Wiki Workshop (10th edition) – May 11, 2023

Automated Detection of Sockpuppet Accounts in Wikipedia

Mostofa Najmus Sakib
Department of Computer Science
Boise State University, Boise – USA
mostofanajmussak@u.boisestate.edu

Francesca Spezzano
Department of Computer Science
Boise State University, Boise – USA
francescaspezzano@boisestate.edu

Abstract

In Wikipedia, sockpuppetry, also known as multiple identity generation, is a prevalent problem that is typically used for deception. With such significance in mind, we have discussed sockpuppet account detection for this Internet-based encyclopedia. To be able to distinguish sockpuppets from benign users, we structure the issue as a binary classification challenge and provide a set of features based on user activity and the semantics of their contributions. Throughout the process, we have developed a dataset of 17K sockpuppet accounts from Wikipedia and inaugurated an ML-based system capable of detecting sockpuppet accounts at an initial stage. Our results are promising as it has overshadowed prior works in terms of F1-score measures.

Keywords: Sockpuppetry, Malicious Activity Detection, Early Prediction, User Behavior Analysis, Language Models.

Introduction

Since its journey in 2001, Wikipedia has provided free access to its enriched content all over the globe. Users can communicate and share knowledge on Wikipedia with great ease. It has a unique open-source management style that allows the voluntary participation of users to create articles and documents. Such a process can minimize the cost but at the same time include vulnerability in terms of malicious activity, including vandalism, spam, undisclosed paid editing, etc. (Kumar et al., 2015; Green and Spezzano, 2017; Joshi et al., 2020) as almost anyone can collaborate or contribute with minimal internet access.

As formally defined in Wikipedia, sockpuppet accounts are accounts whose behavior is dictated by another person or account and used for deception. [1] Taking advantage of a less strict provision of the one-user one-account policy of Wikipedia, very often multiple accounts are operated by one single account then the account is referred to as puppetmaster. The current practice of Wikipedia to prevent such heinous activity is mostly done by manual inspection. Prior research of sockpuppet detection (Solorio et al., 2013; Solorio et al., 2014) from faithful single or multiple accounts have concentrated on the stylistic, syntactic, and social network elements primarily through cross-checking the similarities of different account holders. To counter the sockpuppetry issue from a more holistic view, we have included content-based features specifically the semantic meaning of each user's contribution for the first time. Our suggested method can identify sockpuppet accounts with an F1-score of 0.82 when taking into account the user's first 20 edits and 0.73 when only taking into account the first edit.

Methods

Our approach evolved around the idea of creating a method to automatically recognize and classify suspicious updates made by the same author from several accounts. To start the process we gathered and examined sockpuppetry data using MediaWiki Action API. [2] Specifically we have collected around 17,180 sockpuppet accounts [3] from Wikipedia by aggregating all the users under the "Suspected Wikipedia sockpuppets" category. [4] To counterpart the sockpuppet user accounts in our binary classification system, we have also gathered contributions through the same API for benign user accounts described in Kumar et al. (Kumar et al., 2015). We had 16,496 benign user accounts. For both types of accounts, we gathered the first 20 edits, and the corresponding page id, the parent page id, the page names-pace (article, article discussion, user page, etc.), the page title, the edit timestamp, the text of the user contribution, and the size of the user contribution for each of them.

From this dataset, we computed two types of feature sets, i.e., account-based and content-based. We calculated account-based information such as (i) the number of digits in a username, (ii) the ratio of digits in a username, (iii) the number of leading digits in a username, and (iv) the unique character ratio in a username for each user account. Alongside this, we have also included the average contribution length, average title length, and average

[1] https://en.wikipedia.org/wiki/Sockpuppet_(Internet)

[2] https://www.mediawiki.org/wiki/API:Main_page

[3] Our dataset is available for download at https://github.com/Mostofa-Najmus-Sakib/Wikipedia-Sockpuppetry

[4] https://en.wikipedia.org/wiki/Category:Suspected_Wikipedia_sockpuppets

time difference between two consecutive edits. Sockpuppet users, however, typically remark on a predetermined type of posts where they would gain the most. Exploring the semantics of user contributions through topic modeling and BERT embedding, we have incorporated this intuition into our work. We have used the state-of-the-art BERT transfer model (Devlin et al., 2018) to generate embedding for each user's contribution. Our goal is to tie down the similar contribution from multiple accounts as BERT takes advantage of the attention mechanism to understand the contextual relationship. Sockpuppet users frequently upload similar content, even after being deleted before. Pinpointing the contents' topic can contribute significantly to detecting multiple identities. To support this idea, we generated 20 topics on all the users' comments with an LDA model. Finally, we assigned to each comment the vector with the corresponding topic distribution and used those as features.

With the above-aforementioned methodology, we have developed models with classifiers like Logistic Regression, Gaussian Naive Bayes, Decision Tree, Multilayer Perceptron (MLP) Classifier, Random Forest, and a Long short-term memory (LSTM). We used five-fold cross-validation and the F1-score for the evaluation protocol. Except for LSTM, all the models had input from username-based features, average vector of user contribution's BERT embeddings, and topics. To appraise the sequential nature of this problem, we have developed an LSTM model with all the features previously used in the other models but for each contribution. Finally, we added the username-based features at the last cell of the LSTM before passing for classification.

Results

The Random Forest model had the best F1-score of 0.82, which makes our findings remarkable. The Table 1 displays the performance data for each model. To evaluate the quality of our work we have compared our results with the findings from prior researchers. Specifically, we have included the proposed features listed in (Yamak et al., 2016) and (Solorio et al., 2013). In addition, we have also integrated the Objective Revision Evaluation Service (ORES) which is a web service developed by Wikimedia Foundation that provides a machine learning-based scoring system for edits. We extracted ORES scores from the ORES publicly available API [5]. Since Random Forest (RF) had the highest score with our features, we developed an RF model in addition to the LSTM model to reattach the sequential nature with the features described in the comparison work. Table 2 displays the F1-scores of the competitors with our approach as we have a higher F1-score (0.82 in comparison to 0.54, 0.64, and 0.77) excelling all of them. Further to this, we studied the early detection of sockpuppet accounts shown in figure 1 with 1 to 20 edits. We can achieve an early detection with an F1-score of 0.73 considering only 1 edit and it was consistently better than the competitors while the number of edits was varied.

[5] https://ores.wikimedia.org

Discussion/Conclusions

The challenge of automatically identifying sockpuppet accounts on Wikipedia is discussed in this study. Our developed strategy has been effective two-fold. We can detect sockpuppetry issues at a better consistency as well as early detection is viable. Understanding the deep inherited semantic meaning is particularly important in evaluating sockpuppetry as it helps to join similar contributions from multiple accounts thus better highlighting benign accounts from their counterparts.

References

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Green and Spezzano2017] Thomas Green and Francesca Spezzano. 2017. Spam users identification in wikipedia via editing behavior. In *Proceedings of the Eleventh International Conference on Web and Social Media, 2017*, pages 532–535. AAAI Press.

[Joshi et al.2020] Nikesh Joshi, Francesca Spezzano, Mayson Green, and Elijah Hill. 2020. Detecting undisclosed paid editing in wikipedia. In *Proceedings of The Web Conference 2020*, pages 2899–2905.

[Kumar et al.2015] Srijan Kumar, Francesca Spezzano, and V. S. Subrahmanian. 2015. VEWS: A wikipedia vandal early warning system. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015*, pages 607–616. ACM.

[Solorio et al.2013] Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2013. A case study of sockpuppet detection in wikipedia. In *Proceedings of the Workshop on Language Analysis in Social Media at NAACL HTL*, pages 59–68.

[Solorio et al.2014] Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2014. Sockpuppet detection in Wikipedia: A corpus of real-world deceptive writing for linking identities. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1355–1358, May.

[Yamak et al.2016] Zaher Yamak, Julien Saunier, and Laurent Vercouter. 2016. Detection of multiple identity manipulation in collaborative projects. *Proceedings of the 25th International Conference Companion on World Wide Web*.

Classifier	**F1-score**
Logistic Regression	0.75
Gaussian NB	0.60
Decision Tree	0.75
MLP Classifier	0.77
Random Forest	**0.82**
LSTM	0.75

Table 1: F1-score comparison of different machine learning model with our proposed features in input to predict sockpuppet accounts. Best scores are in bold.

Experimentation	**F1-score**
Our proposed features with RF	**0.82**
Our proposed features with LSTM	0.75
ORES with RF	0.54
ORES with LSTM	0.53
Yamak et al. (Yamak et al., 2016) with RF	0.64
Yamak et al. (Yamak et al., 2016) with LSTM	0.59
Solorio et al. (Solorio et al., 2013) with RF	0.75
Solorio et al. (Solorio et al., 2013) with LSTM	0.77

Table 2: F1-score comparison of our proposed features vs. related work. We compare features in input to Random Forest (RF – which results the best classical machine learning algorithm) and LSTM. Best scores are in bold.

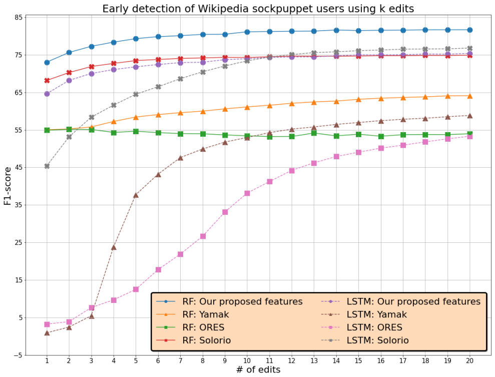

Figure 1: Early detection of Wikipedia sockpuppet accounts.