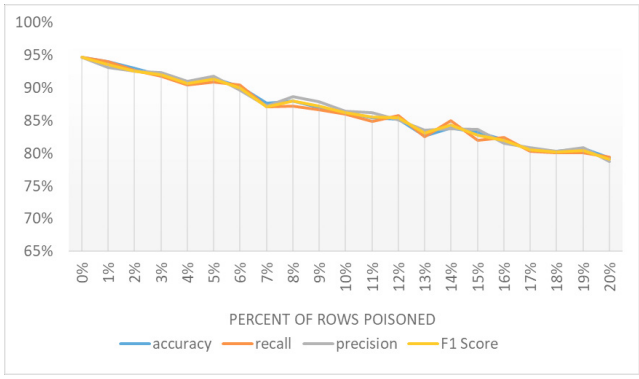

Fig. A.6. Results of Label Flipping Attack Against SVM-Trained Model for Detecting Fake-Followers.

Appendix B. FEATURE SELECTION RESULTS

Table B.1. Calculated F1 Score for Excluding Each Feature

Feature	F1 Score without
favourites_count	0.8945241
Utc_offset	0.9224079
Statuses_count	0.9261423
friends_count	0.9352764
Followers_count	0.9357285
lang	0.9362008
Listed_count	0.9362119
Created_at	0.9365591
Geo_follow_protect_verify	0.9371561
Time_zone	0.9481067

Table B.2. Fisher Score of Each Feature

Feature	Fisher Score
Created_at	0.69
Lang	0.42
Statuses_count	0.2
Favourites_count	0.16
Time_zone	0.14
Utc_offset	0.069
Geo_follow_protect_verify	0.058
Friends_count	0.0011
Listed_count	0.00035
Followers_count	0.000032

Table B.3. Features in Descending Order of Importance

Feature
favourites_count
statuses_count
Lang
utc_offset
created_at
friends_count
time_zone
followers_count
Geo_follow_protect_verify
listed_count