Survey Paper : Sock Puppet Detection

Abhijnan Chakraborty[1] **Vaibhav Saha**[2] **Harshit Kumar**[3] **Rishabh Verma**[4]

[1]Dept. of Computer Science [2]Dept. of Computer Science [3]Dept. of Computer Science [4]Dept. of Computer Science

abhijnan@iitd.ac.in, cs1200403@iitd.ac.in, cs1200347@iitd.ac.in, cs5200123@iitd.ac.in

Abstract -

The term Sock-Puppet refers to the fake or multiple accounts created for malicious use. It is observed on most of the Online social Networking platforms including Wikipedia, X (earlier known as Twitter) etc. Sock-Puppet detection is the identification of whether an account is a sock-puppet or not i.e. it is a binary classification problem. This paper is a research done to note down various methods and feature-sets applied and tested for Sock-Puppet detection and goes on further to talk about the future work in this field.

Keywords -

Sockpuppet, Online Social networking, Binary classification, SVM , Random forest, RTV, SPG, Naive Bayes, KNN, BertTokenizer, LDA model

1 Introduction

The pervasive growth of the internet and the prominence of online social networks have catalyzed an increase in deceptive practices, particularly the use of sockpuppet accounts, across various digital platforms. Sockpuppets, deceptive identities created by a single individual or entity, serve as a tool for a multitude of manipulative endeavors. These deceptive accounts have been observed in diverse scenarios, including but not limited to, influencing public opinion, manipulating online polls, spreading misinformation, generating biased reviews, and even for state-sponsored propaganda. Studies reveal their presence in online forums, collaborative projects like Wikipedia, social media platforms such as Twitter, and review systems utilized by electronic marketplaces. As their proliferation continues, detecting and mitigating the impact of sockpuppets has become a pressing challenge. Various detection methods, encompassing linguistic analysis, behavioral patterns, network structures, and propagation tree structures, have emerged aiming to distinguish these deceptive accounts from legitimate users. However, the dynamic nature of sockpuppet behaviors and evolving manipulation tactics necessitate a comprehensive survey to evaluate and integrate effective detection techniques across diverse digital landscapes.

2 Related Work

In the Paper by Thomas Solorio et al. in 2013, based on Wikipedia dataset the binary classification was done on text authorship identification features focused on comments and edits on talk pages using a total of 239 verbal features that capture stylistic, grammatical, and formatting preferences of the authors and two edit timing features. Such features included total number of characters, parenthesis count, emoticons count, two to three continuous punctuation count, frequency of letters etc. They used the Weka's implementation of SVM with default parameters and claimed to have acheived a F-measure of 0.72 with 84.04% confidence.

In 2017, Suman Kalyan Maity et al. and Srijan Kumar et al. published papers on Sock-puppet Detection. The paper by Suman Kumar Maity was focused around 2016 US presidential election for Twitter (now known as X) dataset capturing the most recent 3200 tweets of the candidates (Donald Trump and Hillary Clinton) and the retweets of their followers. In this paper the features recorded were of two broad categories i.e. tweet features and profile based features. Tweet features included entropy of tweets which was a probabilistic distribution on time taken between two consecutive tweets and normalized retweet count. Profile based features included profile verification status, location, account creation date,friends and followers count, Reputation score which is ration of followers among followers and friends etc. The paper claimed to have achieved a best of 0.68 F1-score with 90.98% accuracy using SVM classifier.

In the Paper by Srijan Kumar et al. the database was taken from Disqus, an online commenting platform that hosted the dicussions from nine different online communities that had a variety of topical interests - from news and politics to sports and entertainment. The feature set was of three types i.e Activity, Community, and Post Features. Activity features included clustering coefficient, reciprocity, number of posts etc. Community features had features like whether account is blocked fraction of up-votes etc. and Post Features included verbal features that were included earlier like number of characters, words, syllables etc. The community features were introduced here. They claimed to have achieved an AUC of 0.68 using all the features together. It also works on the tendency of sock puppets i.e. Sock puppets tend to participate in discussions with more controversial topics, Sock-puppets in a pair interact with each other more, Sock puppets are treated harshly by the community etc.

In 2022, Mostofa et al. published his paper based on Wikipedia dataset. However, they had curated the dataset themselves and divided the features broadly in two categories - Account and Content based. Account based features had verbal and profile based features that were discussed in earlier papers. Content based features included the feature of topic which was identified using two models BertTokenizer(unsupervised classification) of BERT model and LDA model (supervised classification) of Gensim library with 20 topics which detected the topic of user comment and post. Experimenting with the results, they found out that when they didn't use the LDA model the F1-score dropped by 0.1 that was the largest among all the features which pointed that content based classification was a good step though removing BERT model only gave a drop of 0.004. The paper claimed to have achieved a best of 0.82 F1-score using Random Forest and also as the depth of number of edits to which dataset was curated increased the measures tend to approach a constant higher value.

In the context of Sock puppet detection, Zaher Yamak et al.[7] has used some notable features which has contributed their work in the field of sock puppet detection. The dataset used by this paper consists of wikipedia dataset. Dataset extraction consists of 10 TB uncompressed which is compressed to 100 GB by 7-Zip. From this, 118414 sock puppet accounts got filtered. After this, in the account selection, 12088 groups were identified which contains 2-557 members. There is a random selection of 5000 sock puppets who are in a group of more than 3 sockpuppets. They are mixed with 5000 random active accounts. After this there is feature selection process, in the paper the author says that by using these features, the accuracy achieved is 99.8%. Some notable features are:-

- **The number of user's contributions by namespaces:** This system categorizes user contributions into six types, offering insights into their focus areas and communication patterns, aiding in understanding their impact on content and community dynamics.

- **The frequency of revert after each contribution in the articles:**

- **The average of bytes added and removed from each revision:**

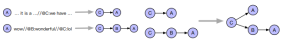


Figure 1. Propagation Tree Growth

- **The interval between the user's registration and his first contribution:** The assumption is that a manipulator creates several dormant accounts upfront, intending to use them as backups if an active account gets blocked during a manipulation attempt.

After this, they do a 10-fold cross-validation technique to assess different models. The algorithm and metrics used are TPR, FPR , F-measure, precision, and MCC. The experimental results are compared with some previous works one is Gao et. al who used the verbal attributes as features which got about 93.9% accuracy, after this Solorio paper which was based on textual features, alphabet count , number of tokens and use of words without vowels have accuracy of about 68.83% using SVM. Also, this is not so accurate if user changes his writinf style. Another paper of Yang that uses non verbal attributes are invitation frequency, outgoing request accept, incoming request are taken into account which gives an accuracy of about 71.3% using SVM classifier. The paper concludes that his technique exceeds many previous algorithms and plan to explore other social media to verify the feature set in other settings, such as forums or twitter.

Jiacheng et al.[11] discusses how its previously ignored a very crucial property present in the sockpuppet behavior, the propagation characteristics. The paper uses the property that the propagation tree of a sockpuppet is deeper and highlights that the message is reposted by sockpuppet will be spread far. The paper talks about the **Difference of pairwise accounts** which is used in the paper as a useful property, shows the sockpuppet pair is more similar than others through three dimensions: size, depth, and width. It is reasonable that the pairwise sockpuppets behave similarly.

The propagation tree is constructed by following as shown in the figure 1, Each reposting log will represents an information propagation process, such as ”wow//!B:wonderful//@C:lol”. Based on the practice of refereeing to another account in a tweet via ”//@username” convention , we extract the usernames from reposting log and construct the propagation trees to represent the information propagation process of an account.

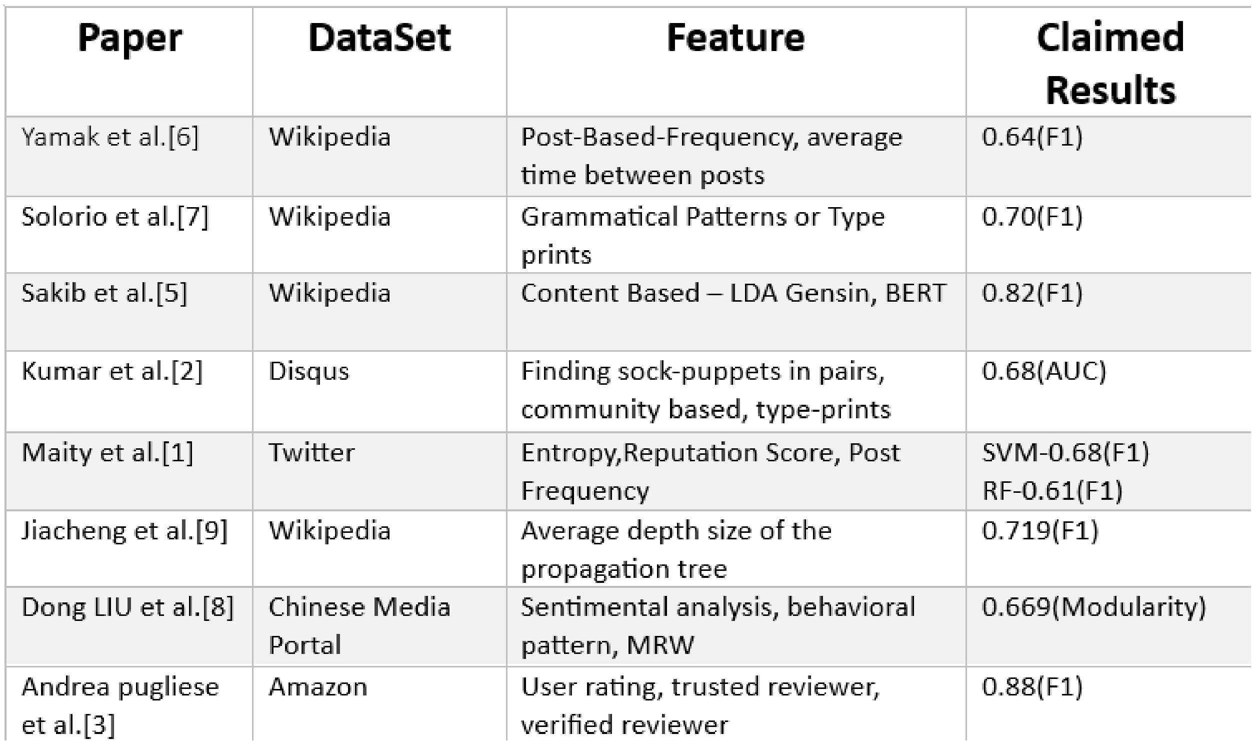
The paper deals the whole problem of sock puppet detection using the graph terminologies in which assume G = (V, E) be a social network, where V is a set of accounts, E $\in V \times V$ is a set of repost relationship, and $e^i_{vu} \in E$ denotes repost relationship of message i between account v and u(v, u $\in V$) which reflects propagation of information over G. In this suppose u represents a user so the sock puppet account detection goes through some notable features such as:-

- **Number of posts:** This is a typical feature that depicts the activity frequency of accounts in social network.
- **Average depth of propagation tree:** This presents the delay in the message i propagation of account u.
- **Average size of propagation tree:** This feature is trying to capture the coverage of message i which the account u is participated.
- **Average number of identical account in tree:** The goal of this feature is that the number of same nickname of type account u is to model the participation rate of similar accounts in a conversation.
- **Average interval between interactions:** This is a normalized feature where we compute the time difference between the t^{th} post p_t and the prior one p_{t-1}

The dataset for the experiment used are the real world data D$_s$ and D$_t$ which the author crawled tweets from 2017.01 to 2018.10. from Sina Weibo. For the comparison purpose the profile feature attributes, verbal attributes , non-verbal attributes and propagation feature is tested using 4 different Machine learning models Logistic Regression,Random Forest, Support Vector machine and Adaptive boosting(ADA) are used. The experimental analysis shows that LR with the propagation feature gave good results with a precision score of 0.840 and F1 score of 0.719.

In their publication, Dong LIU et al.[8] present a sock-puppet gang (SPG) detection algorithm. The authors discovered that sock puppets belonging to the same SPG typically have the same purpose and behavioural pattern, such as a similar writing style, sentimental orientation, and infrequent communication due to the fact that they assume various identities. The sentiment orientation of users towards discussion topics is examined by the authors through the analysis of emotive phrases present in their comments. Orientations can be categorised as neutral, negative, or positive. Orientations of each user with respect to various subjects are denoted by a sentiment orientation vector. Subsequently, an analogous orientation network (SON) is built in which edges between nodes represent sentimental similarity and nodes represent sock puppet accounts. A sockpuppet account in an SPG may not express its sentiment orientation as explicitly as other members do due to the fact that SPG members attempt to conceal their true intentions. To account for this, the SON structure is refined iteratively through the extraction of user behaviour features such as time intervals between comments. SPG accounts frequently exhibit coordinated surges of commentary. The SON undergoes numerous random walks in order to re-weight the edges in accordance with the traversal frequency. This enhances the correlation between accounts that exhibit comparable patterns of behaviour. After many iterations, edges between accounts that are likely to be in the same SPG are assigned a heavy weight, whereas edges that are unrelated begin to lose their weight. This feature enhances the community's organisation and distinguishes SPG members from fortuitous associations. In conclusion, the re-weighted SON incorporated pre-existing algorithms for detecting multiple communities in order to discern densely linked clusters that correspond to sockpuppet groups. The authors utilised a dataset of Chinese social media in order to detect SPG with precision.

The research conducted by Pugliese et al.[11] involved the development of the RTV framework, which was designed to detect sock puppet accounts involved in the fabrication of product ratings. It leverages the concept of "verified reviewers" (who purchased the product) and "trusted reviewers" (hired by the company to provide honest ratings). A bipartite graph is constructed to illustrate the relationship between products and consumers, with ratings acting as the weights. RTV then defines mathematical functions for fairness, goodness and reliability over this graph to model expected behavior. The values at each timestamp are modified iteratively by these functions in accordance with the preceding values, until they converge in order to detect users who deviate suspiciously. RTV effectively mitigates the problem of sock puppet attacks through the implementation of a system which includes verified reviewers and trusted reviewers. Based on the evaluation results, RTV demonstrates superior performance in 72 out of 84 cases when compared to the seven most recent review fraud detection systems. This conclusion is supported by testing conducted on multiple online e-commerce platforms, including Amazon.



Paper	DataSet	Feature	Claimed Results
Yamak et al.[6]	Wikipedia	Post-Based-Frequency, average time between posts	0.64(F1)
Solorio et al.[7]	Wikipedia	Grammatical Patterns or Type prints	0.70(F1)
Sakib et al.[5]	Wikipedia	Content Based – LDA Gensin, BERT	0.82(F1)
Kumar et al.[2]	Disqus	Finding sock-puppets in pairs, community based, type-prints	0.68(AUC)

Maity et al.[1]	Twitter	Entropy,Reputation Score, Post Frequency	SVM-0.68(F1)
			RF-0.61(F1)

Jiacheng et al.[9]	Wikipedia	Average depth size of the propagation tree	0.719(F1)
Dong LIU et al.[8]	Chinese Media Portal	Sentimental analysis, behavioral pattern, MRW	0.669(Modularity)
Andrea pugliese et al.[3]	Amazon	User rating, trusted reviewer, verified reviewer	0.88(F1)

3 Conclusion and Future Work

The collective insights from these papers highlight evolving sockpuppet detection methods across online platforms, emphasizing feature engineering, machine learning, and behavioral analysis. Innovative approaches like comment behavior features and propagation tree structures show promise in detecting deception on social media, collaborative projects, and more, prompting the need for improved techniques against evolving manipulative tactics.

Future work should focus on expanding feature sets for better precision, conducting cross-platform analyses to understand diverse deceptive behaviors, and bolstering algorithms against sophisticated attacks. Ethical considerations, such as user privacy and fraud prevention, need integration into future methodologies, alongside collaborative interdisciplinary research efforts for more effective sockpuppet detection strategies across online domains.

References

[1] Animesh Mukherjee Suman Kalyan Maity, Aishik Chakraborty. Detection of sockpuppets in social media. 2017.

[2] Jure Leskovec V.S. Subrahmanian Srijan Kumar, Justin Cheng. An army of me: Sockpuppets in online discussion communities. 2017.

[3] Quanyuan Wu Dong Liut. Homologous sockpuppet accounts detection based on comment behavior features. 2017.

[4] Francesca Spezzano Mostofa Najmus Sakib. Automated detection of sockpuppet accounts in wikipedia. 2022.

[5] J. Saunier Z. Yamak and L. Vercouter. Detection of multiple identity manipulation in collaborative project. Proceedings of the 25th International Conference Companion on World Wide Web, 2016.

[6] R. Hasan T. Solorio and M. Mizan. A case study of sockpuppet detection in wikipedia. Proceedings of the Workshop on Language Analysis in Social Media at NAACL HTL, 2013, pp. 59–68.

[7] Weihong HAN Bin ZHOU Dong LIU, Quanyuan WU. Sockpuppet gang detection on social media sites.

[8] Jizhong Han Jiacheng Li, Wei Zhou and Songlin Hu. Sockpuppet detection in social network via propagation tree. 2019.

[9] S. Zeadally M. Tsikerdekis. Multiple account identity deception detection in social media using nonverbal behavior. Inf. Forensics Security IEEE Trans. 9 (8)(2014) 1311–1321.

[10] Rui Liu and V. S. Subrahmanian Runze Liu, Andrea Pugliese. Stars: Defending against sockpuppet-based targeted attacks on reviewing systems. Received June 2019; revised March 2020; accepted April 2020.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10]