Instagram Spam Detection

Wuxain Zhang and Hung-Min Sun
Department of Computer Science
National Tsing Hua University
Email: jack482653@is.cs.nthu.edu.tw; hmsun@cs.nthu.edu.tw

***Abstract*— In recent years, Instagram has become one of top 15 online social networks. However, popularity of Instagram also causes advertisement and spam posts flooding. Therefore, it is necessary to build a spam detection model to decrease number of spam posts in Instagram. We present a scheme applying feature-based method and supervised learning technique to detect spam posts from Instagram. We use K-fold cross validation to find best pair of supervised learning model and parameters of the model and accuracy of our best model is 96.27%.**

Keywords*—*Social networks, Instagram, Spam detection, Spam, Machine learning.

I. Overview of Proposed Spam Detection Approach

There are many proposed spam detection approaches (e.g., keyword detection, machine learning). Most spam detections only handle text, but Instagram posts contain media. So, we choose machine learning technique to build our spam detection model. By analysing the dataset, feature vectors are extracted from media posts and user profiles, including 4 user profile features, 1 Colour Difference Histogram feature, and 23 media post features. We collect data from Instagram and label data manually (e.g., advertisement, post with irrelative hashtags). 1983 user profiles and 953808 media posts are acquired. Two-step clustering method is applied to group similar posts into same clusters.

Our method consists of four parts.The following are brief description of these four parts.

A. Data Collection

We collect Instagram users' data as many as possible. Each user data include: 1) profile 2) followers 3) followings 4) media which the users post 5) images of the media posts. We download images to the local directory and store the other data to the database.

B. Clustering Media Posts

Minhash is a technique for quickly estimating how similar two sets are. To reduce the number of cluster which minhash clustering generates, we label the minhash cluster which size is greater than or equal to 5 first, and apply K-medoids method to remain posts in the minhash clusters which sizes are less than 5. As a result, we can limit number of clusters to acceptable size and make labeling data procedure faster.

C. Labeling Data

If media post meets at least one of following conditions, it is likely to be spam:

1) Unsolicited message a) Mention other users arbitrarily. b) Use number of unrelated hashtags (e.g., tag *#night* in the photo shot in the daytime). c) Use more than 20 hashtags. However, some user may prefix # to every word in the text message, and it is not spam.
2) Text message has advertisement words such as a) #buy, #sell or price b) Website link
3) Text message contains repeated words
   a) Use more than 3 synonymous hashtags.
   b) Spam hashtag: i) followforfollow ii) follow4follow iii) f4f iv) likesforlikes v) likeforlikes vi) likesforlike vii) likeforlike viii) like4like ix) l4l x) tagsforlikes xi) followme xii) followus xiii) fftc4life
4) Picture a) Contain official watermark. b) Looks like displaying products. c) Contain advertisement words, contacts, prices of products, request of following some user.

D. Training the Classifier

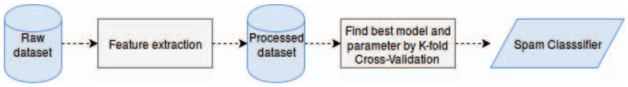

Fig. 1: Flow chart of training the classifier

Flow chart of training the classifier is shown in Figure 1. Raw dataset is the labeled dataset we do in the Section I-C. In the feature extraction phase, each raw data is transformed to feature vector and stored in processed dataset. Each feature vector is extracted from the media post and the profile of user who posts the media. The following are features of a instance:

1) Features of the profile of user who posts the media a) Number of media b) Number of followers c) Number of followings d) Does biography of the user contain website links?
2) Features of the media post a) Number of tags b) Number of likes c) Number of comments d) Number of mentions e) Minhash of uni-gram of the text message f) Minhash of bi-gram of the text message g) Minhash of tri-gram of the text message h) Does the text message contain prices? i) Does the text message contain currency symbols? j) Does the text message contain website links? k) Colour difference histogram of the photo

Next, we determine best model and combination of parameters by K-fold cross-validation, using a part of processed dataset(i.e. training dataset) as input. Finally, entire training dataset is applied to train spam classifier based on best model and combination of parameters. The other part of processed

978-1-5090-5652-1/17 $31.00 © 2017 IEEE
DOI 10.1109/PRDC.2017.43

dataset(i.e. testing dataset) is used to evaluate the performance of the spam classifier.

II. Experimental Result and Analysis

A. Best (model, parameters) pair

After running 10-fold cross validation, the best pair of parameters is (Random forest, (maxDepth: 8, numTrees: 20, impurity: entropy)) and it is average of area under ROC is 0.99238256. We use this pair and whole training data to build new classifier and evaluate it is performance by testing data. The confusion matrix of predicted results is shown in Table I and Table II. All matrics are greater than 90 percent. Accuracy of predicted result is 96.27%, precision rate is 97.48%, recall rate is 95.05%, and F1 measure is 96.25%.

		Predicted cond.		
	$n = 21099$	**Pos.**	**Neg.**	
True condition	**True**	10084	525	10609
	False	261	10229	10490
		10345	10754	

TABLE I: Confusion matrix

Metric	Value
Area under ROC curve	0.992431107606
Accuracy	0.962747049623
Precision rate	0.974770420493
Recall rate	0.95051371477
F1 score	0.962489262193

TABLE II: Performance metrics

B. Execution Time and Throughput

We measure two parts of our program: 1) Feature extraction 2) Rescaling and making prediction. In feature extraction, it split to two parts: 1) Feature extraction of user profiles and 2) Feature extraction of media.

We launch the program of evaluating execution time of feature extraction evaluation as local mode under node 191. It extracts user features from 1622 profiles and media features from 1038 media. We repeat these two extraction procedure 10 times to get average elapsed time. The result of performance of feature extraction is shown in Table III.

The program of rescaling and prediction is submitted as standalone mode under node 174, 191, 196 and 204. It perform rescaling 104620 feature vectors and predict them; we also measure their elapsed time 10 times. In the result of performance of rescaling and prediction (Table IV), average throughput of prediction is twice as much as that of rescaling.

	Elapsed Time (sec.)		**Throughput (inst./sec.)**	
No.	**User**	**Media**	**User**	**Media**
1	0.4620	217.2724	3511.0353	4.7774
2	0.0950	214.2908	17076.7219	4.8439
3	0.0480	209.7811	33796.4664	4.9480
4	0.0530	209.6517	30620.5940	4.9511
5	0.0828	214.8556	19577.5547	4.8312
6	0.0541	215.1801	29957.0716	4.8239
7	0.0836	218.3701	19405.8330	4.7534
8	0.0471	216.2927	34427.2106	4.7991
9	0.0486	213.8768	33401.2229	4.8533
10	0.0687	217.6338	23594.1510	4.7695
Average	0.1043	214.7205	24536.7861	4.8351

TABLE III: Performance of feature extraction

	Elapsed Time (sec.)		**Throughput (inst./sec.)**	
No.	**Rescale**	**Predict**	**Rescale**	**Predict**
1	30.6705	15.6325	3411.1005	6692.4830
2	31.6696	15.9523	3303.4873	6558.3127
3	31.4022	15.5558	3331.6143	6725.4843
4	31.1010	15.4065	3363.8805	6790.6513
5	31.8283	15.3184	3287.0093	6829.6737
6	30.9703	15.6038	3378.0744	6704.7800
7	30.9894	15.2713	3375.9970	6850.7535
8	31.3067	15.1097	3341.7812	6924.0193
9	30.9913	15.7189	3375.7839	6655.7011
10	31.6868	15.0617	3301.6861	6946.0927
Average	31.2616	15.4631	3347.0414	6767.7952

TABLE IV: Performance of rescaling and prediction

III. Conclusion

We present a scheme applying feature-based method and supervised learning technique to detect spam posts from Instagram. We collect user profiles and media posts from Instagram. To mark media posts quickly, we use two-pass clustering method(i.e., Minhash clustering and K-medoids clustering) to group the near-duplicate posts into same clusters. We mark these posts as spam or normal based on the clustering result.

Before training classifiers, raw data are transformed to feature vectors. We consider not only statistics of user profiles and posts, but also information implied in photos, which is different from other researches. Therefore feature vectors is extracted from media posts and user profiles, including 4 user profile features, one Colour Difference Histogram feature, and 23 media post features.

Finally we use 10-fold cross validation to find best pair of supervised learning model and parameters of the model. The best pair is (Random forest, (maxDepth: 8, numTrees: 20, impurity: entropy)) and accuracy of our best model is 96.27%.

In the future, we will purpose to design a scheme of customizing spam classifiers according to users' favour. Someone may view posts promoting clothes as spam while others who love shopping think they are normal. As a result, It is necessary to customize spam classifiers according to users' favour.

Acknowledgment

This research was supported in part by the Ministry of Science and Technology, Taiwan, under the Grants MOST 105-2218-E-007-011 and MOST 103-2221-E-007-073-MY3.