Machine Learning: A Review on Binary Classification

Roshan Kumari
M.Tech. Scholar
Department Of Computer Science & Engineering
ABES Engineering College, Ghaziabad

Saurabh Kr. Srivastava
Sr. Asst. Professor
Department Of Computer Science & Engineering
ABES Engineering College, Ghaziabad

ABSTRACT
In the field of information extraction and retrieval, binary classification is the process of classifying given document/account on the basis of predefined classes. Sockpuppet detection is based on binary, in which given accounts are detected either sockpuppet or non-sockpuppet. Sockpuppets has become significant issues, in which one can have fake identity for some specific purpose or malicious use. Text categorization is also performed with binary classification. This research synthesizes binary classification in which various approaches for binary classification are discussed.

Keywords
Sockpuppets, Non sockpuppets, multiple identity deception, text categorization, NB, SVM, Random Forest, Ensemble methods and Binary Classification

1. INTRODUCTION
Sockpuppets are some fake IDs or accounts, which are created for some specific malicious use. So sockpuppet detection is based on binary classification, in which classification is done on each account to assign a class i.e., sockpuppet or non-sockpuppet. Anyone can have new account with the help of less information. So it's necessary to have some method to find out these sockpuppet cases or suspicious cases because its violates privacy. Wikipedia does not provide any specific facility to detect such malicious accounts. So the current process are done manually which is time consuming and cost effective. So identify such accounts as sockpuppet in Wikipedia, is significant issue. Multiple identity is also an example of binary classification, in which one person has more than one account for malicious use. With the help of multiple account, one can try to alter senders contents for some specific purpose. So multiple identity deception is also a big issues on social media. Text categorization is also done by performing binary classification. Text categorization plays an important role on information retrieval for classification of different documents.

Sockpuppet detection becomes a significant problem in social media environment. Thamar Solorio et.al.[1] has contributed his work towards sockpuppet detection. They have done their work with small case study of automated detection of sockpuppets based on Authorship Attributes where the task consists of analyzing a written document to predict the true author. Some features of authorship attribution are collected and examined. Each user comment is a "document". There are two steps taken for classification process, in initial step, predictions from the classifier on each comment has taken. Then in second step, predictions for each comments and combine them in majority voting schema to assign final decisions to each account. Michail Tsikerdekis et.al. [2] has proposed a novel approach for use of non- verbal behavior to detect multiple account identity deception on social media. New accounts initiated by blocked users are called sockpuppetry. In social media, identity deception is a major issue. Nonverbal communication(user activity or movement) are more powerful than Verbal communication(speech or text). Identity deception focuses on manipulating the senders information and is divided in three categories-identity concealment, identity theft and identity forgery. Major issue with identity deception in social media is the presence of multiple identities by one user. They have used Logs of blocked users on Wikipedia during the period since February 2004 until October 2013 as dataset and used SVM, Random forest and Adaptive Boosting(ADA) method for classification of sockpuppetry or not. They found that Adaptive Boosting provides the best balance between Recall and Precision, and achieved highest Accuracy among all used classification techniques.

2. REVIEWED PAPERS IN THIS DIRECTION
Thamar Solorio et.al.[3] has described a corpus of sockpuppet cases from Wikipedia. A Corpus provides a real world dataset of short messages from malicious users. Sockpuppet investigations in Wikipedia(SPI) are identified using support vector machine. Author has tried to detect SPI, to decide whether to mentioned the editors as belonging to the same person or not on the basis of binary classification. These results are based on observations from comments made by each user. Used complete list of features can be found at the following link:http://docsig.cis.uab.edu/media/2014/03/list-of-features.pdf. Xueling Zheng et.al.[4] has proposed algorithm for detecting sockpuppet pair in one forum and two different forum. Two different online accounts but belong to the same person are referred as sockpuppets pairs. In this paper there are two methods proposed for detecting sockpuppets. The first one is designed for detecting those sock puppets pairs in the same discussion forum while the second one is for detecting sockpuppets pairs that appear in two different forum. Authors has used dataset from Uwants and HK discuss during the period of March 2010 to May 2010 . On the basis of Detection Score, they have tried to find out similar keywords used by different people. Sadia Afroz et.al.[5] Author has proposed a method to detect stylistic deception in written document. It is mentioned in this paper that with the help of large feature set, it is possible to distinguish regular documents from deceptive documents. To detect adversarial writing, it is necessary to identify a set of discriminating features that distinguish deceptive writing from regular writing. After determining these features supervised learning techniques are used to classify new writing styles. Three feature sets are used: Write print feature set, Lying-detection feature set and 9-feature set(Authorship attribution features). SVM and DT techniques are used for analysis. SVM classifier works best with the write prints feature and DT performed well with the Lying

detection features. Dhanyasree P et.al.[6] has contributed their work for detection of identity deception on social networking sites. On social networking sites, one person creates multiple account for malicious use. So this become a very big issue on social sites. So on the basis of verbal and non verbal behavior it can be detect such types of account. So authors has tried to detect such accounts on the basis of verbal and non verbal behavior. They have used algorithms, Calculation of non-verbal variables and model testing using Random Forest method and Identification of time window using PSO. They found that Detecting multiple accounts through nonverbal behavior has more accuracy. The automated system to detect multiple accounts gives good performance . Both the verbal and nonverbal behavior can be combined and used for sockpuppets detection, in which binary classification are done to detect sockpuppet or non sockpuppet cases. M BalaaNand et.al. [7] has proposed a method to detect multiple account and fake identity on social media like WIKIPEDIA using non-verbal behavior(User activity and User Movement). Authors has worked for time independent based non verbal behavior. In which they has used data from Wikipedia and SVM,RF and ADA techniques for binary classification. They found Adaptive Boosting gives the best balance between recall and precision with high accuracy. Sheetal Antony et.al.[8] has proposed a system that can use verbal and non verbal behavioral patterns to detect identity deception. There is an Admin who manages each account for users. The details and activities of the user are analyzed and detect if there is some deception. The details are verified in database. If it detects that there is some deception then there are some security questions that are asked to users. Zaher Yamak et.al.[9] has proposed a detection method in which following steps are taken: first of all data are crawled from Wikipedia, then detect sockpuppet accounts, after that create a set of non-verbal behavior features and then calculate the values of the proposed features and finally used machine learning algorithm for classification. SVM, RF, Naive Bayes, K nearest neighbor, Bayesian Network and Adaptive Boosting are taken for result comparison. Best accuracy given by Random Forest(99.8%) and Bayesian Network(99.6%) for sockpuppet detection. Malware detection is an important issues to save our computer system and communication infrastructure. So, Anti-virus technology is a key player in tackling malware files, based on two methods: signature based and heuristic-based method. Asaf Shabtai et.al.[10] has addressed different challenges i.e., files representation method, feature selection method and classification algorithm. Some additional issues are also mentioned in this paper such as: weighting algorithm(ensembles),imbalance problem ,active learning and chronological evaluation. Authors has proposed a framework for detecting new malicious code in executable files can be designed to achieve very high accuracy while maintaining low false positives. Antu Mary et.al.[11] has proposed a method for detecting identity deception by a single user is based on using Nonverbal behavior. Non verbal behavior explains activities done by each user separately such as Some Wikipedia users create multiple accounts and use them for various malicious purposes such as Number of articles generates, Number of searches done for same articles, Number of bytes added and also removed, Number of times same spelling mistakes carryout constantly, Time taken between each revision, creating fraudulent articles, damaging existing article text etc. So these deceptions cannot easily detected by any authority. Numerous methods have been proposed that can help in detecting multiple accounts owned by the same persons. Using verbal and nonverbal behavior of user can easily detect the sockpuppet with limited amount of time. Ashkan Sami et.al.[12] has provide a framework for analyzing and classifying PE files based on data mining techniques. Windows Application programming interface(API) can be used to extract knowledge describing behavior of executables .Each API call is used as a feature. FISHER SCORE based feature selection process is used. Top 4 categories by Fisher's Score are :File Management, Process and Thread, Console and Registry. 34820 PE files where 31,869 were malicious and 2951 were benign windows PE files. RF,NB and DT techniques are used. Random Forest gives good performance. G.Ganesh Sundarkumar et.al. [13] has done text mining for feature selection. Then Mutual Information is used to extract most influential features. Then data mining models such as Decision tree, Neural network model, SVM , Probabilistic neural network and group method of data handling(GMDH) is used. On the basis of Accuracy, Sensitivity. Specificity all 59 models are compared. DT, SVM, PNN, NN and GMDH techniques are used for comparison. Then again the dataset are balanced using Oversampling and again tested the model .After balancing sensitivity/accuracy improved. Prasha Shrestha et.al.[14] has explained Malware Family Identification process using string information. Classification of malware into correct family is an important task for antivirus vendor. Using term-frequency and inverse document frequency(tf-idf) and using prominent strings extraction classification work are done in this paper. To check accuracy-way vendor agreement are compared with accuracy achieved by used algorithm or techniques. Exact match: Global vocabulary, exact matches: Prominent strings, Prominent strings set and Absence of prominent string are techniques used for this purpose. Data are used from University's malware database(1504 malware files). On the basis of above mentioned experiments it can be easy to detect malware family files. Exact Match: Global vocabulary gives the best result. Michael Bailey et.al.[15] has explained that anti-virus is incomplete in that it fails to detect or provide labels of the malware samples. Authors explained that when these systems do provide labels, theses labels do not have consistent meaning across families and variants within a single naming convention as well as across multiple vendors. Finally they demonstrated that these system lack conciseness in that they provide some little information or sometime too much information about a specific piece of malware. Authors has proposed a novel technique to overcome these problems. On the basis of behavioral fingerprints of malware's activity, automated malware classification are done. To compare and combine these fingerprints, single-linkage hierarchical clustering approach are applied. Gaston L'Huillier et.al.[16] has explained phishing mail classification. Phishing email fraud is to attempt to gain personal/sensitive information such as username, passwords and credit cards details. Algorithm like Support vector machines, naïve Bayes, Random forest algorithm are used for classification of phishing emails. The classification of phishing emails is extension of text mining. In this paper feature extraction methodology for fishing emails are enhances by using latent semantic analysis features and keyword extraction techniques. SVMs ,the naïve Bayes model and the logistic regression method are used in Weka tool to improve accuracy. Rafiqul Islam et.al.[17] has tried to classify malware on the use of static and dynamic features. There are some drawback in static techniques for malware classification. So it focuses to detect some dynamic features which is very useful in classification process. For static features there are two information needed: function length frequency and printable strings information. For dynamic features API functions name are used. SVM,DT,RF and Naive Bayes techniques are used in WEKA tools with 10 fold cross

validation for classification. Random forest gives the highest accuracy with TP,FN and Accuracy parameters. . Ali Danesh et.al. [18] proposed a classifier fusion method to improve text classification. Proposed approach combined Naive Bayes, K-NN and Rocchio methods by Voting algorithms methods and achieve a better classification rate which experimental results shows that the classification error decreases by 15%. 2000 documents from 20 different newsgroups has taken for experiment. Aytug Onan et.al.[19] proposed ensemble approach such as Adaboost, Bagging, Dagging, Random Subspaces and majority Voting. Two way ANOVA test conducted. The experimental analysis shows that the bagging ensemble of Random Forest with the most frequent based keyword extraction method yields promising results for text classification. The experimental result shows that the utilization of keyword based representation of text documents in conjunction with ensemble learning can enhance the predictive performance and scalability of text classification schemes. Baoxun Xu et.al.[20] has proposed an improved Random forest classifier for text categorization. They proposed improved random forest methods with both feature weighting and tree selection methods(WTRF), Breiman's Random forest (BRF) and the random forest with only tree selection method(TRF).Comparisons are based on accuracy and F-measures. All these three improved random forest methods are compared with other widely used text categorization methods i.e., support vector machines(SVM),Naive Bayesian(NB),and K-NearestNeighbor(KNN). M.Sivakumar et.al. [21] proposed a hybrid text classification Approach using KNN and SVM. They proposed SVM-KNN approach aims to reduce the impact of parameters in classification accuracy. The performance analysis shows the accuracy of SVM-KNN method remains optimal for even huge values of the parameters. The accuracy compared to the KNN method is higher in the SVM-KNN. Unlike the conventional KNN classification approach, the SVM-KNN approaches has low impact on the implementation of the parameters. Sundus Hassan et.al. [22] proposed a method for text categorization in which they compared Support Vector Machine(SVM) and Naive Bayes (NB) classifiers. Baseline for the experiment has setup by removing stopwords and stemmed the dataset by using Porter Stemmer. They used micro-average and macro average F-Measure. Experiments shows the improvement in micro average and macro average F-measure in both method i.e., SVM and NB.

Table 1: Summary of related work on sockpuppets detection, multiple identity deception detection and text categorization by performing binary classification

Citation	Dataset used	Classifiers	Measures	Results
Thamar Solorio et.al.[2013]	Data collected from Wikipedia	Support Vector Machine(SVM)	Precision, Recall ,F-Measure and accuracy	On the basis of authorship attributes sockpuppet cases are detected.
Michail Tsikerdekis et.al[2014].	Logs of blocked users on Wikipedia during the period since February 2004 until October 2013	SVM, Random forest and Adaptive Boosting(ADA)	Precision, Recall, Accuracy, F-Measures, false positive rate and Matthews Correlation Coefficient (MCC)	Higher accuracy Rate.
Thamar Solorio et.al[2014]	Wikipedia	Support Vector Machine in Weka tools	Precision, Recall ,F-Measure and Accuracy	F-Measure gives the best result.
Xueling Zheng et.al[2011]	Uwants and HK discuss during the period of March 2010 to May 2010	Detection score	On the basis of similarity and keywords	Sockpuppet pair are detected.
Sadia Afroz et.al[2011].	(1)Extended -brennan-Greenstadt corpus (2)Hemingway-Faulkner Imitation corpus (3)Thomas-Amina Hoax corpus	SVM and DT	Precision, Recall and F-Measure	SVM classifier works best with the Write prints feature and DT performed well with the Lying detection features.
M BalaaNand. et.al[2015]	Dataset used from Wikipedia	SVM, RF and ADA.	Precision, Recall, F-Measure, Accuracy, MCC and False Positive rate.	Adaptive Boosting gives the best balance between recall and precision with high accuracy
Sheetal Antony. et.al[2016]	Not Used	Not Used	Not Used	On the basis of verbal and non verbal behavioral patterns, detected identity deception.
Zaher Yamak et.al[2016]	Dataset used from Wikipedia from Feb2004 to April2015	SVM, RF, Naive Baiyes, K nearest neighbor, Bayesian Network and Adaptive Boosting	TPR, FPR, F-Measure, Precision and MCC	Best accuracy given by Random Forest(99.8%) and Bayesian Network(99.6%).
Asaf Shabtai. et.al[2009]	Not Used	ANN,DT,KNN,BN,SVM ,OneR,Boosted Algorithm	TPR, FP and, Accuracy	This paper includes aspects of different challenges for classifying new malicious code based on static features extracted.

Antu Mary et.al[2015]	Wikipedia	SVM	Recall, Precision and F-Measure	Identity deception detection is more accurate using non verbal behavior in comparison to verbal behavior.
Ashkan Sami et.al[2010]	34820 PE files where 31,869 were malicious and 2951 were benign windows PE files.	RF, NB and DT	Accuracy, Precision, Recall and F Arate(False alarm rate)	Random Forest gives good performance.

Table 1: Continued...

Citation	Dataset used	Classifiers	Measures	Results
G.Ganesh Sundarkumar et al.[2013]	Dataset from CSMINING group is used	DT,SVM,PNN,NN and GMDH	Accuracy ,Sensitivity and Specificity	On the basis of Accuracy ,Sensitivity and Specificity all 5 models are compared. Then again the dataset are balanced using Oversampling and again testing the model. After balancing sensitivity/accuracy improved.
Prasha Shrestha et al.[2014]	Data are used from University's malware database(1504 malware files).	(1)Exact match: Global vocabulary (2) exact matches (3) Prominent strings set (4)Absence of prominent string	Accuracy and correlation between n-way vendor agreement with accuracy	Exact match: Global vocabulary gives the best accuracy around 91.02%.
Michael Bailey et al.[2007]	Data collected from data sources .There are 3 types of dataset used: Legacy, small and large.	Hierarchical clustering algorithm	Consistency, Completeness and conciseness	It's easy to classification of malware on the basis of behavior of fingerprints.
Gaston L'Huillier et al.[2013]	Not Used	SVMs ,the naïve Bayes model and the logistic regression method were used in Weka tool	F-Measure	Accuracy improved
Rafiqul Islam et al.[2013]	Data are used from antivirus vendors time periods 2003-2007 and 2009-2010.	SVM,DT,RF and Naive Bayes in WEKA tool.	TP,FN and Accuracy	Random forest gives the highest accuracy.
Ali Danesh et al.[2007]	100 articles are taken from 20 different newsgroup	NB, KNN, Rocchio ,Voting and OWA AND DT	Accuracy Rate	Fusion of classifiers gives better results in comparison to base classifiers.
Aytug Onan et al.[2016]	Reuters-21578 dataset used	NB, SVM, LR, RF and ensemble methods of SVM and RF.	F-Measure, Accuracy and AUC Values	This paper represents analysis of 5 statistical keywords methods for text classification.
Baoxun Xu et al.[2012]	20 different Usenet newsgroups and contains 18772 documents divided into 20 different classes	SVM,KNN,NB,BRF,TRF and WTRF	Micro F-Measures and Macro F-Measures	It has been observed that proposed method WTRF method outperforms among all other text categorization methods
M.Sivakumar et al.[2014]	Reuters-21578 R8 dataset used.	KNN and SVM	Accuracy	Proposed SVM-KNN method provides high accuracy.
Sundus Hassan et al.[2000]	Dataset from 20 Newsgroup with 1000 documents	NB and SVM	Macro F-Measures and Micro F-Measures	NB gives better performance over SVM.

3. CONCLUSION AND FUTURE WORK

This paper presented a taxonomy for binary classification. Sockpuppet detection is based on binary classification, in which two classes are predefined i.e. sockpuppet or non-sockpuppet. And datasets are classified on the basis of predefined classes. Multiple identity deception is also based on binary classification in which classification process are done on given datasets into two groups i.e., sockpuppet and non-sockpuppet. Text categorization is also done by involving binary classification. So binary classification implies an important role in machine learning process. To get better result, analyze different features sets for binary classification. With different feature sets, better results can be observed in terms of precision, recall, F-Measure and accuracy. Different datasets can be used for experiment with different text features. These feature sets can be used for multilevel classification and multiclass classification.

4. REFERENCES

[1] Thamar Solorio, Ragib Hasan and Mainul Mizan, "A Case Study of Sockpuppet Detection in Wikipedia", Proceedings of the Workshop on Language in Social Media(LASM 2013),Pages 59-68,Atlanta,Georgia,June 13 2013.@2013 Association for Computational Linguistics**.**

[2] Michail Tsikerdekis and Sherali Zeadally, "Multiple Account Identity Deception Detection in Social Media Using Non Verbal Behavior", IEEE Transactions on Information Forensics and Security, Vol 9, No 8, August 2014.

[3] Thamar Solorio, Ragib Hasan and Mainul Mizan, "Sockpuppet Detection in Wikipedia :A Corpus of Real-World Deceptive Writing For Linking Writing", arXiv:1310.6772v1[cs.CL] 24 Oct 2013.

[4] Xueling Zheng, Yiu Ming Lai, K.P. Chow, Lucas C.K. Hui and S.M. Yiu, "Detection of Sockpuppets in Online Discussion Forums", HKU CS Tech Report TR-2011-03.

[5] Sadia Afroz, Michael Brennan and Rachel Greenstadt, "Detecting Hoaxes Frauds and Deception in Writing Style Online". 2011.

[6] Dhanyasree P*, Sajitha Krishnan and Ambikadevi Amma T, "Deception Detection in Social Media through Combined Verbal and Non-Verbal Behavior ", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 5, Issue 4, 2015.

[7] M Balaanand,R Soumipriya,S Sivaranjani and S Sankari, "Identifying Fake Users in Social Networks Using Non-Verbal Behaviour". International Journal of Technology and Engineering System (IJTES)Vol 7. No.2 2015 Pp. 157-161©gopalax Journals, Singapore.

[8] Sheetal Antony, Prof. B. S. Umashankar, "Identity Deception Detection and Security in Social Medium, IJCSMC, Vol. 5, Issue 4, April 2016, pg.499-502.

[9] Zaher Yamak, Julien Saunier and Laurent Vercouter, " Detection of Multiple Identity Manipulation in Collaborative Projects", IW3C2, WWW'16 Companion, April 11-15, 2016, Montreal, Quebec, Canada. ACM 978-1-4503-4144-8/04.

[10] Asaf Shabtai, Robert Moskovitch, Yuval Elovici and Chanan Glezer, " Detection of malicious code by applying machine learning classifiers on static features: A state -of-the-art-survey ", INFORMATION SECURITY TECHNICAL REPORT 14 (2009) 16-29, ELSEVIER.

[11] Antu Mary Kuruvilla1 and Saira Varghese2, "A Survey on detecting Identity Deception in Social Media Applications", International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 02, Issue 04, [April – 2015] ISSN (Online):2349–9745 ; ISSN (Print):2393-8161.

[12] Ashkan Sami, B. Yadegari, N. Peiravian, and S. Hashemi and A. Hamze, "Malware detection based on mining API calls", SAC '10: Proceedings of the ACM Symposium on Applied Computing, pp. 1020-1025, 2010.

[13] G.Ganesh Sundarkumar and Vadlamani Ravi, "Malware Detection by Text and Data Mining".IEEE2013..

[14] Prasha Shrestha,Suraj Maharajan,Gabriela Ramirez de la Rosa,Alan Sprague,Thamar Solorio and Gracy Warner, "Using String Information for Malware Family Identification" @Springer International Publishing Switzerland 2014,A.L.C.Bazzan and K.Pichara(Eds.):IBERAMIA 2014,LNAI 8864,pp.686-697,2014.DOI:10.1007/978-3-319-12027-0_55

[15] Michael Bailey, Jon Oberheide, Z. Morley Mao, Farnam Jahanian and Jose Nazario, " Automated Classification and Analysis of Internet Malware". April 26 2007

[16] Gaston L'Huillier, Alejandro Hevia, Richard Weber and Sebastian Rios, "Latent Semantic Analysis and Keyword Extraction for Phishing Classification".IEEE2010.

[17] Rafiqul Islam, Ronghua Tian , Lynn M. Batten and Steve Versteeg," Classification of malware based on integrated static and dynamic features". Journal of Network and Computer Applications 36 (2013) 646–656. ELSEVIER.

[18] Ali Danesh, Behzad Moshiri and Omid Fatemi, "Improve Text Classification Accuracy based on Classifier Fusion Methods".2007 IEEE Xplore.

[19] Aytuğ Onana, Serdar Korukoğlub and Hasan Bulutb, " Ensemble of keyword extraction methods and classifiers in text classification". A. Onan et al. / Expert Systems With Applications 57 (2016) 232–247.

[20] Baoxun Xu, Xiufeng Guo, Yumming Ye and Jiefeng Cheng, "An Improved Random Forest Classifier for Text Categorization", [JOURNAL OF COMPUTERS] VOL. 7, NO. 12, DECEMBER 2012.

[21] M. Sivakumar, C. Karthika and P. Renuga, "A Hybrid Text Classification Approach using KNN and SVM", [IJIRSET] Volume 3, Special Issue 3, March 2014.

[22] Sundus Hassan, Muhammad Rafi and Muhammad Shahid Shaikh, "Comparing SVM and Naive Classifiers for Text categorization with Wikitology as knowledge enrichment". IEEE Xplore 2012.