# Detecting Undisclosed Paid Editing in Wikipedia

Nikesh Joshi, Francesca Spezzano, Mayson Green, and Elijah Hill
Computer Science Department, Boise State University
Boise, Idaho, USA
francescaspezzano@boisestate.edu
{nikeshjoshi,maysongreen,elijahhill}@u.boisestate.edu

## ABSTRACT

Wikipedia, the free and open-collaboration based online encyclopedia, has millions of pages that are maintained by thousands of volunteer editors. As per Wikipedia's fundamental principles, pages on Wikipedia are written with a neutral point of view and maintained by volunteer editors for free with well-defined guidelines in order to avoid or disclose any conflict of interest. However, there have been several known incidents where editors intentionally violate such guidelines in order to get paid (or even extort money) for maintaining promotional spam articles without disclosing such.

In this paper, we address for the first time the problem of identifying undisclosed paid articles in Wikipedia. We propose a machine learning-based framework using a set of features based on both the content of the articles as well as the patterns of edit history of users who create them. To test our approach, we collected and curated a new dataset from English Wikipedia with ground truth on undisclosed paid articles. Our experimental evaluation shows that we can identify undisclosed paid articles with an AUROC of 0.98 and an average precision of 0.91. Moreover, our approach outperforms ORES, a scoring system tool currently used by Wikipedia to automatically detect damaging content, in identifying undisclosed paid articles. Finally, we show that our user-based features can also detect undisclosed paid editors with an AUROC of 0.94 and an average precision of 0.92, outperforming existing approaches.

## CCS CONCEPTS

• **Information systems → Wikis**; **Data mining**.

## KEYWORDS

Wikipedia, Detection of abusive content, Malicious editors, Sockpuppet accounts.

**ACM Reference Format:**
Nikesh Joshi, Francesca Spezzano, Mayson Green, and Elijah Hill. 2020. Detecting Undisclosed Paid Editing in Wikipedia. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366423.3380055

## 1 INTRODUCTION

Wikipedia is the free online encyclopedia based on the principle of open collaboration; for the people by the people. Anyone can add and edit almost any article or page. However, volunteers should follow a set of guidelines when editing Wikipedia. The purpose of Wikipedia is "to provide the public with articles that summarize accepted knowledge, written neutrally and sourced reliably" [1] and the encyclopedia should not be considered as a platform for advertising and self-promotion. Wikipedia's guidelines strongly discourage any form of *conflict-of-interest (COI) editing* and require editors to disclose any COI contribution. *Paid editing* is a form of COI editing and refers to editing Wikipedia (in the majority of the cases for promotional purposes) in exchange for compensation. The guidelines set by Wikipedia are based on good faith, and malicious editors who earn a living through paid editing Wikipedia choose to ignore the requirement to disclose they are paid. Moreover, these malicious editors often use *sockpuppet accounts* to circumvent a block or a ban imposed on the person's original account. A sockpuppet is an "online identity used for the purpose of deception." [2] Usually, several sockpuppet accounts are controlled by a unique individual (or entity) called *puppetmaster.*

The first discovered paid editing case was the "Wiki-PR editing of Wikipedia", in 2013. [3] Wiki-PR is a company, which still exists but is banned by Wikipedia, whose core business is to offer consulting services to create, edit and monitor "your" Wikipedia page. The 2013 investigation found out that more than 250 sockpuppet accounts were related to and controlled by the company. On August 31, 2015, Wikipedia community uncovered a bigger set of 381 sockpuppet accounts, as part of an investigation nicknamed "Orangemoody" [4], operating a secret paid editing ring where participants extorted money from businesses who had articles about themselves rejected. The Orangemoody accounts themselves may have been involved in the deletion of some articles.

When undisclosed paid articles or editors are identified, such pages are removed from Wikipedia, and accounts are blocked. However, the Wikipedia community still relies on administrators who manually track down editors and affected articles. The differences between good faith editing and spam can be hard for even experienced editors to see, and, with hundreds of articles to be examined each month, the review process can be tedious, inefficient, and possibly unreliable.

In this paper, we focus, for the first time, on automatically detecting Wikipedia undisclosed paid contributions, so that they can be quickly identified and flagged for removal. We make the following contributions. (1) We propose a machine learning-based framework to classify undisclosed paid articles that uses a set of features based on both article content, metadata, and network properties, as well

*WWW '20, April 20–24, 2020, Taipei, Taiwan*

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380055

[1] https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest
[2] https://en.wikipedia.org/wiki/Sockpuppet_(Internet)
[3] https://en.wikipedia.org/wiki/Wiki-PR_editing_of_Wikipedia
[4] https://en.wikipedia.org/wiki/Orangemoody_editing_of_Wikipedia

as the patterns of edit behavior of users who create them. (2) To test our framework, we built a curated English Wikipedia dataset containing 73.9K edits by undisclosed paid editors (including deleted edits) and 199.2K edits by genuine editors, with ground truth on undisclosed paid articles. (3) Through our experimental evaluation, we show that our proposed method can efficiently identify undisclosed paid articles with an AUROC of 0.98 and an average precision of 0.91. We also show that our approach outperforms ORES, [5] the state-of-the-art machine learning service created and maintained by the Wikimedia Scoring Platform team to detect content damage on Wikipedia. Finally, we demonstrate that our proposed user-based features can be used to detect undisclosed paid editors as well, achieving an AUROC of 0.94 and an average precision of 0.92 and outperforming other existing approaches for sockpuppet detection in Wikipedia.

## 2 RELATED WORK

Different forms of content damage on Wikipedia have been studied in the literature, including vandalism, hoaxes, and spam. Wikipedia vandalism is "the act of editing the project in a malicious manner that is intentionally disruptive", e.g., through text that is humorous, nonsensical, or offensive. [6] Detecting vandalism was the very first problem studied in the context of Wikipedia content deception. Research shows that linguistic, metadata and user reputation features are all important to detect vandal edits in Wikipedia [1, 2, 11, 18]. Kumar et al. [8] addressed the problem of detecting vandal users and proposed VEWS, a warning system to early detect these users that leverages editor's behavioral patterns.

Kumar et al. [9] studied the characteristics and impact of Wikipedia hoaxes, articles that deceptively present false information as fact. They showed that Wikipedia hoaxes can be detected by using features that consider the article structure and content, hyperlink network properties, and hoaxes' creator reputation.

Spam on Wikipedia refers to the unsolicited promotion of some entities such as external link spamming and advertisements masquerading as articles (as the promotional articles written by undisclosed paid editors). The majority of the work on spam detection on Wikipedia has focused on detecting link spamming via metadata, URL properties, landing site characteristics [16, 17], or spam users by using behavioral-based features [5]. *To the best of our knowledge, there is no work addressing the problem of detecting promotional Wikipedia articles or undisclosed paid edits.*

Some bots and tools run on Wikipedia to detect vandalism or general damaging edits. ClueBot NG [7] and STiki [8] [18] are designed to detect vandalism. ClueBot NG is a bot that analyzes edit content, scores edits and reverts the worst-scoring edits. STiki is an intelligent routing tool that suggests potential vandalism to humans for definitive classification. It works by scoring edits by metadata and reverts and computing a reputation score for each user. These days, Wikimedia ORES[5] is the state-of-the-art approach to classify the quality of Wikipedia articles. Specifically, given an article, ORES evaluates the content of the article according to one of the following classes: spam, vandalism, attack, or OK. Thus, we compare our proposed approach to detect undisclosed paid articles with ORES in Section 5.2.2.

**Table 1: Size of Positive and Negative Data. Positive data refers to newly created paid articles or known undisclosed paid editors (UPEs).**

| | Positive Data | Negative Data |
|---|---|---|
| Newly Created Articles | 748 | 6,984 |
| Editors | 1,104 (UPEs) | 1,557 |
| Total Num. of Edits | 73,931 | 199,172 |

As explained in the Introduction, undisclosed paid editors typically act as a group of sockpuppet accounts. In the literature, several works have analyzed and detected sockpuppet accounts in online social networks and discussion forums [4, 7, 10, 15]. Specifically to Wikipedia, Solorio et al. [12, 13] have addressed the problem of detecting whether or not two accounts are maintained by the same user by using text authorship identification features. Other approaches have focused on classifying sockpuppet vs. genuine accounts by using non-verbal behavior and considering editing patterns [14, 19].

## 3 DATASET

This section describes the dataset we used to perform this study. We collaborated with an English Wikipedia administrator [9] active in reviewing articles that may have a conflict of interest (especially paid editing) to collect and curate a dataset of newly created positive articles, created by known undisclosed paid editors, and newly created negative articles, created by genuine users who are not paid editors. We collected the data through the publicly available Wikipedia API. We were able to access currently deleted edits from known undisclosed paid editors, thanks to our administrator's account. Deleted edits are not visible to general users through the Wikipedia API.

To gather the set of positive articles, we started by considering a manually curated set of 1,006 known undisclosed paid editor (UPE) accounts from English Wikipedia, which includes accounts from 23 different sockpuppet investigations [3]. Another set of 98 additional known UPE accounts were manually added by our Wikipedia administrator, resulting in a total of 1,104 UPE accounts. Among the set of new articles created by these UPE accounts, our administrator manually classified 748 of these articles (authored by 330 different editors) as paid articles (positive data). [10]

To collect the set of negative articles, we started by retrieving accounts of users who created a new article (or moved pages created in their user page or draft page to the article namespace as some UPEs do) in March 2019 (time of data collection) and who, similarly to UPEs, had made relatively few edits (less than 200 edits) in their account lifetime. 1,557 of these users resulted in being genuine, i.e., they are not known paid editors (or even Wikipedia blocked users [11]), or potentially paid editors as manually verified by our Wikipedia administrator. Then, we considered as the set of negative

[5] https://www.mediawiki.org/wiki/ORES

[6] https://en.wikipedia.org/wiki/Vandalism_on_Wikipedia

[7] https://en.wikipedia.org/wiki/User:ClueBot_NG

[8] https://en.wikipedia.org/wiki/Wikipedia:STiki

[9] Smart SE, https://en.wikipedia.org/wiki/User%3ASmartse

[10] Not all the articles created by a UPE are paid articles as these editors may create "genuine" articles to build a reputation.

[11] List of all Wikipedia blocked users: https://en.wikipedia.org/wiki/Special:BlockList

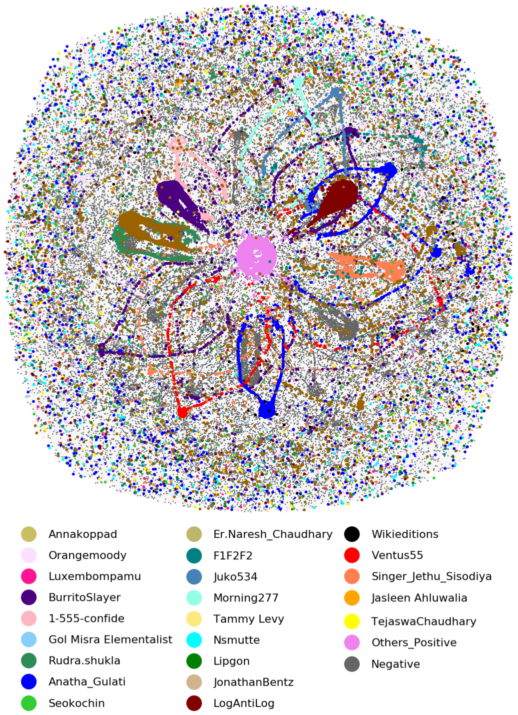

**Figure 1: Article network: two articles are connected by an edge if they have been edited by a common user. Colors indicate articles create by the same sockpuppet group of undisclosed paid editors (UPEs). Negative articles (in gray) are articles never edited by an UPE.**

articles, all the newly created articles by these genuine users. This resulted in 6,984 articles.

For each article in the positive and negative sets, we built a dataset containing the username of the user who created the page, the creation timestamp, the content of the article corresponding to the last edit by the article creator, and computed its size (in bytes). Further, in order to be able to compute features about the article creator account, we collected the time when the account was created and the whole edit history of all the genuine and UPE accounts collected as explained above. For each of these edits (or revisions), we collected timestamp, edited page title, revision ID, revision size, and size difference with respect to the previous version of the edited page. As explained above, deleted edits by UPEs are included in this data, providing us with a complete edit history for the UPE accounts. In total, we collected 73,931 edits by our 1,104 UPEs and 199,172 edits by our 1,557 genuine editors. Table 1 summarizes the size of the collected data.

### 3.1 Article Network Analysis

We built an article-article network where Wikipedia articles are nodes, and there is an edge between two articles if the same user has edited them. We considered the edit history of all the users in our dataset for creating this network. Figure 1 shows the resulting network (93,406 nodes and 44,264,072 edges) where colored nodes represent articles where at least one undisclosed paid editor contributed (referred as positive articles in the rest of this section), and gray ones indicate articles edited only by benign users (referred as negative articles in the rest of this section). Two positive nodes have the same color if the same sockpuppet group has created them. We used the list of sockpuppet investigations in the context of undisclosed paid editing provided by Ballioni et al. [3].

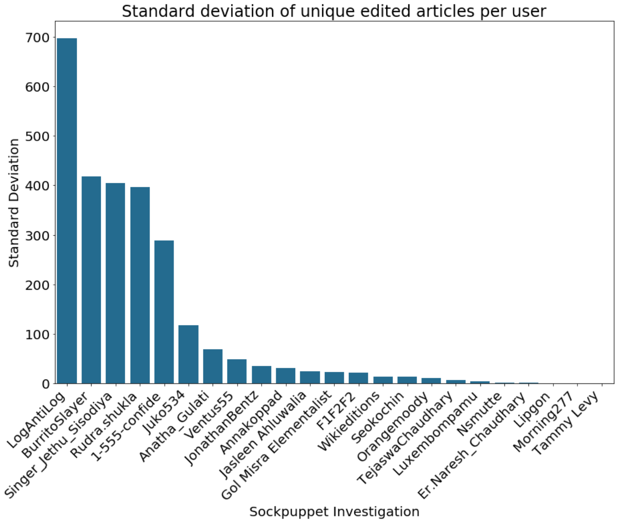

**Figure 2: Standard deviation of unique edited articles per user per sockpuppet investigation.**

By studying the network, we found that positive articles are less central in the network than negative ones. On average, positive articles have a PageRank of 1.15e-05 (vs. 1.17e-05 for negative ones) and an average local clustering coefficient (LCC) of 0.966 (vs. 0.974 in the case of negative articles). In both cases, the means are different with a $p$-value $< 0.001$ according to an independent $t$-test. This means that there is less user collaboration among positive articles. UPEs only work on a limited number of Wikipedia titles that they are interested in promoting, whereas genuine users edit more pages related to their field of expertise. That results in negative pages being more tightly knit in the network. This result also shows that sockpuppets accounts’ behavior in Wikipedia is different from sockpuppetry in online discussion communities where sockpuppets’ main goal is to interact with each other to deceive other users, and they have higher PageRank and LCC than benign users [7].

By looking at the articles edited by the same sockpuppet group in Figure 1, we observe that, for some investigations, the corresponding pages are more clustered than others. To further understand the meaning of different cluster shapes, for each investigation, we computed the number of unique articles edited by each sockpuppet account and used the standard deviation (SD) to measure the amount of variation of unique edited articles per account in each one of the investigations. Figure 2 shows the SD values for each investigation. We found out that the investigations with higher standard deviation tend to form denser clusters with a “drop” shape. This is the case, for instance, of the LogAntiLog investigation that has the highest standard deviation of 697 among all the investigations with fewer sockpuppet accounts contributing to most of the articles. For investigations with standard deviations in the intermediate range, corresponding clusters tend to form "bracket" shapes as UPEs' contributions are more distributed among affected articles. For example, Singer_Jethu_Sisodiya (SD=405)/Rudra.shukla (SD=397) and Ventus55 (SD=49)/Anatha_Gulati (SD=70) investigation pairs have similar standard deviation values and form near-identical shapes. As we move to investigations with a lower standard deviation of unique edited articles per account, clusters start to lose shape (e.g., Orangemoody investigation with SD=12). This analysis suggests that different undisclosed paid editor groups may adopt different editing strategies, which makes the problem of detecting undisclosed paid articles more challenging.

## 4 FEATURES FOR IDENTIFYING UNDISCLOSED PAID ARTICLES

In this section, we propose two sets of features to be used with our framework based on properties of articles as well as the edit history of users who created such articles. The features we chose are based on the patterns or behaviors that are more likely to be associated with malicious behavior and paid editing to distinguish them from a benign one. The list of features considered in our approach is described in the following two subsections.

### 4.1 Article-based Features

This first set of features we propose includes features related to the article such as metadata, content, and network-based features:

*Age of user account at article creation (user_age)* - Since sockpuppet accounts are more likely to be created at the time of the creation of an article, the age of user account at time of article creation can be considered as one of the features in detecting undisclosed paid articles.

*Infobox* - This feature checks if the article contains the infobox. The infobox is "a fixed-format table usually added to the top right-hand corner of articles to consistently present a summary of some unifying aspect that the articles share and some time to improve navigation to other interrelated articles." [12] Undisclosed paid editors tend to add the infobox to the pages they create to increase the exposure of the entity they are promoting as the presence of an infobox is an easy way for humans to grasp a summary of article content.

*Number of references* - This feature indicates the total number of references (including URL links) present in a given Wikipedia article. Regular Wikipedia articles (especially newly created ones) have a lot of missing references, and researchers have been addressing the problem of suggesting proper references [6]. On the other hand, the purpose of creating undisclosed paid articles is promotional, hence several explicit references to the promoted item are added at time of page creation. Therefore, a higher number of references in a given article can be a useful indicator of undisclosed paid editing.

*Number of photos* - This feature refers to the number of photos present in a given Wikipedia article. Uploading images on Wikipedia is relatively complicated as it requires copyright verification. The majority of images added to Wikipedia articles are removed within hours or days of being uploaded because of inappropriate, insufficient, or inaccurate copyright information. Then, to avoid that a promotional article looks suspicious because of its associated images, undisclosed paid articles tend to have fewer images than regular articles.

*Number of categories* - This feature represents the number of categories associated with a given article. Articles that belong to many categories deal with more complex topics and are less likely to be undisclosed paid articles.

*Content length* - This feature indicates the total length, in bytes, of the content of the given article. As regular pages are more curated and edited collaboratively by many editors, they tend to have more content and being longer in size than undisclosed paid ones.

*Network-based features* - We also consider the article *PageRank* and *Local Clustering Coefficient (LCC)* as additional features for the article (cf. Section 3.1).

### 4.2 User-based features

The second group of features we propose refers to characteristics, such as choice of username and editing behavior, of the user account that created the article. All the features but the username-based ones are computed by considering the history of contributions made by the editor.

*Username-based features.* Characteristics of usernames can be linked to malicious users that could create undisclosed paid articles [20]. For instance, Green and Spezzano [5] showed that username-based features are important to detect Wikipedia spammers. Thus, we consider the *number of leading digits*, the *number of digits*, the *ratio of digits to characters*, and the *ratio of unique characters* in username as features indicating a suspicious account.

*Average size of added text (avg_size_added)* - Given an editor, this feature computes the average size of text added to an article by the editor. Undisclosed paid editors are more likely to create new article content offline and then add it to Wikipedia at once, while benign users edit Wikipedia directly with smaller additions over time.

*Average time difference (avg_time_diff)* - This feature indicates the average time between two consecutive edits made by the same user. As explained in the above feature, undisclosed paid editors do not regularly edit Wikipedia. They work mainly offline and then add the content whenever they are ready. Thus, we expect the average time difference to be higher for these malicious editors than benign editors.

*Ten-byte ratio* - This feature computes the percentage of edits made by a user that are less than 10 bytes. Undisclosed paid editors try to become *autoconfirmed* users; thus they typically make around 10 minor edits before creating a promotional article. A registered user account becomes automatically autoconfirmed if the account is more than four days old and has made at least 10 changes. Autoconfirmed users are considered benign users that are therefore allowed to move pages to a different title and make changes to pages that have been semi-protected by administrators. The main reason for having autoconfirmed status on Wikipedia is to prevent vandalism and other types of disruptive editing. [13]

*Percentage of edits on User or Talk pages (user_talk_edits)* - This feature computes the percentage of edits a user has done on a User or Talk page. Undisclosed paid editors may want to edit User or Talk

[12] https://en.wikipedia.org/wiki/Help:Infobox

[13] See https://en.wikipedia.org/wiki/Wikipedia:User_access_levels

**Table 2: Performance of our proposed features to detect undisclosed paid articles according to different classification algorithms (best scores highlighted in bold) and comparison and combinations with ORES features (which are article-based) according to AUROC and average precision metrics.**

| | Article-based Features | | User-based Features | | Article + User Features | |
|---|---|---|---|---|---|---|
| | AUROC | Average Precision | AUROC | Average Precision | AUROC | Average Precision |
| **Our Features** | | | | | | |
| Random Forest | **0.856** | **0.507** | 0.971 | **0.893** | **0.983** | **0.913** |
| Logistic Regression | 0.734 | 0.230 | 0.675 | 0.171 | 0.656 | 0.153 |
| Support Vector Machine (SVM) | 0.555 | 0.171 | **0.980** | 0.823 | 0.556 | 0.172 |
| **ORES (Random Forest)** | 0.844 | 0.424 | - | - | - | - |
| **Our Features + ORES (Random Forest)** | 0.905 | 0.597 | 0.974 | 0.877 | 0.981 | 0.907 |

pages for several reasons: they want to have a User page to look like genuine editors; they may draft some content on the article Talk page before moving it to the main article page. Further, the content of an article is discussed by Wikipedia editors on the article Talk page or the contributor's User page. As the contribution of undisclosed paid editors may be disputed by administrators and genuine editors, we expect these malicious editors to engage more in editing these types of pages than genuine editors.

## 5 EXPERIMENTAL RESULTS

In this section, we report on the classification performances of our proposed features to detect undisclosed paid articles, in comparison with ORES, the scoring system tool currently used by Wikipedia to automatically detect damaging edits such as vandalism and spam. Also, in addition to the above main task, we also perform experiments in Section 5.3 to investigate the performances of our user-based features to predict undisclosed paid editors.

### 5.1 Experimental Setting

We tested our features for the classification task by using three different classification algorithms, namely Logistic Regression, Support Vector Machine (SVM), and Random Forest. We used class weighting to deal with class imbalance in all the classifiers. Class weighting is a way to learn from an unbalanced dataset where the classification imposes, during training, a penalty proportionally inverse to the class distribution on the model for making classification mistakes. To evaluate the performances, we considered the Area Under the Receiver Operating Characteristics curve (AUROC) and the Average Precision metrics, which are well-suited to measure classification results in case of unbalanced data, and performed stratified 5-fold cross-validation.

### 5.2 Detecting Undisclosed Paid Articles

Our results on the main task of detecting newly created undisclosed paid articles are reported in Table 2. In this experiment, the user-based features describing the behavior of the user, e.g., average time difference, have been computed by considering the user edit history up to the time of page creation. We observe that Random Forest is the overall best performing classification algorithm. When we consider article-based features only, we achieve an AUROC of 0.856 and an average precision of 0.507. Instead, when we consider the user-based features alone, we obtain better performances than article-based features. In this case, this group of features achieves an AUROC of 0.971 [14] and an average precision of 0.893. Moreover, when we combine both article and user-based features, we improve our classification results upon each group of features individually: AUROC of 0.983 and average precision of 0.913. This means that both article content and information about the account that created the article are important for detecting undisclosed paid articles.

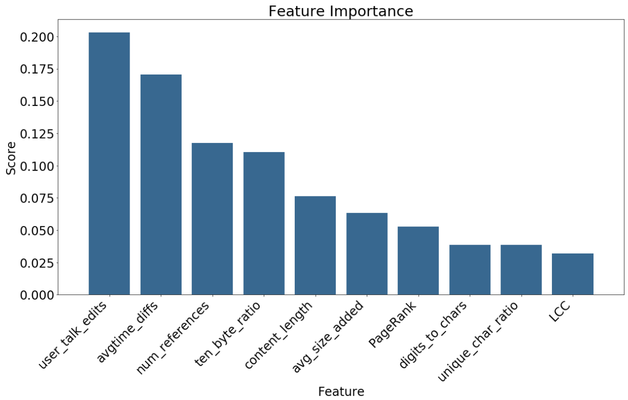

**Figure 3: Top-10 most important features for detecting undisclosed paid articles.**

*5.2.1 Feature analysis.* To analyze our features, we computed feature importance via a forest of randomized trees. Let $F$ be a set of features. The relative importance (for the classification task) of a feature $f \in F$ is given by the depth of $f$ when it is used as a decision node in a tree. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples they contribute to can thus be used as an estimate of the relative importance of the features. Figure 3 shows the importance of our set of features for the undisclosed paid articles classification task. The red bars in the plot show the feature importance using the whole forest. The variability of feature importance scores across the trees in the forest is minimal (less than 0.0001).

Among all the features we defined (article and user-based), the top four most important features are: *percentage of edits on User or Talk pages*, *average time difference*, *number of references*, and *ten-byte ratio*. We observe that, on average, the value of the percentage

[14] A slightly higher AUROC of 0.98 is achieved with SVM.

of edits on User or Talk pages feature is higher for positive articles (0.008) than for negative ones (0.0007). These values confirm our hypothesis that users who create undisclosed paid articles are more engaged in editing User and Talk pages than genuine users. Further, we see that, on average, users who create undisclosed paid articles edit more slowly than genuine users: the value of the average time difference feature is 2.8 days for regular articles and 9.2 days for undisclosed paid articles. Also, the percentage of edits that are less than 10 bytes in size is higher for users who created undisclosed paid articles: the value of the ten-byte ratio feature is, on average, 0.38 for positive articles and 0.34 for negative ones. This pattern aligns with the typical behavior of UPEs who make around 10 minor edits, then remain quiet for a few days waiting for becoming *autoconfirmed* users (the process takes 4 days), and then create a promotional article followed by the account going silent [3]. The third most important feature is the number of references in the newly created article. We observe that, on average, positive articles have more references than negative ones: 7.06 vs. 4.88. As explained in Section 4.1, this aligns with the fact that regular Wikipedia articles have more missing references than undisclosed paid ones that instead use references to the promoted item.

*5.2.2 Comparison with ORES.* As explained in Section 2, ORES is a web service developed by Wikimedia Foundation that provides a machine learning-based scoring system for edits. It analyzes the edit content and metadata. One of the offered services is to evaluate the quality of an article draft by assigning to the article a probability distribution of being in one of the following four classes: spam, vandalism, attack, or OK. To compare our proposed approach with ORES, we retrieved the draft quality scores for the positive and negative articles in our dataset by using the ORES publicly available API [15] and used them in input to a classifier to predict undisclosed paid articles. We found that Random Forest performed better than both logistic regression and SVM; hence, we report random forest results only at the bottom of Table 2. Our proposed approach significantly outperforms ORES that achieves an AUROC of 0.844 and average precision of 0.424 in detecting undisclosed paid articles. Specifically, we drastically improve the average precision (+49%), mainly thanks to inclusion on the user-based features in our approach. We argue that adding user-based features to ORES (which are not currently considered) would be beneficial to increase its performances in detecting damaging edits. For instance, if we combine our user-based features with ORES, we increase ORES AUROC and average precision scores to 0.974 and 0.877, respectively, on the task of detecting undisclosed paid edits.

## 5.3 Detecting Undisclosed Paid Editors

As we have seen in the previous section, the user-based features described in Section 4.2 are better than content-based ones in detecting undisclosed paid articles. Thus, in this section, we want to investigate the effectiveness of these features on the different, but related task of detecting undisclosed paid editors. As reported in Table 1, we have 1,104 UPEs and 1,557 benign users in our dataset. In this experiment, we consider the whole user edit history to compute the user-based features. To compare with ORES, given a user, we retrieved and averaged the draft quality scores of all their edits.

**Table 3: Performance of our user-based features to detect undisclosed paid editors and comparison with related work according to AUROC and average precision. Results are computed with Random Forest (best classifier).**

| | AUROC | Average Precision |
|---|---|---|
| Our User-based Features | 0.937 | 0.916 |
| ORES | 0.718 | 0.620 |
| Our User-based Features + ORES | **0.950** | **0.931** |
| Yamak et al. [19] | 0.934 | 0.885 |

Results are reported in Table 3 for Random Forest (best classifiers). As we can see, our features achieve an AUROC of 0.937 and an average precision of 0.916 and outperform ORES. By combining our features with ORES we slightly improve the performances to an AUROC of 0.950 and an average precision of 0.931.

*5.3.1 Comparison with other work on detecting sockpuppet accounts in Wikipedia.* As we have seen in Section 2, there is work addressing the problem of detecting sockpuppet accounts specifically to Wikipedia. As undisclosed paid editors use sockpuppet accounts, we compare our approach with this related work as well. Yamak et al. [19] showed that their approach based on the contribution behavior of the users achieves better performances than other works based on the analysis of the contribution text [12, 13] or using nonverbal behavior [14]. Hence, we compare our approach with the one by Yamak et al. [19] only. [16] As we can see from Table 3 (last row), our approach has a comparable AUROC score but achieves a better average precision (+3%) than the competitor.

# 6 CONCLUSIONS

In this paper, we addressed the problem of identifying undisclosed paid articles in English Wikipedia. Our proposed approach relies on article-based and user-based features that describe potential malicious behavior. Through our experimental evaluation, we showed that we can detect such articles with an AUROC of 0.98 and an average precision of 0.91. As our features are independent of linguistic barriers, our proposed approach can work on any Wikipedia language version. We also showed that our user-based features can be used to identify undisclosed paid editors with 0.94 AUROC and 0.92 average precision. Our results in detecting undisclosed paid articles and editors improve over state-of-the-art approaches.

# ACKNOWLEDGMENTS

The authors would like to thank Smart SE (English Wikipedia administrator) for providing the dataset and contributing to the paper with his domain expertise. They also thank Aaron Halfaker for his help in retrieving ORES scores for deleted articles.

# REFERENCES

[1] B. Thomas Adler, Luca de Alfaro, Santiago Moisés Mola-Velasco, Paolo Rosso, and Andrew G. West. 2011. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In *Computational Linguistics and*

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

[16] Yamak et al. included a feature that considers whether an edit has been reverted by another user, making the detection not completely automated as human input is required. As we propose an automatic detection approach that does not rely on human input, we did not include the reverted-based feature in our implementation of Yamak et al. approach for a fairer comparison.

*Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II.* 277–288.

[2] B. Thomas Adler, Luca de Alfaro, and Ian Pye. 2010. Detecting Wikipedia Vandalism using WikiTrust - Lab Report for PAN at CLEF 2010. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy.*

[3] Tony Ballioni, James Heilman, Brian Henry, and Aaron Halfaker. 2018. Known Undisclosed Paid Editors (English Wikipedia). (4 2018). https://doi.org/10.6084/m9.figshare.6176927.v1

[4] Zhan Bu, Zhengyou Xia, and Jiandong Wang. 2013. A sock puppet detection algorithm on virtual spaces. *Knowledge-Based Systems* 37 (2013), 366–377.

[5] 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, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.* 532–535.

[6] Abhik Jana, Pranjal Kanojiya, Pawan Goyal, and Animesh Mukherjee. 2018. WikiRef: Wikilinks as a route to recommending appropriate references for scientific Wikipedia pages. *arXiv preprint arXiv:1806.04092* (2018).

[7] Srijan Kumar, Justin Cheng, Jure Leskovec, and V. S. Subrahmanian. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017.* 857–866.

[8] 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, Sydney, NSW, Australia, August 10-13, 2015.* 607–616.

[9] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016.* 591–602.

[10] Dong Liu, Quanyuan Wu, Weihong Han, and Bin Zhou. 2016. Sockpuppet gang detection on social media sites. *Frontiers of Computer Science* 10, 1 (2016), 124–135.

[11] Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic Vandalism Detection in Wikipedia. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings.* 663–668.

[12] 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.* 59–68.

[13] Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2013. Sockpuppet detection in wikipedia: A corpus of real-world deceptive writing for linking identities. *arXiv preprint arXiv:1310.6772* (2013).

[14] Michail Tsikerdekis and Sherali Zeadally. 2014. Multiple account identity deception detection in social media using nonverbal behavior. *IEEE Transactions on Information Forensics and Security* 9, 8 (2014), 1311–1321.

[15] Bimal Viswanath, Ansley Post, Krishna P Gummadi, and Alan Mislove. 2011. An analysis of social network-based sybil defenses. *ACM SIGCOMM Computer Communication Review* 41, 4 (2011), 363–374.

[16] Andrew G. West, Avantika Agrawal, Phillip Baker, Brittney Exline, and Insup Lee. 2011. Autonomous link spam detection in purely collaborative environments. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, 2011, Mountain View, CA, USA, October 3-5, 2011.* 91–100.

[17] Andrew G. West, Jian Chang, Krishna K. Venkatasubramanian, Oleg Sokolsky, and Insup Lee. 2011. Link spamming Wikipedia for profit. In *The 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, CEAS 2011, Perth, Australia, September 1-2, 2011, Proceedings.* 152–161.

[18] Andrew G. West, Sampath Kannan, and Insup Lee. 2010. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *Proceedings of the Third European Workshop on System Security, EUROSEC 2010, Paris, France, April 13, 2010.* 22–28.

[19] Zaher Yamak, Julien Saunier, and Laurent Vercouter. 2016. Detection of multiple identity manipulation in collaborative projects. In *Proceedings of the 25th International Conference Companion on World Wide Web (Companion).* 955–960.

[20] Reza Zafarani and Huan Liu. 2015. 10 Bits of Surprise: Detecting Malicious Users with Minimum Information. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015.* 423–431.