# One Month with the Devils: A Re-Analysis of Content Polluters on Twitter in 2010

Matthew Chang
*Department of Computer Science & Engineering*
*Texas A&M University*
College Station, TX 77843 USA
matthewchang@tamu.edu

Eric Nunes
*Department of Computer Science & Engineering*
*Texas A&M University*
College Station, TX 77843 USA
eric27n@tamu.edu

***Abstract*—Detecting spam accounts, or "content polluters" is a necessity if we are to protect the social media space from becoming overrun with unwanted content. In this paper, we build upon previous research of content polluter detection on Twitter, and analyze another important feature that has proved to be invaluable to spam detection- hashtags. In recreating and training a random forest model that combines hashtag features on top of old methods we find that although the paper's accuracy of 98.42% could not be reproduced, however, adding hashtag features increased the accuracy of our base model from 95.84% to 96.25%. In spite of the minor increase in accuracy, this is observed as an indicator in favor of hashtag-oriented spam analysis. View our project presentation in video form here: https://www.youtube.com/watch?v=gm0baVN95SU. View the notebooks used in the paper here: https://github.com/eric27n/439-704-research-fall-2023.**

***Index Terms*—Twitter, spam, hashtags, machine learning**

## I. Introduction

Like other social medias and online communities, Twitter (now X) faces the problem of content polluters. These content polluters can be criminals, spammers, or other accounts who post unwanted content or content that goes against Twitter's community guidelines.

The detection of content polluters is vital for a healthy and sustainable Twitter community. Content polluters can spread misinformation or distribute malware, harming legitimate users who do not recognize these malicious tactics. For legitimate Twitter users who are aware of the tactics of content polluters, spam accounts and Tweets dilute the content that users come to Twitter to see. Ultimately, this pushes users away from using the platform, weakening the bonds of a legitimate Twitter community. For Twitter, a loss of a legitimate base is detrimental, decreasing the reputation of the platform and revenue as a business.

In this paper, we focus on a Twitter dataset from 2010 which contains user and Tweet data and corresponding labels of legitimate user or content polluter. Using this dataset, we extract several features which can be used to differentiate content polluters from legitimate users using machine learning techniques.

We first attempt to replicate the results of the previous paper by extracting the same features and using the random forest model, which performed the best. We then proceed to add hashtag-related features to our feature-set to see how we can improve the performance of our classifier. We analyze how adding hashtag features contributed to our model and why they might be useful to the detection of content polluters on Twitter.

## II. Related Work

The primary research that our paper builds off of was performed in a 7-month long study published in 2011 [1]. In this paper, they tracked 41,499 users and over 5 million of their tweets over the course of seven months. Their goal was to build a classifier to distinguish between content polluters and legitimate users using the information they collected. After trying several classifiers on the data, they found that the random forest performed best with an accuracy of 98.42% and an F-score of 0.984.

Along with the results of their classifier, the paper contributed a dataset with user profile, follower, and Tweet data for both legitimate users and content polluters as well as an analysis of Twitter user features that are important to training a classifier. Our work tries to replicate the results from this paper, and then expand on it with ideas from another paper.

The idea for performing hashtag analysis with tweets was inspired by papers like Hspam14, conducted by Sedhai and Sun [2], which created their own dataset for the purposes of studying what they coined *hashtag-oriented spam*. However, their dataset lists shortcomings such as a lack of full profiles in their dataset, and it can be hard to verify the veracity of individual tweets being spam or not. As a result, and in part because over half of tweets in the dataset are no longer publicly available, their dataset is not used in this paper, though some ideas on hashtag analysis have been taken into consideration when researching for this paper.

Finally, a group in Brazil lead by Benevenuto et al. detected spam surrounding popular Tweets a few months before the paper by Caverlee and others [3]. Unlike Caverlee who took an account-based approach, Benevenuto uses an SVM to create a hybrid approach, analyzing behaviors of an account and the contents of Tweets to make an inference on identifying users as spammers. Some of the features he analyzed used analysis of features involving hashtags, and one of the ten most important features was the average number of hashtags per tweet.

## III. Methodology

Our research process can be divided into two stages. In the first stage, we build a base classifier which tries to replicate the results of a 2011 study [1]. In the second stage, we methodically add hashtag features to our dataset and re-evaluate the performance of the model[1].

### A. Dataset

The dataset we worked with comes from the 2011 study. It contains 22,223 content polluter user profiles and 2,380,059 of their tweets as well as 19,276 legitimate user profiles and 3,263,238 of their tweets. The data was harvested over 7 months using a method called a social honeypot- a way of luring in content polluters and tracking their activity.

TABLE I: Dataset

| Class | User Profiles | Tweets |
|---|---|---|
| **Content Polluters** | 22,223 | 2,380,059 |
| **Legitimate Users** | 19,276 | 3,263,238 |

### B. Feature Extraction

To create our base classifier, we first determined what features to use from the original paper. These features are descriptors for the Twitter accounts that can help us differentiate between a content polluter or legitimate user. Of the 18 features used in Dr. Caverlee's experiment, only 15 were extractable from the dataset.

5 of the 18 features were directly extractable from the dataset:

1) length of user's screen name
2) length of user profile's description
3) number of following
4) number of followers
5) total number of posted tweets

10 of the features had to be recalculated from the dataset based on the information given:

1) longevity
2) ratio of number of following to number of followers
3) number of posted tweets per day
4) average number of links per tweet
5) average number of unique links per tweet
6) average number of mentions per tweet
7) average number of unique mentions per tweet
8) average content similarity over all pairs of tweets
9) ZIP compression ratio of posted tweets
10) change rate of number of following

While we assume that the calculations reflect the authors' original intent with feature extraction, it is not clear if we calculated the above metrics in the same way that they did. They did not include documentation on their calculations so we have used common metrics to capture the descriptions of their features as best as possible. For example, for "average content similarity over all pairs of tweets" a metric of similarity or method of calculation was not defined. For this, we used the Jaccard Index for measuring similarity between every pairwise combination of a user's tweets and then took the average index. This is just one example of a calculation we made that may not align perfectly with the original authors' calculations. Please refer to our Jupyter Notebook for more details on how each feature was calculated.

3 of the features from the original paper were not extractable at all due to missing data:

1) percentage of bidirectional friends (both following and follower)
2) standard deviation of the unique numerical IDs the user is following
3) standard deviation of the unique numerical IDs that follow the user

In order to calculate the missing features, we needed some sort of information on the exact user's who were followed by and following each given user. We assume that the original authors retrieved this information using the Twitter API. However, reconstructing the data now with the Twitter API is nearly impossible due to the accounts getting deleted or suspended, username changes, follower changes, etc that have occurred in the nearly 15 years since initially extracting the data. So instead, we will leave these 3 features out and see how the model performs. According to the past experiment, standard deviation of following was the single most important feature in differentiating content polluter from legitimate user, and standard deviation of followers was the third most important of the eighteen features extracted. Not including these features may be detrimental to the performance of the model, but unfortunately, there is not a clear solution on how to add them.

### C. Base Classifier

We split the data into 80% for training and 20% for testing. For our base classifier, we trained a random forest model using the 15 features from above. To obtain maximum performance on our training data, we ran a grid search hyperparameter tuner and found the best parameters over 12,150 models. Then, we measured the performance of the RF using the best hyperparameters.

### D. Adding Hashtags

On top of the features that we could translate from Caverlee's paper, the main focus of this paper is to see if improvements to the model could be made through analysis of the hashtags in use. Of the 5.6 million tweets available in the dataset, 470,563 tweets, or a little more tha 8% of all tweets, incorporated the use of hashtags. However, because the objective is to identify profiles as either legitimate users or content polluters, and not individual tweets, it was hypothesized that a portion of users, both legitimate and polluter, would hold different patterns with hashtags and their usage. Each tweet was analyzed for hashtags, and the following five features were added to each account during training:

[1]The dataset and all notebooks/models can be found at https://github.com/eric27n/439-704-research-fall-2023

1) number of tweets with at least one hashtag

2) total number of hashtags divided by number of tweets given

3) total number of unique hashtags divided by number of tweets given

4) average number of characters in a user's hashtags

5) average number of words in a user's hashtags

The number of hashtags that exist among a user's tweets can potentially provide information on the person's use (or perhaps overuse) of hashtags to identify them as a legitimate user or content polluter. As seen in Figure 1, there is a slight tendency for content polluters to use more hashtags in their tweets when compared to legitimate users.

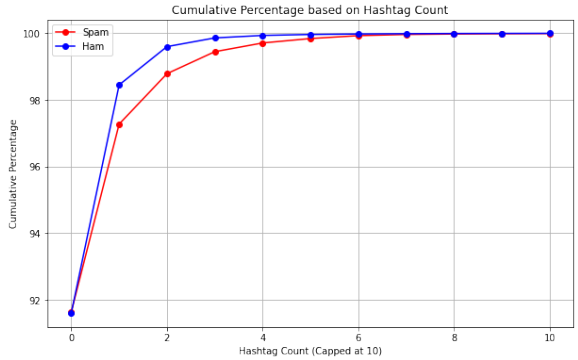



Fig. 1: Cumulative percentage of the number of hashtags per tweet, separated by tweets from legitimate users (blue) and content polluters (red).

The average number of characters and words in hashtags was a feature worth exploring upon learning of the top 20 most popular tweets among legitimate users and content polluters, as seen in Table II.

Though there is some overlap among the two categories - hashtags that content polluters use since they know they are popular and can get traction - there is a remarkable difference between the differing hashtags in the top 20 between the two categories. Legitimate users appear to use hashtags in more niche, concentrated subjects, as can be seen with hashtags like *epicpetwars*, *theresway2many*, and *p2*. By contrast, content polluters tend to stick with more broader, generic topics, such as *business*, *travel*, and *apple*, perhaps to achieve a wider reach when discussing these large ventures. It is also observed that legitimate users, although they can stick with shorter hashtags, are also more likely than content polluters to promote hashtags with more words. This trend can be observed in Figure 2.

## IV. Results

The original model constructed from the 15 extractable features performed comparably to the classifier in the original

TABLE II: Top-20 Most Popular Hashtags by User Class

| **Legitimate Users** | | **Content Polluters** | |
|---|---|---|---|
| ***Hashtag*** | ***Frequency*** | ***Hashtag*** | ***Frequency*** |
| ff | 8667 | quote | 8704 |
| fb | 7668 | ff | 6582 |
| jobs | 4679 | nowplaying | 6069 |
| musicmonday | 3111 | news | 2747 |
| tcot | 2650 | followfriday | 2653 |
| followfriday | 2377 | fb | 2400 |
| iranelection | 1998 | worldcup | 1485 |
| mm | 1927 | jobs | 1464 |
| fail | 1779 | free | 1321 |
| whatsbetter | 1597 | nfl | 1295 |
| justbecause | 1415 | health | 1290 |
| tebakbandtranslate | 1403 | twitter | 1262 |
| tinychat | 1351 | tcot | 1261 |
| news | 1321 | business | 1198 |
| shjobs | 1200 | marketing | 1190 |
| epicpetwars | 1195 | musicmonday | 1188 |
| quote | 1187 | iphone | 1163 |
| classicmoviequotes | 1159 | teamfollowback | 1128 |
| theresway2many | 1097 | travel | 1112 |
| p2 | 1002 | apple | 1101 |

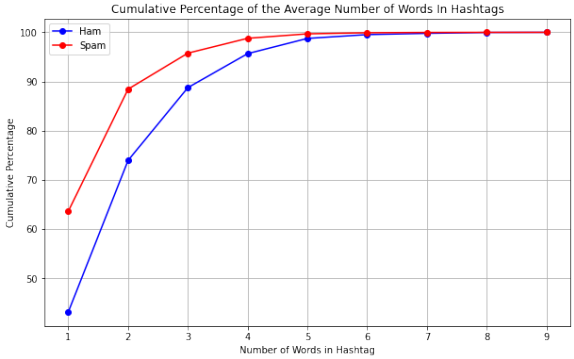



Fig. 2: Cumulative percentage of words inside hashtags, separated by tweets from legitimate users and content polluters.

paper. Our classifier achieved an accuracy of 95.84% and an F-score of 96.16%. While these metrics are not nearly as good as the 98.42% accuracy of the original paper's best RF classifier, they are consistent with the 95% to 98% accuracy achieved in the upper echelon the models they tested, and perform especially well considering that it lacks three features that the authors considered important to prediction, two of those being considered part of the top three overall most important features.

Fortunately, the second most important feature, "Following Change Rate" was extractable from the data. Following change rate is a measure of how much the number of accounts a user is following changes over time. So accounts that follow and/or unfollow a large number of accounts is a short period of time will have a higher following change rate. Fig 1 shows the difference in the distribution of the two classes for following change rate. It appears that content polluters tend to follow and unfollow a lot of people in a short amount of time as compared to the organic methods of a legitimate user.

The base classifier performs slightly better on content

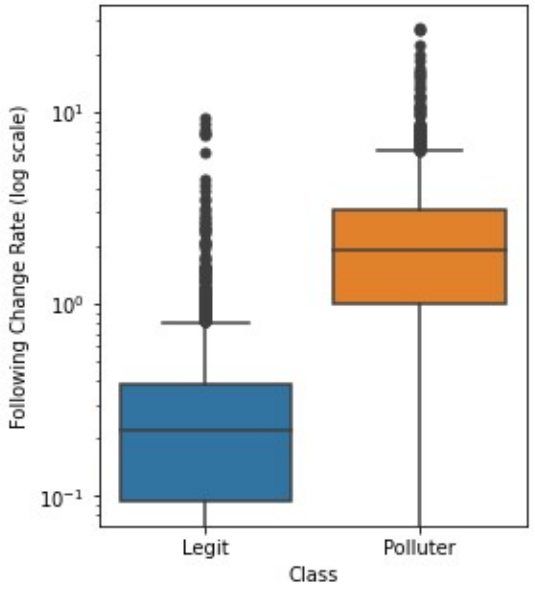


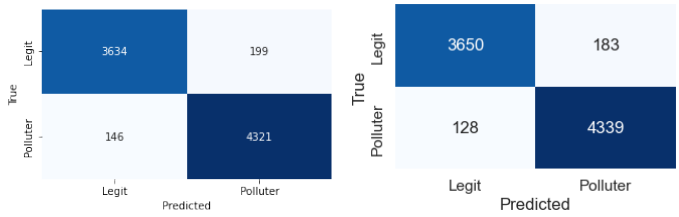

Fig. 3: Boxplot for the comparing the different distributions for legitimate users and content polluters over the "Following Change Rate" feature. The data in the plot represents a random sample of 10% of the dataset for easier visualization of outliers.

polluters than legitimate users, as can be seen in Fig. 3. 96.7% of the polluters were classified correctly compared to 94.8% for the legitimate users. This may be attributed to the fact that there is more content polluter data than legitimate user data in the dataset.

The improved model with hashtag analysis performed slightly better than the base model, with an overall accuracy of 96.25%, and an F-score of 96.54%. This is an improvement to the base model, and still falls within the range of the 95-98% accuracy reported in the original paper, though it still lags behind the 98.42% accuracy achieved in the original paper, and the improvement from the base model ultimately comes down to a few extra cases.

The exact hyperparameters for splitting and sampling the dataset were not published in the previous paper, so it is difficult to compare exact numbers, but it appears that our performance differs in that we had more false negatives (content polluter predicted as legitimate user) than false positives (legitimate user predicted as content polluter) whereas they had more false positives than false negatives. If they kept they kept the original distribution of the dataset in their training, then it seems that they had a more even performance across the two classes than we did.

In spite of the minor improvements, the overall importance of the hashtag features to the improved model were small. Upon gathering feature importances using mean decrease in impurity, out of the twenty features observed in the improved model, not one of the five hashtag features made the top 10 most important. Some of the most important features, such as the following change rate and number of followings, were




(a) Base Model (b) Improved Model

Fig. 4: Confusion matrices for the base and improved random forest models. There is a higher proportion of false negatives than false positives in both models, with the improved model seeing a greater disproportion.

features identified as among the most important in the original paper.

## V. Conclusion

Though neither the base model nor the improved model could achieve as high an accuracy as the original paper, given how some important data has been lost to time, a score of approximately 96% is not far from the initial goal. In between the two models, though an improvement to the accuracy could be seen by including hashtag analysis, the improvement was minimal at only 0.3% that of the base model. The reasoning behind this could be for a few reasons. First and foremost, over 90% of the tweets did not incorporate the use of hashtags, and hence, there was a strong number of users, both legitimate and content polluters, that did not use hashtags. Second, in spite of a few missing features, the base model has an accuracy above 96%, and with less than 4% of test cases proving to be

However, in spite of the small improvements, more observation into this analysis would be beneficial, particularly for more modern tweets. For starters, a limitation of this study is that the data used comes primarily from 2009-2010, and thus is no longer much relevant outside of reproduction. With the rise of artificial intelligence and robots gaining a better ability to appear human through tweets, a more rigorous analysis on hashtags and other tweet-based features could provide further insights into modern content polluters. In addition, a limitation of this study was not being able to utilize graph networks to analyze relationships among users, mentions, and hashtags. Further research can be done based on graph networks to better help distinguish legitimate users and content polluters.

## References

[1] K. Lee, B. Eoff, and J. Caverlee, "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter", ICWSM, vol. 5, no. 1, pp. 185-192, Aug. 2021.

[2] S. Sedhai and A. Sun, "HSpam14: A Collection of 14 Million Tweets for Hashtag-Oriented Spam Research", Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 223-232, Aug. 2015.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter", in Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), 2010, vol. 6, p. 12.