Credibility Assessment of Public Pages over Facebook

Himanshi Agrawal and Rishabh Kaushal(✉)

Department of Information Technology,
Indira Gandhi Delhi Technical University for Women, Delhi, India
hagarwal.281@gmail.com, rishabh.kaushal@gmail.com

Abstract. With the growing use of online social media and presence of users on many such platforms, their interaction with social networks is huge. They are free to spread wrong information without any accuracy, integrity and authenticity checkpoints masquerading as legitimate content. All these wrong, unrelated, unwanted, manipulated information are distributed for some hidden reasons. Even, their distribution network is not limited to one social media platform. Sometimes, they use the social network as a market place either for advertising, promotion of particular website, product and an application. But these advertisements do not provide any incentive to Facebook as this content is just spam, irrelevant for Facebook. Dissemination of unwanted, unrelated information has become a huge problem not only on blog, discussion forum but also on online social network like Facebook. Due to lack of marking over content posted, this information become online and reader has no barometer to check either the credibility of commenter or poster or the credibility of facts. To address these issues, we have derived an equation to weigh the credibility of public pages and applied machine learning algorithms (MLA) over collected data to validate our prediction.

Keywords: Credibility · Online social media · Unrelated content · Machine learning algorithms · Public pages · Facebook

1 Introduction

Among the various social media platforms, Facebook is very popular. While there are many advantages that are offered by online social media, there are few issues faced as well. In this virtually connected world of online social media platforms, many aspects of our social interactions are disguised by garb of anonymity. With whom we are on chat, what kind of people are sending us emails, to whom we are friends on Facebook, Twitter, Google+ and most importantly our decisions, opinions are influenced or diverted by what kind of misguiding information is diffusing over social space. In spite of thousands of innovation and policy implementation, fraudulent users target the information that is available for public view on online social media platforms like Facebook, Twitter. These users indulge in fraudulent activities like spreading rumors, advertising content for promotion

P. Mueller et al. (Eds.): SSCC 2016, CCIS 625, pp. 188–198, 2016.
DOI: 10.1007/978-981-10-2738-3_16

of their business, etc. On public pages of Facebook, users that are not allowed to post typically contribute by commenting over the posts. It is often observed that users put unrelated and unwanted comments over multiple posts, multiple pages and in very limited duration. Such misplaced spam comments with this information can misguide the other users. In our work, we focus on this problem and assign credibility to these facebook pages taking into account the spam comment count out of the total comment count and fake user count out of total user count. As an extension, we performed prediction over collected dataset by using three conventional classifiers namely J48, Naive Bayes and Random Forest to take comment level, user level, userfeed level and page level features into consideration.

Key contributions of our work are as follows.

- We have used a comprehensive set of features taking into account all three dimensions namely message, source and media to assess credibility over Facebook public pages.
- We have taken care of *hinglish* text in comments.
- We proposed a metric to assess the credibility taking spam comment count over total comments count and fake user count over total users count into account and calculated score for public pages.

Further, paper is structured as follows. Related Work is elaborated in Sect. 2. Our proposed work and its results and observations have been described in Sect. 3 and Sect. 4, respectively. Conclusion and Future Work is discussed in Sect. 5.

2 Related Work

The problem of unsolicited content no longer restricted to blogs and discussion forums but has marked its presence over online social media with their offensive footprints. Along with this, other derivatives challenges have emerged like credibility of message content, publisher of content and media, where these activities are taking place, etc. Following are some previous works that address both base and derived issues.

2.1 Unrelated Spam Content

Wang et al. [1] have categorized diversionary comments into five types based on their observations and propose an effective framework to identify and flag them. They have also conducted a user study to verify the effect of identifying diversionary comments by asking the following question "Assume that you hold interest in the post discussion topic, is this comment of interest to you? [1] In their proposed work, they compute the relatedness between a comment and the post content, and the relatedness between a comment and its reply-to comment, which involves co reference resolution, extraction from Wikipedia, and topic modeling. They have evaluated 4,179 comments from Digg and Reddit. Dewan and Kumaraguru [2] classified the unsolicited malicious content generated

during news making events. For classification, an extensive feature set was taken comprising of 42 features. They had provided REST API and browser plug-in as well. URLs were extracted from the post and then those URLs were visited using Python request or LongURL API to remove the invalid URLs to find the final destination URL of URL mentioned in the post. Each URL was passed to six blacklist lookups e.g. Google SafeBrowsing, PhishTank etc. Abu-Nimeh et al. [3] assessed the prevalence of malicious and spam posts in Facebook. They have analyzed more than half a million posts with the help of Defensio, a Facebook application that protects users from such content as well as filters profanity and blocks URL categories. They surveyed the temporal and network-level properties of those posts containing URLs that Defensio had determined to be malicious or spam. They have concluded by their research that much more re- search is needed to gain a better grasp of the true extent and nature of security threats in online social networks.

2.2 Credibility Assessment of Source, Content and Media

Metzger et al. [4] has defined *"credibility is defined as the quality of being trustworthy. In communication research; information credibility has three parts, message credibility, source credibility, and media credibility.* Thai et al. [5] have worked to identify the origin point of misinformation on OSN platform. They had studied k-suspector problem which works for identification of the top k most suspected origin point of misinformation. For this, they had proposed two efficient approaches, ranking-based and optimization-based algorithms. They had applied their approaches on differently structured networks. Thus their experiments and observations show that their proposed approach is very helpful to discover the origins of misinformation up to 80 % accuracy and hence increase the solidity of facts sharing on OSN. Lin et al. [6] have identified the need of a framework for all online social media to analyze consequences for the dissimilar practices of posting on content generated by user over different OSN workspace. This discovery of dissimilar practice had become prior knowledge for ensemble classifier to measure the quality of content. The proposed approach had shown good results over different OSN like Slashdot and Apple discussion forum. Ismail and Latif [7] have tried to address the issues like uncertainty sources of it and last but not the least is quality of content published over online social network. Basically, they had arranged a questionnaire survey to get an outline of social network trends in Association of Southeast Asian Nations (ASEAN) countries and in Malaysia particularly. Through this, they had identified features like credibility of publishers, quality of information they shared and lack of continuity in online content. For the analysis and proof of the framework, they had applied multiple regression method and from the results they had concluded that lack of continuity of online contents create more uncertainty among users in comparison to lack of quality of information and even the publisher credibility. AlMansour et al. [8] had examined the different organizations for credibility of information based on methodologies and parameters used and classify Twitter surveys based on parameters used for credibility assessment. They had brought out a model for

assessment of credibility in different context and will help Arab people. Abbasi and Liu [9] have proposed a CredRank algorithm to examine the user credibility in online network based on the users behavior. In their research work, they had detected arranged behavior and mark that users less credibility score who are involved in these actions. Their proposed work will help to identify those personnel who have their accounts on multiple social networks and distribute wrong information using those accounts and ranking algorithm will help to save from rumors trap, bogus product feedbacks. For better result, they had suggested that all three; messages, source and media credibility should be taken into consideration. Kang [10] have proposed 14-components evaluation for credibility assessment of blog. They had taken two dimensions source and content for credibility assessment. Based on the observation results, authority and reliability were highly effective factor for blogger credibility and accuracy and focus are best indicators of message credibility. The proposed methodology is helpful to examine behavior effects on consumers of blog information.

From the study of above researchers work over OSN workspace, it is evident that to measure credibility, it is important to address all three dimensions message, source and media. In our proposed work, we have tried to include all three addressing comment, user and Facebook public pages and along with this validate our results using Machine learning algorithms in Weka.

3 Proposed Approach

In this section, we propose an approach to quantify the scope of distribution of unrelated content over public pages posts and to label each public page with a credibility score. We apply machine learning algorithms using Weka. Figure 1 shows the workflow of our proposed approach. Every step of flow chart is described in detail in the following subsections.

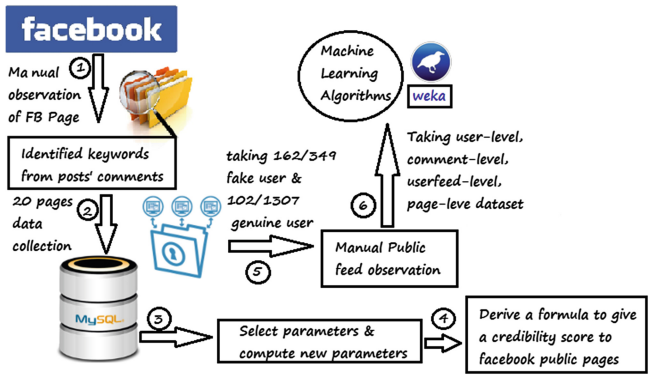

Fig. 1. Proposed approach

3.1 Manual Observation of Public Pages

We manually observed 56 public pages over Facebook. Out of them, advertisements activities were detected on 40 pages where it was found that users are placing advertisement comments which are unrelated to posts on those public pages. Even URLs are shared with text. Spammers are posting same comment over multiple pages. We have identified some patterns in comments for data collection. Here we have shown one such public page's post in Fig. 2 as an example of our problem:

Fig. 2. Unrelated content over Wikipedia

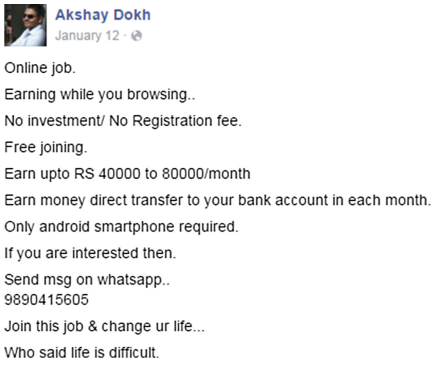

Fig. 3. Public feed example

3.2 Data Collection with and Without Identified Keywords

After keyword identification, we collected two types of comments. One type of comments contain those identified keywords and second type do not contain those keywords using Facebook Graph API v2.5 and stored all collected attributes for posts and comments in MySQL database. As some of comments are in *hinglish*, so for translation, we used Text Blob. Amount of data collected is shown in Table 1.

Identified Keyword Patterns in Comments: Here in Table 2, we have listed those keywords observed manually and taken into consideration during data collection:

Table 1. Data set with and without keywords

Data collection	#Pages	#Posts	#Posts per Page	#Comments	Duration
With keywords	20	1997	100	1523	Jan-Feb, 2016
Without keywords	20	1997	100	1626	Jan-March, 2016

Table 2. List of identified keywords

Identified keywords patterns
Good knowledge of the Internet, spiritual astrologer,
Unemployment in India, All World Famous Astrologer Tulsi Das,
Champcash, Many way to earn unlimited, play store,
Earntalktime, Android mobile, Minimum Recharge,
SPONSOR ID, Just download and install 8–9 Apps,
Level income, ONLINE JOB Digital India, Investment
Part time job, Online Home Based Job, Friends Refers Someone

3.3 Public Feed Observation of Fake and Genuine Users

To strengthen our results, we also analyzed the public feeds of both users who have posted the comments having identified keywords and those who have not. Observations of fake user's public feeds are shown in Table 3.

Table 3. Observations of public feed dataset of fake user

Observed points over fake user	Count
No of users with advertisement comments	349
No of users for whom public feeds are observed	162
Users commented present on more than one page	98 [28.08 %]

Observations of genuine users' public feed are shown in Table 4.

Table 4. Observations of public feed dataset of genuine user

Observed points over genuine user	Count
No of users with genuine comments	1307
No of users for whom public feeds are observed	102
Users commented present on more than one page	28 [2.14 %]

3.4 Selected Parameters

After data collection and manual observation of public feed of both fake and genuine users, we selected some parameters from our collected dataset and calculated some new parameters. All selected and calculated features on all levels are shown in following tables. Table 5 shows the features of comment.

Table 5. Comment level features

Features	Description
Is_Fake	Does comment contain spam keywords or not
Has_URL	Comment containing URL or not

All responses to users' comments like reply count, like count and other features like total comments per day by user, total urls in users' comment etc. are shown in Table 6.

Table 6. User level features

Features	Description
Total_Cmtcount	Total comment posted by user
Total_Likecount	Total Likes on their Comments
Total_Urlcount	URL posted In the comments
Totalcmt_Perday	Count of Comment posted by each user per day
Total_Replycount	Total reply to their comments
Avg_Likecount	Number of total like count over total comment count
Avg_Replaycount	Number of total reply count over total comment count
Avg_Urlcount	Number of total url count over total comment count

Amount of spam comments with fake users and genuine comments with genuine users on Facebook pages are shown in Table 7.

Table 7. Page level features

Features	Description
Totalcmt_Posted	Cumulative comment count on all collected posts
Spam_Cmt	Total number of Spam comment out of total comments
Genuine_Cmt	Cumulative genuine comment count on all collected posts
URL_Count	Cumulative url count on all collected posts
Fakeuser_Count	Cumulative fake user count on all collected posts
Genuineuser_Count	Cumulative genuine user count on all collected posts

Table 8 shows the details of users' feeds which are observed manually:

Table 8. User feed level features

Features	Formulas
Posted_Comment_ Count	Total comment posted by users
Comment_Type	What identified keyword, comment contain or not contain
IS_PublicFeed	Does user contain such unrelated content in their feed or not
Active_On_Facebook	Whether user is active on Facebook or not
Feed_Count	Total count of such spam feed
CmtCount_OnFeed	Total comment posted on such spam feed
LikeCount_OnFeed	Total like count on such spam feed
ShareCount_OnFeed	Total share count on such spam feed
DiffPageCount	Total no of public pages on which user posted comment

3.5 Formulation of Credibility Score for Public Pages

We propose a formulation (metric) to give credibility score to Facebook public pages. The algorithm takes as input all the 20 pages in the list PL and keywords list $Keylist$ based on which comments of one type are separated from others. It outputs corresponding *Credibility* score.

Algorithm. Find-Credibility-Score (PL, $Keylist$)

$ScalingFactor \leftarrow 10$
for all $Page_i \in PL$ **do**
  $spamcmtCount \leftarrow 0$
  $genuinecmtCount \leftarrow 0$
  $fakeuserCount \leftarrow 0$
  $genuineuserCount \leftarrow 0$
  $Post_i \leftarrow getPosts(Page_i)$
  for all $p_{ij} \in Post_i$ **do**
    $Comment_{ij} \leftarrow getComments(p_{ij})$
    $TotalComment \leftarrow getLength(Comment_{ij})$
    for all $c_k \in Comment_{ij}$ **do**
      if any (word in c_k for word in $Keylist$) **then**
        $spamcmtCount \leftarrow spamcmtCount + 1$
        $fakeuserCount \leftarrow fakeuserCount + 1$
      else
        $genuinecmtCount \leftarrow genuinecmtCount + 1$
        $genuineuserCount \leftarrow genuineuserCount + 1$
      end if
    end for
  end for
  $Score(Page_i) \leftarrow (\frac{spamcmtCount}{TotalComment} + \frac{fakeuserCount}{fakeuserCount+genuineuserCount}) * ScalingFactor$
end for

Score less than and equal to 0 shows page is very high credible and score value between 16 to 20 shows page is very low credible. Other labels like high, medium and low lie between these ranges.

4 Results and Observations

After applying credibility formula over pages data set, public pages are labeled as "Very High, "High, "Medium, "Low and "very Low credibility as shown in the following Table 9.

Table 9. Credibility label for public pages

Credibility score	Label	For pages
0 and <0	Very high	Star One, Fox, HBO India, Star World
1–5	High	Life OK, Inidiaforum.com, Channel [v] India
6–10	Medium	And TV, Zee Cinema, sony entertainment television, Zee business, Etc Bollywood, Star Plus, National Geograhic, Colors TV
11–15	Low	Star Gold, Bindass, 9XM, Zoom TV
16–20	Very low	Sony Max

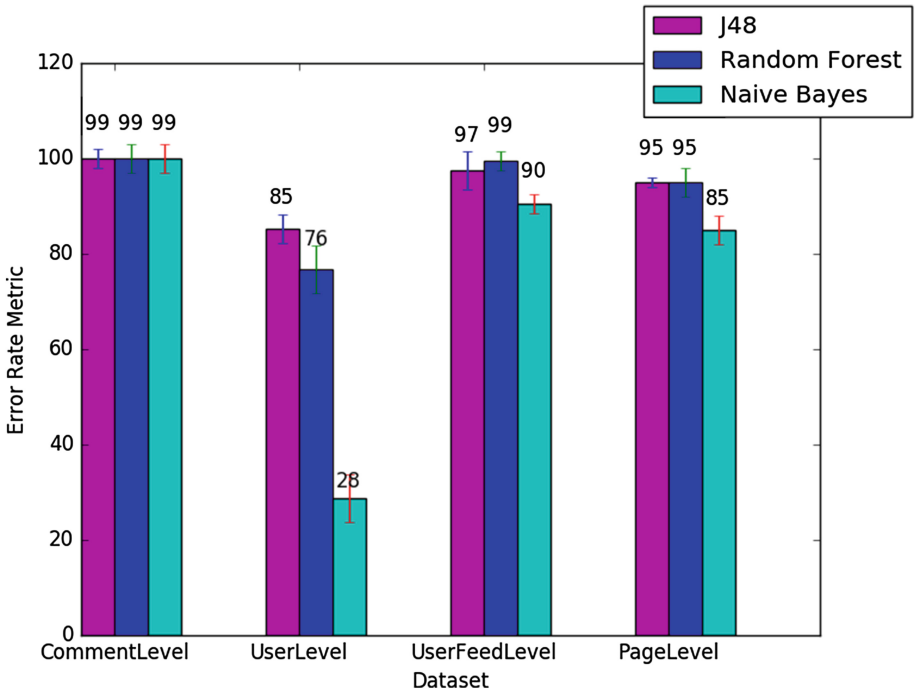

Fig. 4. Experimental result after execution of Machine Learning Algorithms

Using comment level, user level, userfeed level and page level features, we apply classifiers based on three machine learning algorithms namely J48, Naive Bayes and Random Forest using Weka to test our training dataset. Here we have shown the classification accuracy resulted from this experiment in the following graph.

4.1 Other Observations

Along with the above results, we have listed some of users (out of total 337 users) who had posted the same comments over different public pages in Table 10.

Table 10. Commenter on different public pages

User	On_diff_page
Abhishank Goyal	AndTV, ETC Bollywood, Bindass, Sony MAX, Zoom TV
Ravi Sharma Khajuria	AndTV, STAR Gold, Sony Entertainment Television, 9XM, National Geographic Channel
Vanita Gore	AndTV, STAR Gold, Sony Entertainment Television, ETC Bollywood, India-Forums.com, 9XM, Sony MAX, STAR Plus, Zoom TV
Anu Gaurav Gupta	ZEE Cinema, STAR Gold, Sony Entertainment Television, STAR Plus, Zoom TV
Atul Jain	Sony Entertainment Television, 9XM, Sony MAX, STAR Plus, Zoom TV

5 Conclusion and Future Work

Unsolicited and unrelated content over online social platforms have been analyzed in depth in the above research work. Using the collected data set of source (spammer), content (unrelated comment) and medium (public pages), user activities have been classified with maximum 85 % accuracy, user feed with 97.6 % accuracy, comments with 99 % accuracy as genuine or fraudulent and pages with 95 % accuracy as credible or not credible. But to use our analysis, we plan to build an application for Facebook that can show the credibility score for the public page.

References

1. Wang, J., Yu, C.T., Yu, P.S., Liu, B., Meng, W.: Diversionary comments under blog posts. ACM Trans. Web (TWEB) **9**(4), 18 (2015)
2. Dewan, P., Kumaraguru, P.: Towards automatic real time identification of malicious posts on Facebook. In: 2015 13th Annual Conference on Privacy, Security and Trust (PST), pp. 85–92. IEEE, 21 July 2015

3. Abu-Nimeh, S., Chen, T.M., Alzubi, O.: Malicious and spam posts in online social networks. Computer **12**(9), 23–28 (2011)
4. Metzger, M.J., Flanagin, A.J., Eyal, K., Lemus, D.R., McCann, R.M.: Credibility for the 21st century: integrating perspectives on source, message, and media credibility in the contemporary media environment. Commun. Yearb. **20**(27), 293–336 (2003)
5. Nguyen, D.T., Nguyen, N.P., Thai, M.T.: Sources of misinformation in Online Social Networks: who to suspect? In: 2012-MILCOM Military Communications Conference, pp. 1–6. IEEE, 29 October 2012
6. Lin, C., Huang, Z., Yang, F., Zou, Q.: Identify content quality in online social networks. IET Commun. **6**(12), 1618–1624 (2012)
7. Ismail, S., Latif, R.A.: Authenticity issues of social media: credibility, quality and reality. In: Proceedings of World Academy of Science, Engineering and Technology, vol. 74, p. 265. World Academy of Science, Engineering and Technology (WASET), 1 February 2013
8. AlMansour, A.A., Brankovic, L., Iliopoulos, C.S.: A model for recalibrating credibility in different contexts and languages-a twitter case study. Int. J. Digit. Inf. Wirel. Commun. (IJDIWC) **4**(1), 53–62 (2014)
9. Abbasi, M.-A., Liu, H.: Measuring user credibility in social media. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 441–448. Springer, Heidelberg (2013)
10. Kang, M.: Measuring social media credibility: a study on a measure of blog credibility. Inst. Public Relat. 59–68 (2010)
11. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference. ACM (2010)
12. Stein, T., Chen, E., Mangla, K.: Facebook Immune System in European Conference on Computer System (EuroSys) (2011)
13. Robertson, M., Pan, Y., Yuan, B.: A social approach to security: using social networks to help detect malicious web content. In: 2010 International Conference on Intelligent Systems and Knowledge Engineering (ISKE). IEEE (2010)
14. Cvijikj, I.P., Michahelles, F.: Monitoring trends on facebook. In: 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC). IEEE (2011)
15. Abu-Nimeh, S., Chen, T.M.: Proliferation and detection of blog spam. IEEE Secur. Priv. **8**(5), 42–47 (2010)
16. Agrawal, H., Kaushal, R.: Analysis of text mining techniques over public pages of Facebook. In: 2016 IEEE International Advance Computing Conference(IACC). IEEE (2016)