Machine Learning Based Classification of Twitter Accounts

Chinmay Sumant
Dept. of Computer Science
NYU Tandon School of Engineering
New York City, USA

Vishesh Kakarala
Dept. of Computer Science
NYU Tandon School of Engineering
New York City, USA

Suril Shah
Dept. of Computer Science
NYU Tandon School of Engineering
New York City, USA

***Abstract*— The emergence of automation in the domain of Social Networks and information sharing presents an incredibly dynamic Machine Learning problem. If automated accounts which are referred to as bots are present only for information sharing, it is a welcome prospect. However, another aspect of bots is spamming and false information spreading. For these or other reasons, classifying a twitter account as a bot or human operated account poses an interesting problem. To obtain an accurate, scalable solution, we are building a machine learning based solution which uses multiple algorithms and chooses the most suitable and accurate model for the given dataset.**

Keywords—machine learning

I. Introduction

Twitter is a widely used social networking site that today doubles as a source of information. The versatility and ease of use of twitter have attracted a large user base who generate millions of tweets every day. Twitters readily available API and its simplicity have attracted many automated accounts, also known as bots. There are two types of bots, harmless bots which mainly aggregate information or perform a task, the other spreads spam and may also be used to bloat user engagement statistics. We will conduct a series of analysis to identify these bots and differentiate them from Human users.

II. Motivation

In today’s world of automation, humans are trying to automate whatever we can set our eyes on. Social Media is one such field where automation is being enforced. On social media, the use of bots has become prevalent. On Twitter, there are many handles that relay or retweet, tweets and just do that. They are many handles that are not fully operated by humans but are run governed by a block of code. Such handles can be classified as bots. As per the definition of a bot given above, it is near impossible to physically/manually classify multiple accounts as bots and human operated accounts. But, this can be achieved and automated using machine learning. A Machine Learning solution is also scalable to thousands, millions, and even all of twitter using modern day Big Data processing solutions.

Hence, the motivation for this project is to classify a handle or twitter account that we are following or plan to follow as a bot or human operated account, using Machine Learning methods.

III. Related Work

A lot of work has gone in to identifying spam bots on internet user platforms. Our work will be closely related to spam detection used to identify automated profiles. Zi chu et al [4] have worked on identifying the differences in Humans, cyborgs and bots on twitter. Lokot .T et al [5] studied the use of news bots in social medias shift towards automated journalism. Olof Larsson et al [6] worked on understanding the popularity of online news media through understanding the mitigation of social bots. Alex wilkie et al[7] developed their own social bot to understand the and document the way in which a bot can evoke instability in its interactions. Stefanie Haustein et Al[8] studied the implication of the presence of such bots from the perspective of social media metrics (altmetrics), where mentions of scholarly documents on Twitter have been suggested as a means of measuring impact that is both broader and timelier than citations. Emilio Ferrera [9] discusses research on potentially malicious social bots that use artificial intelligence on social media networks.

IV. Data

The training dataset for the analysis has been extracted through the twitter API. We used a combination of crowdsourcing efforts to identify whether the accounts are bots or nots. Twitter lists has a functionality where a user can share publicly a list of other users. We used these publicly shared and curated lists to identify bots and human user. For the not bots we extracted users from Mashables list interns for fall 2013 [11], assuming Mashable does not employ Bots as interns and Mashable list of speakers at connect 2012[10]. For the Bots, we extracted accounts from a list of bots[13] that are related to various scientific literature curated by the Lille university of science and Technology.

V. Algorithm(s) used

We plan to use and compare the following algorithms to classify twitter users as bots or not bots.

- Gaussian Naïve-Bayes Classification: - Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. If the Naïve Bayes conditional independence assumption holds, a

Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. And even if the Naïve Bayes assumption doesn't hold, a Naïve Bayes classifier still often does a great job in practice.

- Support Vector Machine: - Support Vector Machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space.

- Logistic Regression: - Logistic Regression, is a regression model where the dependent variable (DV) is categorical. It uses binary dependent variable—that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. You also have a nice probabilistic interpretation, unlike decision trees or SVMs, and you can easily update your model to take in new data, again unlike decision trees or SVMs.

- Random Forest Classifier: - Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier. Random Forest runs efficiently on large databases. It can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification. Random Forest generates an internal unbiased estimate of the generalization error as the forest building progresses.

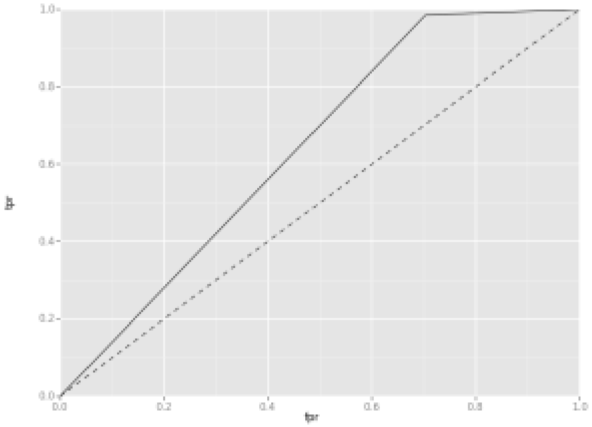
- Multinomial Naïve-Bayes: - Multinomial Naive Bayes simply lets us know that each *p(f$_i$|c)* is a multinomial distribution, rather than some other distribution. This works well for data which can easily be turned into counts, such as word counts in text. Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.

- Gradient Boosting: - Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

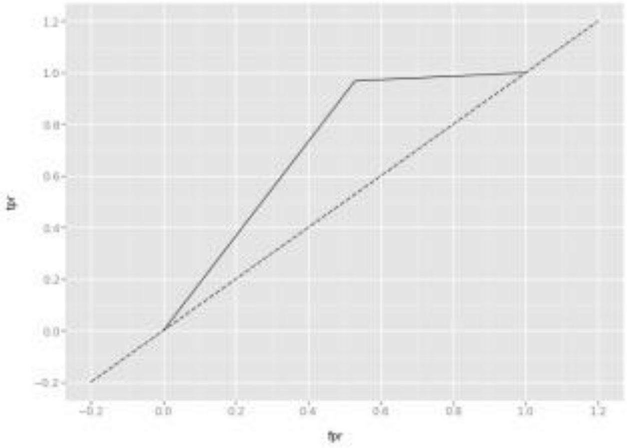
VI. Result

10-fold cross-validation score obtained for the algorithms above is as follows: -

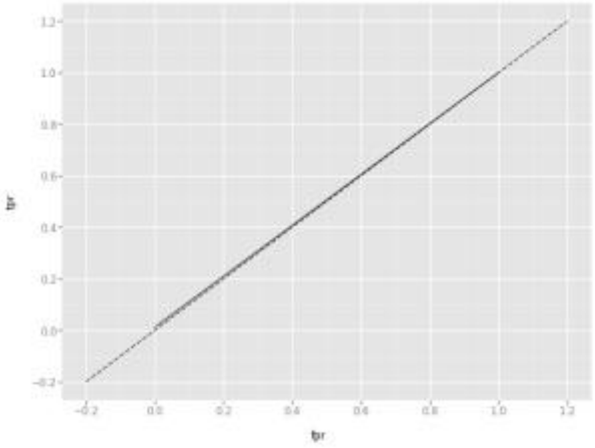
- Gaussian Naïve-Bayes: - 63%



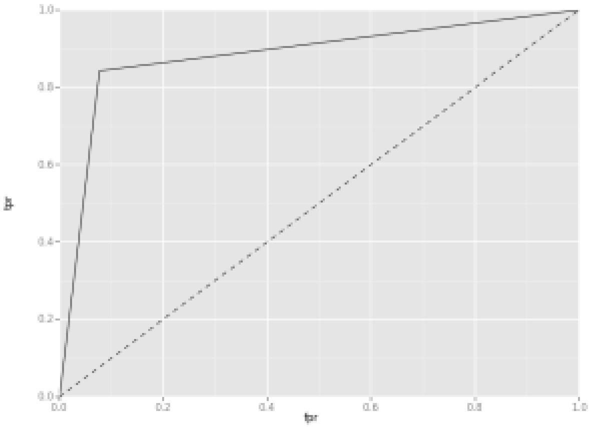
- Logistic Regression: - 80%
- Multinomial Naïve-Bayes: - 73%



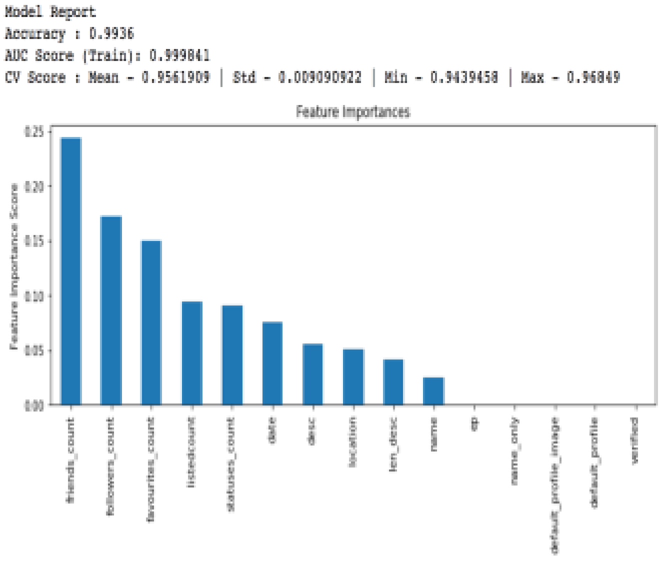
- Support Vector Machine: - 61%



- Random Forest Classifier: - 99%

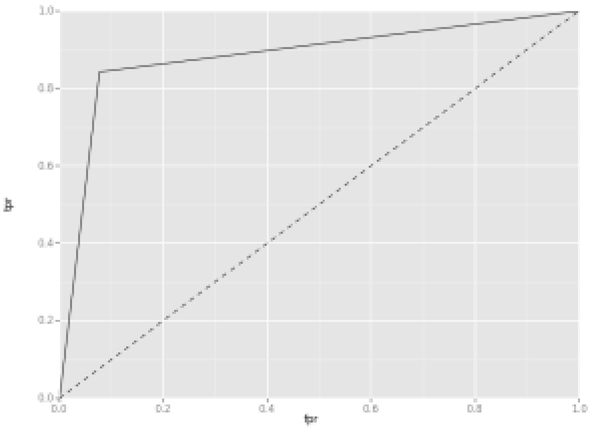


- Feature Importance (Random Forrest)

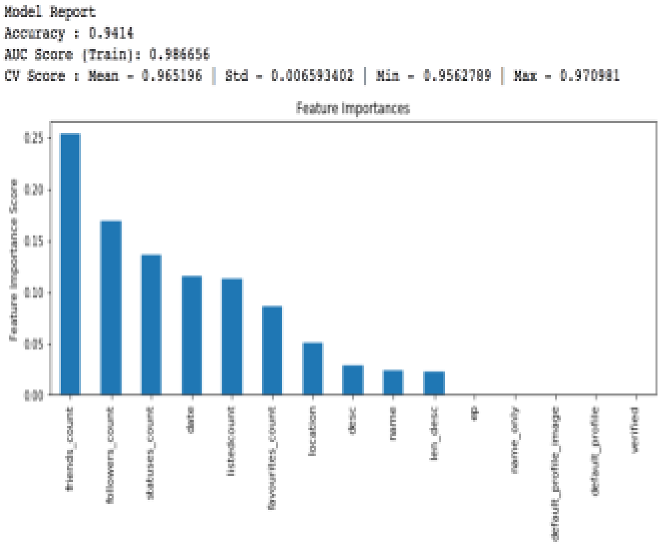


```
Model Report
Accuracy : 0.9936
AUC Score (Train): 0.999841
CV Score : Mean - 0.9561909 | Std - 0.009090922 | Min - 0.9439458 | Max - 0.96849
```

- Gradient Boosting: - 94%



- Feature Importance (Gradient Boosting)



```
Model Report
Accuracy : 0.9414
AUC Score (Train): 0.986656
CV Score : Mean - 0.965196 | Std - 0.006593402 | Min - 0.9562789 | Max - 0.970981
```

VII. CODE

https://github.com/chinmayms/twitterbot

VIII. VIDEO LINK

https://www.youtube.com/watch?v=_eKls1t1TNY

IX. EVALUATION

- After our mid-way analysis, we concluded that Random Forest classifier would be the best fit for the given problem.
- The numerical features fitted perfectly into the algorithm and did not require cleaning.
- Playing around with different features and parameters of random forest using sklearn provided different results every time and gave us an opportunity to learn more about the dataset and tune parameters for max results and accuracy.
- Text data in spite cleaning, tokenizing and stemming did not provide any boost to the classifier.
- TFIDF tokenizing and NLTK stemmers reduced the accuracy of the model which is very odd.
- Given more time, we would have worked on integrating text features smoothly into the model so as to provide meaningful insights inside the text patterns of bots and non-bots.
- Another aspect is, to add more features using twitter api to look at other account characteristics which were not looked at during this attempt over the semester and could add more value to the current model to increase accuracy further.

X. CONCLUSION

In this paper, we have presented a feature-based prediction model that can automatically identify whether a given twitter handle is a bot or not. Since numerical features are dominant in comparison to text based features, we found Random Forest Classifier to be the best fit for this prediction. Overall, we learnt that in such problems, given the current dataset, numerical and boolean features dominate the model by providing standardized characteristics of each account. Text data on the other hand is highly volatile and much more difficult to normalize in order to make characteristics comparable. This, in spite the 140-character limit given by twitter.

REFERENCES

[1] Ferrara, et. Al. The Rise of Social Bots. https://arxiv.org/abs/1407.5225 o Lee, et al. Who Will Retweet This?

[2] Automatically Identifying and Engaging Strangers on Twitter to Spread Information. https://arxiv.org/ftp/arxiv/papers/1405/1405.3750.pdf

[3] Shellman, Erin. Bot or not. http://www.erinshellman.com/bot-or-not/

[4] Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2010. Who is tweeting on twitter: human, bot, or cyborg? In Proc. 26th Annual Computer Security Applications Conf. (ASAC)

[5] Lokot, T., and N. Diakopoulos. "News Bots: Automating news and information dissemination on Twitter." Digital Journalism 4, no. 6 (August 17, 2016): 682-699. Scopus®, EBSCOhost (accessed March 13, 2017).

[6] Olof Larsson, A, & Moe, H 2015, 'Bots or journalists? News sharing on Twitter', Communications: The European Journal Of Communication Research, 40, 3, pp. 361-370, Communication & Mass Media Complete, EBSCOhost, viewed 13 March 2017.

[7] Wilkie, A, Michael, M, & Plummer-Fernandez, M 2015, 'Speculative method and Twitter: Bots, energy and three conceptual characters', Sociological Review, 63, 1, pp. 79-101, SocINDEX with Full Text, EBSCOhost, viewed 13 March 2017.

[8] Haustein, S, Bowman, T, Holmberg, K, Tsou, A, Sugimoto, C, & Larivière, V 2016, 'Tweets as Impact Indicators: Examining the Implications of Automated "bot" Accounts on Twitter', Journal Of The Association For Information Science & Technology, 67, 1, pp. 232-238, Business Source Complete, EBSCOhost, viewed 13 March 2017.

[9] FERRARA, E, VAROL, O, DAVIS, C, MENCZER, F, & FLAMMINI, A 2016, 'The Rise of Social Bots', Communications Of The ACM, 59, 7, pp. 96-104, Business Source Complete, EBSCOhost, viewed 13 March 2017.

[10] https://twitter.com/mashable/lists/connect-2012-speakers/members

[11] https://twitter.com/mashable/lists/fall-2013-interns

[12] https://twitter.com/Alexis_Verger/lists/twitterbot