

# Identify the Human or Bots Twitter Data using Machine Learning Algorithms

## Nambouri Sravya<sup>1</sup>, Chavana Sai praneetha<sup>2</sup>, S. Saraswathi<sup>3</sup>

<sup>1,2</sup>Student, Department of Computer Science and Engineering, Panimalar Engineering College, Tamil Nadu, India <sup>3</sup>Associate Professor, Dept. of CSE, Panimalar Engineering College, Tamil Nadu, India \*\*\*

**Abstract** - we implement this technique to identify the malicious activities in social contact. The increasing number of accounts in social media platforms is a serious threat to the internet users. To detect and avoid fake identities it is need to understand the dynamic contagion. In exist; there are many models to detect the fake identities by bots or humans. Sybil identities are generally focused on famous social media platforms. The proposed system discussed in this paper is to detect the Sybil and troll identities using machine learning engineered techniques.

#### Key Words: Machine learning, fake accounts, data sets, social platforms

## **1. INTRODUCTION**

The platforms of social media have a great impact on many areas today. In this we are focusing to identify the Sybil and troll identities in the platforms of social networks. There are many identities that are threats and malicious to the people on internet. So to identify the platforms of fake identities we use this supervised machine learning techniques to overcome of these fake identities.

In this the data sets are collected by the large data collection blogs. The data is stored and if any data is found malicious the data is cleaned and stored again. This gets the data more accurate of the user whether the account is a Sybil or troll identities/accounts using advanced techniques. This makes the platforms free of malicious activities to some extent.

Once the data is cleaned the spaces where the data is missing is filled. This shows that the missing spaces are fake identities and filling space are the cleaned fake identities. Before, the data is cleaned it is stored in non-relational database. Therefore, gets the data sets in a collection for future reference and remove the fake profiles.

Then they predict the accounts of social networks that are threats or ward. Using machine learning helps to find the fake identities of many social platforms. This growth in areas of internet makes the accounts more reliable and trustworthy for the users. Then the accounts are iterated in machine learning algorithms to identify the fake profiles over the internet.

There is iterative training in machine learning to get the data and store in database. The activities in the accounts are identified as menace or protected in SPM. Finally, the results of identifying bots and troll identities are visualised and resulted by supervised machine learning algorithms.

### **1.1 Proposed System**

Create a social media tweets, hash tags, social media posts, feeds, comments. Create non-relational databases. Using a data set preparation and cleaning. Then create a dataset. Applying the Ml supervised machine learning algorithms. Finally evaluate and visualize the results. It gives accuracy more than 90%. It is an real time data analytics.

#### **1.2 Existing System**

During the process of detecting the fake identities humans and bots have same behavior. These are applied to many supervised machine learning models. Many engineered features are existing but are not much successful in implementing to detect the malicious accounts. Existing system use only two parameters.

"Friend-to-followers ratio."

Friend count

Less prediction accuracy

Not an real time analysis

Existing system not used for an long dataset.

Accuracy in supervised algorithm is 68 %

The existing system is not much featured to detect troll accounts then the bots accounts.

The prediction of identity is not much accurate.

The existing system focused on twitter to identify the fake identities.

Create a social media tweets, hashtags, social media posts, feeds, comments. Create non relational databases. Using data set preparation, cleaning .Then create a dataset. Applying the MI supervised machine learning algorithms. Finally evaluate and visualize the results. Its gives



accuracy more than 90%. Its an real time data analytics. The existing system detect fake identities to 50% of accuracy. Three types of machine learning algorithm are used to detect the fake identities. The model is dependent on features (name, location, profile image).

2 cross validation and resampling methods are used in machine learning to detect fake identities.

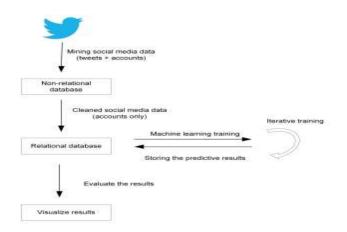


Fig. 1 process architecture

Data collection is the first activity. It is collected from various social media networks (twitter, kaggle, data.gov) etc. Then create non-relational databases. Then cleaning process is started after that the data is stored in relational databases. Then train the dataset using supervised machine learning algorithms (Linear regression, Navies Bayes). Finally the results are visualised and evaluated.

#### 2. Modules

- 1) **Data collection:** Real time data collected from Twitter, kaggle, UCI, Data.gov
- 2) **Data Cleaning:** fill the missing data and cleaning the noise data.
- 3) Machine learning algorithm: In this module we use linear regression and Naive bayes supervised algorithms
- 4) **Compare the machine learning model**: Finally we create a compare model for other algorithms and also visualize the results

### DATA COLLECTION:

Real time data collected from Twitter, kaggle, UCI, Data.gov. Collection of data is one of the major and most important tasks of any machine learning projects. Because the input we feed to the algorithms is data. So, the algorithms efficiency and accuracy depends upon the correctness and quality of data collected. So as the data same will be the output.

-											
-	-					-					
<sup>-</sup>	And in case of the local division of the loc			_			Larger and		and the second second	Contract of the	
	and a failed at								100	Contraction of the	
-	The second secon						The second second			Contraction of the	
	and the local data		the local statistics				The second second				
	Contraction in contraction		tally street.	-			- manufacture		the second	international and the	
1.00	The second		the second se					······································	10.00	International Action	
100	10011						A DECEMPTOR	5.22		Contraction of the local	
	10000	_	trant di platigi presi					······································			
	Land Sectors in the		The second se	-							
	president and the second strength		Married Street, Street, or other				al la la second			Contraction in the	
	interest opposite		all shares a local sector of the				-		100.0	Contract of the	
	And in the second second			-			100000000000	10 Aug		Danked of the line	
	and the second second		the state in the second second				Contraction of the	1.		interesting the last	
	ALC: NOT THE OWNER.		and the				100000000	1			
	and the second second	the second second			1.2	-		1.22			
_	same in the second second						a later for the		1000	in the second se	
_	Internet Concernity of						- In contraction	C	100	in the second section of the	
	and the second second										
	where the second second						Contract of the local sectors			- 100	
	New Yorks		-			-	Contract States	······································		" their way	
	and the second second		wanted one opposite state		- 2		the second se			Company of the	
-	with the local local		instance if a large state in the	-			- system		1.000		
_	the second second second		the second s							Complex 11	
	the second						S provide				
_	- The second second		the second se					· · · · · · · · · · · · · · · · · · ·		and the second second	
	and state in some states		second of the second second					· · · · · · · · · · · · · · · · · · ·		Contraction of the later	
_	and made classifier		Addressed reprint to							Contraction of the	
	- Contract Contractor		Inclusion and part was in addition		- 2-			······································		Contraction of the	
_	Institute Constitute		second and had been been as				1000 00.00000		0.000	International Arrists	
	Colorado Antonio Anto		series series which have been				10.040.00000	1 104		internet by the	
_	the second se		and the second s		100		in beingen		tall and	the second second second	
	And shares in succession of the second		the second second second		-	-	the contract of	· · · · · · · · · · · · · · · · · · ·		internet internet	
	and a second sec		concerning in the second second				1 10000000		-		
	manual second						a lot of the lot of th			Contract in the local division of	
_	and a second sec		Charles with the Alastantic Street Street Street Street				10.000 million (1.000 million)	C 20 Mill	100.00	interesting of the local sector	
_	Contractor and the local division of the loc	- and a second second	Support and an and a support of the local division of the local di		-		A		-	the summer of the	
	delle se contracte			-	-	-	·				
-	Contraction of the local division of the loc		Internet of the second second second				index and			Contractor of the	
	and the second s		factor and the second se				Internal Constants		1.000	interesting the line	
	pretty and an internal state	distance in the	Charles of Anna Street Street, etc.				a Linterne				
	and the second second	and the second s	a disc data and the second states					1.122		The second second	
	I manufacture in the second second		the second se		12		a providence				
-	States and the second s	distant and	the second second second second					12.22			
1.00	the second se						100/14/14		-	Contract of the last	
100.000							and the second second				
											All and a second second

Fig. 2 Data collection (Bots data)

#### DATA CLEANING:

Collecting the data from one task and making it useful to another data is an-other vital task. Data collected will be in an unorganized format and there may be lot of null values, in-valid data values and unwanted data from various means. Cleaning all the data and replacing them with the approximate data and filling the null and missing data with some fixed alternate values are the basic steps in preprocessing of data. Even data collected may contain completely garbage values. It is not necessary to be in exact format what it want to be can be in any format. This process is made to keep the data meaningful and for further processing. Data must be kept in an organized format.

	and t									
	1000 Aug 201	lands.		Strain load dos		Concernant of the	and in such that	North Add	me. this prime in a second	
and the second	of the local diversion of	- market and	minutes and states in the	in the second se		-	100.00	- miles	county have need that provide	
- and the second			internet of the second s	-		and the second	and had		closed and used has been	
and and	and the implication.				- 17	1.0000	1 I I I I I I I I I I I I I I I I I I I	1.00	intent the titl are started	- 1
rine i	(Plant) manufactory	and the second s	promotion delivery or a		14	ind woman is	100 100	1000	county had look this makes	
the second second	Contract Income in the	100			+-	-	100 100	100.0	and field 1940 fair sheets	- 1
and the second second	COLUMN AND A					C. Distant.	11 Parts.	1.000	Number Till 1911, Mult Augen-	
and the second s	manufacture interactions	1000	Management in the street			1.000	11.000	4.0	ment, but the set mouth	
and the second s	CONCERNING AND	(Phillip A), Prycholaet	which are the total commercial data.				100 100	100.00	control tool that they be the	
_	and the second second		Warman and Street and Street S			10 mm	plan when		come this last this mean	
	1010 mile	110.00	The Rev Proc. Name & Address of the	- 1004		100.00	-0.1004	46.0	Tradit, No. 1988, No. 494-144	
-	and the second second	interested and	The second s			10000	AL 1998		many the real and many	
-	And in case of the	10100.0	and it was a second the second of				1.1 1946	194.00	loader that that we could	
-	the second second		Internet Street Street	40.		11/10/00	1.94	100	times the last had peaks	
-	State Incode	Second States		1000		1992 - 114	100.000		time of the set have	
-		terms of the local division of the	Contract on the contract of			10.000	12		100 10 10 10 10 10 10 10 10 10 10 10 10	
		Section 2	State in the second second		- 2	1.000	2.2	- 12.	frame and lost and include	
	-	1100	In the order of the local division of the		- 2		2.22	100	where the last we been	
-	distant.	-	internal and internal sectors.	in an	- 2	Man Includes			income first link fact from	
	the second se	· Contract and	the second second	-			10.00	10.0	timber the lots had been	
-	Contrast Anna		Internal course of the line.	and the second s		100.000	100.00	-	instant bird long land land	-
	Cardinan Incontraction	internal second	strate and in the local dist	-		101		-	times the last his hand	
	COLUMN CONTRACTOR	family and the second	and the second sec		14		10.000		many the last had been	_
1001	and the local day		the lot of the lot of the lot	100		-	1.0	100	the second second second	
-	second and stands	1444	Water for the part of the local division in some			1.000	in test	100.00	the last tail had then	- 1
and the local division of the local division	instant and and		ing beings the station				10. Peeb		mainer, day hard , first second	- 1
-	William A Write P	Teaching and the	ing station was specify	-		industry. In	the loss	10012-000	county were that they reprise	
and the local division of the local division	CONTRACTOR AND ADDRESS	in the second se	and the second s	10.04	- 44	10110-001	ALC: THE	1000	county have been been been been been been been be	
-	dealers have been	100 million (100 million)	The second second second second			(1990 mil)	140.100		carrie had test the heart	
-	1001100-0004		- marine	10000		desident in the second	at. 100		plane had like had been	
	contract count			1000		10.000			Anne 194 198 No. Web.	
-	and the second second	100	100-014			10.00	. 4. 66	104.00	comment and comments where a	
-	and the second second		100-01	1000					man has los on march	
-	2010 particular	where an other	And a part of the second second second			1.000	10,000	100.00	inter hid 14.6 hid place	
1000	and the second second	and a second	and the divertised of the			7444	10.00	-	come that this the burns	-
-	and the second s								"balance from the second	
	COMPANY AND ADDRESS OF	(MIN)	And the second s		-	·*=	52		Append Non 1987 and 1978	
_	and the set				- 75		- C - C - C	- 22-	income the local first framework	
Later -	A REAL PROPERTY.	Terror of Contract	1-100 PT-100 T				12.22	100	inter the last but have	
and -	and the second s		Contraction of the local of the	- 21			20		internal first light have been	
Land I	Contraction of the local division of the loc	(and the second	and the contract lots				- 2 - 2 -		Trans. The last 1 mile	
-	and the second sec			-	- 2	-	10.000	100	instant, but the loss much	
-	And in case of the local division of the loc	And and a local division of the	Industry of the local division of	-				10.0	"Inner had that had been	
Real Property lies	Index and the			-	- 7	-		444	server last that the band	_
10.00						Constitution of the				

Fig.3Non-bots data

	膨	THE SALE	huiv	interiories and	alisions'	斑	2	218)	82
-	135	(Elline)	łzł	istene para la		π	B	1===	重现
	<u>88</u>	<b>EFE</b> rpites	38	देव्हीलक, पेत्र 🕴	taliaketa	語	1	Eleks)	11
Ē.	12	TORN ADDR	iel .	in data netrati at, ktarti	trichell	33	ŝ	新設施	11
2	38	व्याः कृति	8000	Biaeladinitidunal I	vERolat	<u>19</u>	Ø	Bing	81
10	1287	19-Jaid	kies,17	wategines, cones, intel percent	æ	4	ž	(==	22
5	192	Marka	Silte	kanifeter Billio richner	talialisitt	12	20	Wiele.	81
5	1781	服制的		ntitalipti		痴	ų.	327128	93
10	807	2017,enging	leitä.R	uet) (services) (auto	9.869.60	3	В	1	12
ī.	TELE	60000	1 april 1	lessi 1050'n witzlich	talian Mail	囲	¥	ENG	21
8	BER	HER Stabel		ispesių įšekurikasi.		Я	Į.	5=	1 R
5	281	<b>BREB</b>	<b>hint</b> 系	Velapera Star Alexand Star	talia Mis	34	3	Tint	81
8	333	385 terret	<b>West</b>	lare.		I	3	(=	12
I.	愿	and the second s	NASICIE	id naisiosite, le 1936 il	alia%eko	题题	85	Mah	31
2	367	Electi	HEAR	lerisfishtesp & hearit	eleide	36	1	Xieri	51

ISO 9001:2008 Certified Journal

Page 1403



International Research Journal of Engineering and Technology (IRJET)Volume: 06 Issue: 03 | Mar 2019www.irjet.net

e-ISSN: 2395-0056 p-ISSN: 2395-0072



Fig.4 Sample source code

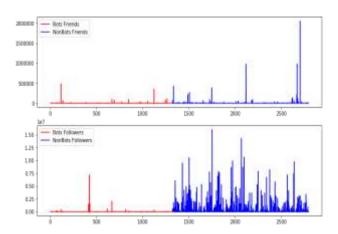
#### MACHINE LEARNING ALGORITHM TRAINING:

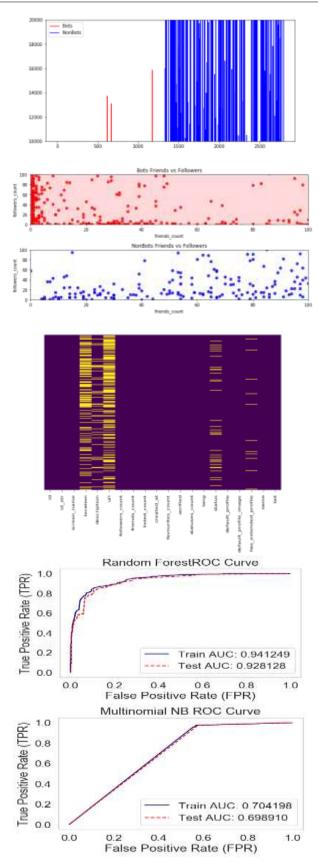
The next step is algorithms are applied to data and results are noted and observed. The algorithms are applied in the fashion Mention in the diagram so as to improve accuracy at each stage.

#### **TRAINING AND TESTING:**

Finally after processing of data the next task is obviously testing. In this process where performance of the algorithm, quality of data, and required output all appears out. 80 percent of the data is utilized for training from the huge data set collected and 20 percent of the data is reserved for testing. Training is the process of making the machine to learn and giving it the capability to make further predictions based on the training process. Whereas testing means already having a predefined data set with output also previously labeled and the model is tested whether it is working properly or not and is giving the right prediction or not. If maximum number of predictions is right then model will have a good accuracy percentage and is reliable to continue with otherwise better to change the model.

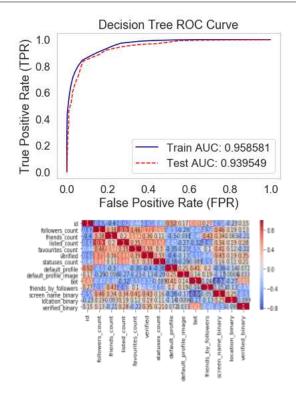
### ExperimentalResults







International Research Journal of Engineering and Technology (IRJET)e-ISSN: 2395-0056Volume: 06 Issue: 03 | Mar 2019www.irjet.netp-ISSN: 2395-0072



### **3. CONCLUSIONS**

A dataset of all media platforms are collected and maintained. In this paper the overall concept of the process has been explained as a summary. The accuracy of the process is ensured. For the future use they may extend their accuracy even more with another algorithm.

A powerful web application can be developed where inputs are not given directly instead student parameters are taken by evaluating students through various evaluations and examining. Technical, analytical, logical, memory based, psychometry and general awareness, interests and skill based tests may be designed and parameters are collected through them so that results will be certainly accurate and the system can be used realiably.

Also decision trees have few limitations like overfitting, no pruning, lack of capability to deal with null and missing values and few algorithms have problem with huge number of values. All these can be taken into consideration and even more reliable and more accurate algorithms can be used. Then the project will be more powerful to depend upon and even more efficient to depend upon.

### REFERENCES

[1] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, "Profile characteristics of fake Twitter accounts," Big Data Soc., vol. 3,no. 2, p. 2053951716674236, 2016, doi: 10.1177/2053951716674236.

[2] C. Xiao, D. M. Freeman, and T. Hwa, "Detecting clusters of fake accounts in online social networks," in Proc. 8th ACM Workshop Artif. Intell. Secur., 2015, pp. 91–101.

[3] S. Mainwaring, We First: How Brands and Consumers Use Social Media to Build a Better World. New York, NY, USA: Macmillan, 2011.

[4] V. S. Subrahmanian et al. (2016). "The DARPA Twitter botchallenge."[Online]Available:https://arxiv.org/abs/1601. 05140

[5] Y. Li, O. Martinez, X. Chen, Y. Li, and J. E. Hopcroft, "In a world that counts: Clustering and detecting fake social engagement at scale," in Proc.25th Int. Conf. World Wide Web, 2016, pp. 111–120.

[6] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: An analysis of Twitter spam," in Proc. ACM SIGCOMM Conf.Internet Meas. Conf., 2011, pp. 243–258.

[7] T. Tuna et al., "User characterization for online social networks," Social Netw. Anal. Mining, vol. 6, no. 1, p. 104, 2016.

[8] P. Galán-García, J. G. De La Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying,"Logic J. IGPL. vol. 24, no. 1, pp. 42–53, 2015.

[9] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: Characterizing and identifying fake images on Twitter during hurricane sandy," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 729–736.

[10] J. P. Dickerson, V. Kagan, and V. S. Subrahmanian, "Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?" in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2014, pp. 620–627.

[11] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in Proc. Int. Conf. Social Media Soc., 2015,p. 9.

[12] B. Viswanath et al., "Towards detecting anomalous user behavior in online social networks," in Proc. Usenix Secur., vol. 14. 2014, pp. 223–238.

[13] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on Twitter: Human, bot, or cyborg?" in Proc. 26th Annu. Comput. Secur. Appl.Conf., 2010, pp. 21–30.

[14] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks," in Proc. NDSS, 2013,pp. 1–17.



[15] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in Proc. Int. Conf. Social Inform., 2016, pp. 22–39.VOLUME 6, 2018 6547

E. van der Walt, J. Eloff: Using Machine Learning to Detect Fake Identities: Bots versus Humans [16] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake Twitter followers," Decision Support Syst., vol. 80, pp. 56–71, Dec. 2015.

[17] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in Proc. Collaboration, Electron. Messaging, AntiAbuse Spam Conf. (CEAS), vol. 6. 2010, p. 12.

[18] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in Proc. 19th Int. Conf. World Wide Web, 2010,pp. 591–600.

[19] Z. Zhang and B. B. Gupta, "Social media security and trustworthiness: Overview and new direction," Future Generat. Comput. Syst., to be published, doi: 10.1016/j.future.2016.10.007.

[20] A. K. Jain and B. B. Gupta, "Phishing detection: Analysis of visual similarity based approaches," Secur. Commun. Netw., vol. 2017, Jan. 2017, Art. no. 5421046. [Online]. Available: https://doi.org/10.1155/2017/5421046

[21] R. J. Oentaryo, A. Murdopo, P. K. Prasetyo, and E.-P. Lim, "On profiling bots in social media," in Proc. Int. Conf. Social Inform., 2016, pp. 92–109.

[22] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," IEEE Trans. Inf. Forensics Security, vol. 9, no. 8, pp. 1311–1321, Aug. 2014.

[23] J. T. Hancock, "Digital deception," in Oxford Handbook of Internet Psychology. London, U.K.: Oxford Univ. Press, 2007, pp. 289–301.