Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram

Indira Sen*
IIIT-Delhi
indira15021@iiitd.ac.in

Anupama Aggarwal
IIIT-Delhi
anupamaa@iiitd.ac.in

Shiven Mian
IIIT-Delhi
shiven15094@iiitd.ac.in

Siddharth Singh
IIIT-Delhi
siddharth14105@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT-Delhi
pk@iiitd.ac.in

Anwitaman Datta
NTU, Singapore
Anwitaman@ntu.edu.sg

ABSTRACT

Instagram is a significant platform for users to share media; reflecting their interests. It is used by marketers and brands to reach their potential audience for advertisement. The number of likes on posts serves as a proxy for social reputation of the users, and in some cases, social media influencers with an extensive reach are compensated by marketers to promote products. This emerging market has led to users artificially bolstering the likes they get to project an inflated social worth. In this study, we enumerate the potential factors which contribute towards a genuine like on Instagram. Based on our analysis of liking behaviour, we build an automated mechanism to detect fake likes on Instagram which achieves a high precision of 83.5%. Our work serves an important first step in reducing the effect of fake likes on Instagram influencer market.

CCS CONCEPTS

• **Networks** → **Online social networks**; • **Information systems** → Social networks;

KEYWORDS

Fake Social Engagement, Online Social Networks, Instagram

ACM Reference Format:
Indira Sen, Anupama Aggarwal, Shiven Mian, Siddharth Singh, Ponnurangam Kumaraguru, and Anwitaman Datta. 2018. Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3201064.3201105

1 INTRODUCTION

Apart from being used as a medium of communication, Online Social Networks (OSNs) are also used to gain popularity, increase social self-worth and promote businesses. Even brands, advertisers and the background recommender algorithms of OSNs rely on the popularity metrics of users and content shared on these services. To get more benefits, users often artificially increase the popularity and engagement on their content in several ways. Some of the prevalent ways are – to leverage bots, purchase social metrics such as – likes, followers, and shares from black market services, and become part of collusion networks which can be used to trade inorganic engagement. Such artificial bolstering of popularity can cause brands to lose money [1], advertisers to not reach the relevant audience, and recommender algorithms to give poor suggestions [13].

There have been several prior efforts to identify fraud [6], spam [4] and fake users [6] on OSNs. In this study, we instead focus on inorganic engagement received by a user. Previous studies aiming to detect fake liking behaviour, assume that if a user has given one or two fake likes, all her likes are fake [2, 8]. However, we believe this is a limited understanding of fake engagement since a single user can generate organic, as well as inorganic engagement. For instance, an Instagram user can *like* content which she is genuinely interested in, and in addition, the same user can also be a part of a colluding 'like-back' network, where she likes unrelated and random content only to receive back some likes and increase her own popularity. Therefore, we propose that the true reach / social-worth of the user should be determined by canceling out the effect of fake engagement which she receives, and should largely depend only on the organic engagement (we use the terms organic and genuine interchangeabley). We define organic liking engagement on Instagram as a *like*-action which a user gives to a post when she has genuine interest in the content, or in the user posting the content (poster). In this study, our goal is to identify the ingenuity of likes by determining user's intention of liking a post. In particular, we define the goal as — *Given* a liker $\mathcal{L}$, who likes a specific post p of a poster $\mathcal{S}$ – *Find out* the features of $\mathcal{L}$, p and $\mathcal{S}$, *to determine* the probability of liker $\mathcal{L}$ *genuinely* liking a post p.

Here, we find attributes of fake liking which can help us distinguish such behavior from organic liking activity. Unlike previous studies on spam detection which assume that a single spam post translates into the user being a spammer [2, 8], we infer the reach of an Instagram user as a function of the organic and fake likes, reducing the effect of fake likes (to some extent - subject to the efficacy of our approach) in the process. Our contributions are –

Characterizing Fake and Organic Likes. We discern fake and organic likes by determining the factors which can lead to a user genuinely liking a post on Instagram. While previous studies have looked at meta-features of profiles, content, and structure, we focus

*the work was done partly while the author was visiting NTU Singapore as part of the NTU-India Connect Research Internship Programme.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WebSci '18, May 27–30, 2018, Amsterdam, Netherlands
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5563-6/18/05…$15.00
https://doi.org/10.1145/3201064.3201105

our efforts in identifying the probability of a user liking particular content based on various factors like topical interest and proximity with the source user. We study an extensive list of features indicating genuineness of a like instance. Our findings show that topical interest of liker with the post, and profile quality are most helpful.

Automatic Detection of Fake Likes. As next step, based on the understanding we gain in the previous step, we build a machine learning based model to automatically distinguish a fake like from an organic like. We are able to detect fake liking instances with a precision of 83.5% using a neural network model.

2 RELATED WORK

User Behaviour. While not as widely researched as other OSNs, such as Facebook and Twitter, studies on Instagram explore user behaviour [11, 20]. Especially close to our work is Jang et al.'s analysis of Instagram but without a focus on fraud [12] where they note the lack of reciprocity when it comes to liking behaviour.

Malicious Engagement on OSNs. Malicious entities on OSNs have been widely studied with a particular focus on Facebook [9, 15], Twitter [4, 21], and to a lesser extent, on Instagram [6]. Our work differs from these since we aim to spot suspicious *behaviour*, specifically, fake liking. While fake liking and other fraudulent activities may co-occur, the former can exist without the latter.

While not explored as widely as detection of fake entities, fake engagement on OSNs has been previously studied on Facebook [2, 5, 8], Twitter [10] and Youtube [16]. Beutel et al. presents an understanding of the 'lock-step' behaviour in co-ordinated fake likers on Facebook, using temporal snapshots [5]. Giatsogolou et al. study fraudulent retweeting behaviour on Twitter by analyzing network and temporal patterns [10] while Li et al. detect fake engagement on Youtube using spectral clustering [16]. While network and temporal features are effective, they are often difficult to obtain. In such cases, content-based analysis can yield fruitful results. Badri et al. aims to weed out fake liking on Facebook pages using profile and post features of likers using a supervised classification model [2].

This present work adds to the use of content-based features in two ways - *first*, by also taking into account the relationship between a poster and a liker. *Second*, we account for Instagram's visual nature by taking the properties of the image into consideration.

3 DATA

Fake Like Instances (`FakeLike_data`): There are multiple sources of fake likes such as paid web-services or apps, trading platforms where a user participates in a giving likes in exchange for likes, and bots which are triggered based on hashtags. Instagram also allows users to post videos and maintains its view count and like count [1]. We assume that if a video has received likes, but has zero views, then the like instances are fake, because they were generated without properly seeing the content. We capture 16,448 such *like instances* (information about the liker, post, and source user), and add it to `FakeLike_data`. Such fake likes instances could have been generated from any of the aforementioned sources. We acknowledge that our fake like instances dataset can be much more comprehensive; we leave it for future work.

[1] View count increases by one when a user watched the content for more than 3 seconds. Like count increases by one on an explicit like action by the user.

Random Like Instances (`RandLike_data`): It is hard to obtain a true positive dataset of genuine likes. Therefore, instead we collect a much larger random set of like instances to draw comparison with fake likes, and to use as negative class to build a machine learning model to identify fake likes. Since Instagram does not provide a direct way to sample random users/posts, we obtain a seed set of Instagram users, [2] and extract their follower and followee connections in a breadth-first-search manner. This gives us a sample of 1 million Instagram users, from which we take a smaller subset of users and extract their posts, and likes on each of those posts. In this manner, we obtain a dataset of 134,669 like instances in `RandLike_data`. Note that this sample is much larger (more than 8 times) than the fake like instance dataset. Therefore, despite the noise, we assume that predominantly, the like instances in `RandLike_data` would be genuine. Though a noisy dataset is one of our current limitations, but with a clean negative dataset, our results showing differences between fake and other likes, and supervised learning based identification of fake likes would only improve. We summarize the data collected in Table 1.

Table 1: Dataset of Instagram like instances. We also collect meta-information of likers, posts and posters.

	#likes	#posts (p)	#Likers ($\mathcal{L}$)	#Posters ($\mathcal{S}$)
`FakeLike_data`	16,448	9,932	9,301	7,822
`RandLike_data`	134,669	1,717	47,233	738

4 ANALYSIS

In this section, we present and validate hypotheses that explain liking behaviour on Instagram. While it is virtually impossible to know why a user might like a post, it is possible to understand how the user could have come across the post, which is a non-trivial prerequisite for liking. Based on this intuition, we enumerate the plausible reasons behind a user genuinely liking another user's post. We begin by giving our definition of a *like instance — Given a poster $\mathcal{S}$ whose post p has been liked by liker $\mathcal{L}$, we define a like instance as the tuple ($\mathcal{L}$, p, $\mathcal{S}$).* A like instance is designed to contain post properties to ensure that a liker is evaluated on the basis of individual posts she likes. We do not assume that if a single like generated by a liker is fake, then all her other likes are also fake. Next, we define the following set of hypothesis which shed light on genuinely garnering likes on Instagram.

4.1 Network Effects

Instagram enables the user to view posts by her followees in her feed, and also *explore* the content liked by the followees. Previous studies have noted the role of social ties in reinforcing engagement on OSNs [3]. We take cues from this to propose the following –

H1.1: A liker $\mathcal{L}$ is more likely to genuinely like $\mathcal{S}$'s post if $\mathcal{L}$ is a follower of $\mathcal{S}$. In this case, $\mathcal{L}$ will receive the content posted by $\mathcal{S}$ in her home feed, and hence there is a higher chance of $\mathcal{L}$ genuinely liking that post. In addition, if $\mathcal{L}$ is following $\mathcal{S}$, we can assume that $\mathcal{L}$ is interested in $\mathcal{S}$'s content.

H1.2: A liker $\mathcal{L}$ is more likely to genuinely like $\mathcal{S}$'s post if $\mathcal{L}$ is a follower of $\mathcal{S}$'s followers. Instagram also lets the users

[2] Seed set contains users whose content was featured on Instagram's official account.

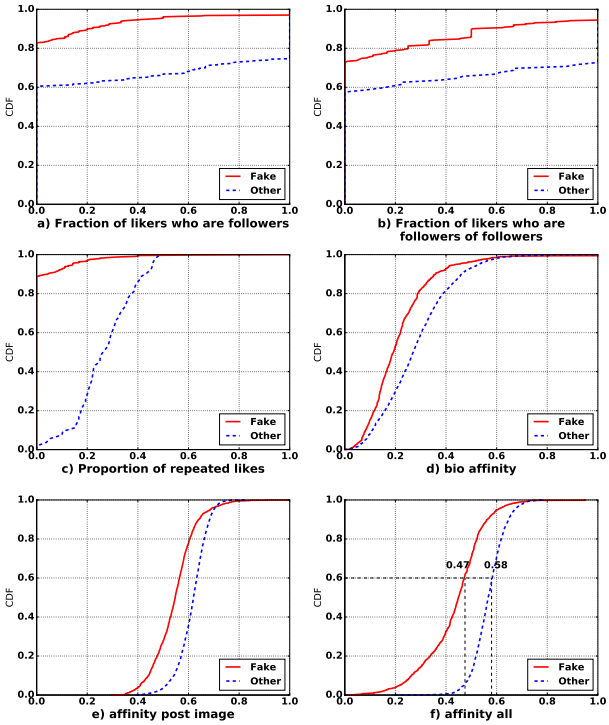

Figure 1: Cumulative distribution function of the like properties for fake likes and other likes.

follow the activities of the users being followed. Therefore, a liker $\mathcal{L}$ can also come across the liked post p if it is liked by one of the users which $\mathcal{L}$ follows. We also consider such an instance to indicate a higher level of confidence in the genuineness of the like instance.

To test H1.1 and H1.2, we study the *follower*, and *follower-of-follower* connections of the posters in our dataset of fake and random likes, and found that genuine likers do indeed like their followees post more than fake likers do (KS-test =0.303, $p < 0.001$). For fake like engagements, there are significantly less proportion of likers which are followers of the poster. In case of fake engagements, only 16.8% of likers of a post are followers of the poster, as compared to a much higher fraction of 39.1% likers being followers in case of random like engagements (Figure 1(a)). We see similar observation in a two-hop network in Figure 1(b). Only 2.8% of likers of a post in case of fake likes are follower-of-follower of the poster, as compared to 42.4% in case of random like engagement [3].

4.2 Interest Overlap

H2: A user $\mathcal{L}$ will have a higher chance of genuinely liking $\mathcal{S}$'s post if $\mathcal{L}$ and $\mathcal{S}$ share interests. To capture interest overlap between two Instagram users, we first define their *Interest Profile* and the extent of overlap as *Affinity*.

Interest Profile: . *Given a user u we define u's interest profile as a set of topics* $(t_u^1, t_u^2, ..., t_u^n)$*, where these topics are inferred from u's bio* b_u *and posts* $(p_u^1, p_u^2 ..., p_u^k)$.

[3]Due to strict API restrictions, we were unable to collect the entire 2-hop network of all like instances in our dataset. The results reported here are of 55% of the likes.

Topic Extraction. Inspired by previous studies, which have used Instagram bios to detect user interests [19], we infer topics from textual sources such as bio and post captions using Wikification [17]. Instagram is an image-based OSN and the post images may also have latent topical informataion, which we leverage using Densecap captioning [14] to obtain meaningful captions. Wikification is applied on these captions to extract fine-grained topics.

Topic Matching. To match topics we utilize word2vec similarities [18] between two tuples of interests [4]. We define a post's attributes t_i as the wikified topics of the post, and define the topical similarity between users as follows –

Affinity. *Given user a and b's interest profile as* $I_a = (t_a^1, ..., t_a^n)$ *and* $I_b = (t_b^1, ..., t_b^m)$*, we define*

$$Affinity(T_a, T_b) = \frac{1}{n} \sum_{i=1}^{n} \max_{1 \le j \le m} w2vec(t_a^i, t_b^j)$$

We extract a user's interest profile from her bio, and by converting the post image into relevant text using Densecap. Note that this metric is not commutative and therefore penalizes likers with a very wide variety of interests, which is an indicator of suspicious behavior. It has been seen previously that genuine likers tend to like things related to a limited set of topics and are more discerning [2]. Therefore, we consider topical affinity as one of the distinguishing features to identify fake liking engagement. Note that the value of affinity is between [0, 1], with higher affinity value indicating high topical matching. We found that 60% of fake likers have an affinity value of 0.475, as compared to 0.58 affinity for same proportion of random set of likers (as seen in Figure 1(f) We have similar observations for affinity calculated over only bio (Figure 1(d)), and only text extracted from post image (Figure 1(e)). However, we combine both of these to enhance the discriminative effect.

4.3 Liking Frequency

H3: A liker $\mathcal{L}$ will genuinely like more than one post of the poster $\mathcal{S}$. We observe that legitimate likers keep coming back to the same poster. Figure 1(c) shows that 90% posters with fake likes get 7% repeated likers on their posts, as compared to the same fraction of posters with other likes getting 42% repeated likes.

4.4 Influential Poster

H4: A user $\mathcal{L}$ will have a higher chance of genuinely liking $\mathcal{S}$'s photo if $\mathcal{S}$ is an 'influential' user or a celebrity. It has been observed that celebrities garner a higher number of likes on their content [7]. We use the Instagram verification badge as a proxy for celebrity users. We observe that in our dataset, only 1.9% users were celebrities who got fake likes, as compared to 7.5% celebrities who got genuine likes, indicating that celebrities are more likely to attract a higher number of likes. Therefore, we consider a poster being a celebrity as a feature to gauge the genuineness of a like.

4.5 Link Farming Hashtags to get Fake Likes

Hashtags have been shown to play an important role in Instagram in spreading the reach of posts and attracting more likes [20], therefore we examine their role in fake liking behvaiour.

[4]$w2vec(i, j) \in [0, 1]$ is the word2vec similarity between the two entities, i and j.

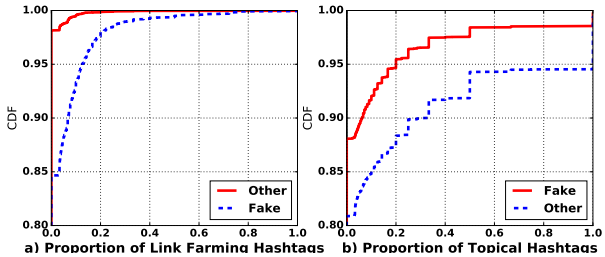

Figure 2: CDF of a) Proportion of Link Farming Hashtags and b) Proportion of Topical Hashtags used by posters. Fake like instances show a high proportion of link farming hashtags which can be used to solicit likes, but show a relatively lower proportion of topical hashtags.

H5.1: A user $\mathcal{S}$ is more likely to attract fake likes if she uses link farming hashtags in her posts. *Link farming* hashtags, such as 'like4like', 'like2follow', can be used to elicit fake likes. Such hashtags are often used by a community of users who collude to inflate each other's likes and followers. We curate a list of 112 such hashtags [5] and find that 20.8% posts with fake likes have at least one link farming hashtag as compared to 1.8% posts with random likes. Figure 2(a) shows that fake like instances tend to have a higher proportion of link farming hashtags.

4.6 Topical Hashtags

H5.2: A user $\mathcal{S}$ with genuine likes will have topical hashtags in their posts. Another aspect of posting behaviour is related to topical hashtags used in a post, instead of occasional (#throwbackthursday, #ootd), trending (#mayweather) or link farming hashtags (#like4like). We adopt a two-step process to detect topical hashtags. We first filter out all link farming hashtags as well as popular non-topical hashtags [6]. Next, we segment these hashtags and use Wikifier to see what proportion of hashtags pertain to a topic. Figure 2(b) shows that fake like instances tend to have a lower proportion of topical hashtags (KS test = 0.342, $p < 0.001$).

5 DETECTING FAKE LIKES

In addition to the features based on the hypotheses previously described, we also consider user-based attributes, viz, volume of posts generated, average number of posts per day [2] and profile completeness (presence of profile picture, name and bio) [8]. Finally, we compute the Chi-square values to rank features in the order of their efficacy in distinguishing between fake and other likes. Topical affinity scores high (16%) as do user based features, particulary profile completeness (11%), and fraction of topical hashtags (9%). [7]

5.1 Building a Classification Model

While the actual ratio of fake to genuine likes in Instagram is unknown, based on previous literature on spam detection [4], we maintain a ratio of roughly 1:8. This proportion ensures that any machine learning model trained on such a dataset can perform well 'in-the-wild' where the ratio of likes would be highly imbalanced. Therefore, we obtain the aforementioned features from `FakeLike_data` and `RandLike_data`, and train a supervised model on these features with fake likes as the positive class. We experiment with different classification algorithms viz. Logistic Regression, Random Forest, SVM (RBF kernel), AdaBoost (with Random Forest as base initiator), and XGBoost. We also use a simple feed-forward neural network – Multi-Layer Perceptron (MLP) which gives us the best performance. In all our experiments we perform 10-fold cross-validation, using 80% of the dataset as training data and 20% for validation. For the MLP based model, we use 2 hidden layers with 200 neurons each. Both layers use sigmoid activation function, and the output layer has a dropout of 0.2 to prevent overfitting.

Baseline. As a baseline, we use Badri et al.'s method to detect fake likers on Facebook [2]. Their work focuses on the properties of the liker, without considering the relationship between a liker and a poster. As the source code was unavailable, we implemented this method on our own based on the features detailed in the paper. The authors propose a supervised method for the detection of fake likes based on profile (length of biography, lifespan of account, number of bidirectional connections), posting activity (average number and maximum posts per day, total posts and skewness of posting), page liking (category entropy of pages liked, proportion of verified pages) and social attention (average number of likes and comments received) of the liker. We discard two features: (1) proportion of shared photos and (2) average number of shares received, since there is no concept of sharing posts on Instagram, unlike Facebook. The authors experiment with multiple classifiers and find XGBoost to perform best. We use Precision, Recall and Area under the ROC curve (AUC) to measure the performance of all models in detecting fake likes, with the latter being especially important in understanding the performance of a classifier trained on imbalanced data.

5.2 Experimental Results

We observe that boosted trees, Adaboost and XGBoost (also used in the baseline), perform well due to robustness against outliers. However, we achieve highest performance using the MLP with an average precision of 83% and recall of 81% (AUC of 89%) in detecting fake likes. Our results are inline with previous studies on fraud detection where neural networks have shown competitive performance [21]. We compare validation loss and training loss across training epochs and observe that they are close to each other, almost converging with increasing epochs. We, therefore, conclude that our fake like detection system does not overfit and generalizes well to unseen likes. Table 2 summarizes our results.

The baseline model gives a precision of 61%, a recall of 69%. Compared to the baseline model, our model obtains an 83% precision and 81% recall to detect fake like instances. The strength of our method lies in effectively capturing the parameters which affect genuine liking behavior on Instagram. We perform a manual investigation of 200 randomly sampled fake likes labeled genuine (false negatives) by the baseline. We find that these likes are given by users with well-formed profiles who superficially appear to be

[5] This list was shortlisted from a list of popular hashtags by two active Instagram users and we make it publicly available here: goo.gl/UshiXk

[6] Like the link farming hashtags, popular hashtags were curated by the same annotators and are available in the same url.

[7] We exclude the follower-of-follower from the feature list since the 2-hop network of the entire dataset could not be collected due to API restrictions.

Table 2: Results of various classifiers in differentiating between fake and random like instances across different feature types. We report precision and recall in detection of fake likes. The MLP model performs best.

Classifier	LogReg			Random Forest			SVM(RBF)			Adaboost			XGBoost			MLP		
Feature Type	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**
H1: Network Effect	0.1	0.71	0.32	0.12	0.67	0.35	0.16	0.72	0.66	0.15	0.75	0.59	0.17	0.70	0.64	0.24	0.73	0.67
H2: Interest Overlap	0.36	0.5	0.49	0.38	0.54	0.56	0.54	0.59	0.73	0.56	0.54	0.72	0.60	0.59	0.79	0.68	0.68	0.80
H3: Liking Frequency	0.11	0.62	0.38	0.12	0.53	0.36	0.18	0.6	0.65	0.13	0.35	0.61	0.15	0.43	0.68	0.22	0.51	0.71
H4: Influential Poster	0.11	0.62	0.33	0.11	0.5	0.35	0.12	0.71	0.6	0.16	0.6	0.68	0.22	0.59	0.69	0.31	0.70	0.72
H5: Hashtag features	0.21	0.56	0.54	0.21	0.55	0.51	0.28	0.6	0.69	0.49	0.26	0.72	0.44	0.34	0.71	0.60	0.51	0.76
User based features	0.32	0.16	0.57	0.3	0.21	0.57	0.4	0.22	0.70	0.53	0.61	0.74	0.54	0.67	0.73	0.61	0.74	0.79
All Features	0.39	0.67	0.67	0.39	0.64	0.62	0.58	0.65	0.77	0.65	0.60	0.78	0.69	0.65	0.81	**0.83**	**0.81**	**0.89**

benign users. On the other hand, our method classifies 78% of these likes as fake likes. We find that the affinity between the likers and posters in these like instances are lower than average. We also note the presence of link farming hashtags in the posts of these like instances. Therefore, we believe our system can detect fake likes given by genuine looking entities, as well, unlike current methods which entirely rely on the user properties to determine fake liking.

Error Analysis. To understand why our model is not able to detect 19% of the fake likes, we randomly sample 100 undetected fake likes and manually inspect them. We find that in 27 fake like instances, the likers were followers of the poster, potentially leading our model to misclassify such like instances as genuine. It suggests that some posters have fake followers and the fake likes are from such followers, something our current methodology is unable to capture. However, our approach can be modularly applied in a cascade, after detecting fake followers using previous techniques [6]. Furthermore, we found that 61 likers had a high topical interest overlap with the posts they had liked. A more thorough analysis showed that this was happening due to small set of interests (just one or two) of the liker, which results in high affinity value.

6 CONCLUSION

To complement previous studies on fake liking in other OSNs, we perform the first exploration of fake liking behaviour on Instagram. We build on existing content-based techniques of fraud detection on OSNs by incorporating factors that motivate liking on Instagram, such as liker-poster interest overlap. Additionally, we also account for Instagram's visual aspect by examining the contents of images. Our automated method is able to detect fake likes with 83% precision (22% increase on the baseline).

Limitations and Future Work. For collecting our ground truth data, we restrict ourselves to videos with likes but no views. In future, we plan to explore other sources such as trading web services and mobile apps. Our preliminary investigations show that there is a thriving underground ecosystem for fake liking on Instagram, including paid services and trading platforms, which we can analyse using our existing model. Another limitation and area of improvement is our affinity metric which has unpredictable behaviour when user interest tuples are small. Finally, other than these improvements, we plan to leverage our detection model to nullify or penalize the effect of fake likes and provide the actual reach of an Instagram user in terms of the genuine likes they get.

REFERENCES

[1] 2017. Social Media Experiment reveals how easy it is to create fake Instagram accounts and make money from them. http://www.independent.co.uk/life-style/gadgets-and-tech/social-media-experiment-fake-instagram-accounts-make-money-influencer-star-blogger-mediakix-a7887836.html. (2017).

[2] Prudhvi Ratna Badri Satya, Kyumin Lee, Dongwon Lee, Thanh Tran, and Jason Jiasheng Zhang. 2016. Uncovering fake likers in online social networks. In *ACM International Conference on Information and Knowledge Management.*

[3] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *ACM Conference on Human factors in Computing Systems.*

[4] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In *Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference.*

[5] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. 2013. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *International Conference on World Wide Web.*

[6] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. 2014. Uncovering large groups of active malicious accounts in online social networks. In *ACM SIGSAC Conference on Computer and Communications Security.*

[7] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *International AAAI Conference on Web and Social Media.*

[8] Emiliano De Cristofaro, Arik Friedman, Guillaume Jourjon, Mohamed Ali Kaafar, and M Zubair Shafiq. 2014. Paying for likes?: Understanding facebook like fraud using honeypots. In *ACM SIGCOMM conference on Internet Measurement.*

[9] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. 2010. Detecting and characterizing social spam campaigns. In *ACM SIGCOMM Conference on Internet Measurement.*

[10] Maria Giatsoglou, Despoina Chatzakou, Neil Shah, Christos Faloutsos, and Athena Vakali. 2015. Retweeting Activity on Twitter: Signs of Deception. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining.*

[11] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types. In *International AAAI Conference on Web and Social Media.*

[12] Jin Yea Jang, Kyungsik Han, and Dongwon Lee. 2015. No reciprocity in liking photos: Analyzing like activities in Instagram. In *ACM Conference on Hypertext & Social Media.*

[13] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *International Conference on Web Search and Data Mining.*

[14] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[15] Kyumin Lee, James Caverlee, and Steve Webb. 2010. Uncovering social spammers: social honeypots+ machine learning. In *International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[16] Yixuan Li, Oscar Martinez, Xing Chen, Yi Li, and John E Hopcroft. 2016. In a world that counts: Clustering and detecting fake social engagement at scale. In *International Conference on World Wide Web.*

[17] Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *ACM International Conference on Information and Knowledge Management.*

[18] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

[19] Aditya Pal, Amaç Herdagdelen, Sourav Chatterji, Sumit Taank, and Deepayan Chakrabarti. 2016. Discovery of topical authorities in instagram. In *International Conference on World Wide Web.*

[20] Ramine Tinati, Aastha Madaan, and Wendy Hall. 2017. InstaCan: Examining Deleted Content on Instagram. In *ACM Conference on Web Science.*

[21] Svitlana Volkova and Eric Bell. 2017. Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter Across Languages. In *AAAI International Conference on Weblogs and Social Media.*