Expert knowledge for automatic detection of bullies in social networks

Maral Dadvar Dolf Trieschnigg Franciska de Jong

Human Media Interaction Group, University of Twente,
POBox 217, 7500 AE, Enschede, The Netherlands

Abstract

Cyberbullying is a serious social problem in online environments and social networks. Current approaches to tackle this problem are still inadequate for detecting bullying incidents or to flag bullies. In this study we used a multi-criteria evaluation system to obtain a better understanding of YouTube users' behaviour and their characteristics through expert knowledge. Based on experts' knowledge, the system assigns a score to the users, which represents their level of "bulliness" based on the history of their activities, The scores can be used to discriminate among users with a bullying history and those who were not engaged in hurtful acts. This preventive approach can provide information about users of social networks and can be used to build monitoring tools to aid finding and stopping potential bullies.

1 Introduction

Cyberbullying is an alarming problem among teenagers and adolescents. Cyberbullying comes in different forms and through a variety of modalities depending on the characteristics of the people involved. Bullying can happen through public postings of a private picture without the consent of the person(s) in the picture, it can happen through vulgar text messages and threatening phone calls, and one of the most common ways is by posting nasty and hateful comments about someone in social networks such as YouTube. As negative effects caused by bullying among youths are increasingly reported, an increasing number of studies is dedicated to dealing with cyberbullying in online environments [1-2]. However, cyberbullying remains a growing concern and the existing approaches are still inadequate. Most current approaches towards cyberbullying detection concentrate on the content of textual comments. By applying conventional sentiment analysis techniques they try to identify cyberbullying incidents [3-5]. However, these approaches fail to incorporate personal and contextual features for the users involved. (From now on we refer to "user" as someone who is active in social networks, and uses these networks for communication, entertainment and other purposes.)

Cyberbullying is a multi-dimensional problem and environmental characteristics (for example, the degree to which a person is active in social networks) and personal features (for example, age) influence how cyberbullying takes place. Extra information, other than what can be extracted from online comments can be of added value to improve the accuracy of bullying incidents detection. There is ample information available: users' characteristics and intentions can be traced and recognized in their writings and online activities However, this information might not be transparent enough and not all human characteristics can be extracted in the form of a set of features that can be interpreted by machine learning models. In contrast, deductive approaches, such as expert systems, can benefit from more sources of information and can offer an analysis based on elements conveying information about human characteristics that goes beyond online activity patterns.

A candidate approach that integrates human reasoning and experts' opinions is a so-called Multi-Criteria Evaluation System (MCES). A MCES is a deductive approach that combines different sources of information, to make a decision among alternatives. Multi-criteria evaluation systems are commonly applied in variety of evaluation and decision support system research fields [6-7], information retrieval [8], and measurement of document content reliability [9]. In this paper we propose to integrate expert knowledge into a multi-criteria evaluation system for cyberbullying detection. By the trust commonly put in human knowledge, this may bring a human touch to a problem that is rooted in human behaviour .This would overcome the limitation inherent to approaches based on machine learning, namely that it is often difficult to understand why a machine learning algorithm gives certain results.

In this study we focus on measuring the degree of "bulliness" of YouTube users. The degree of bulliness represents the extent to which a user can be expected to act as a bully. We hypothesize that expert knowledge can be used to assign a score to a YouTube user which represents the likeliness of future bullying behaviour of that user. YouTube is considered to be a sample of the general internet population in terms of the audience demographics [10]. There is a broad audience from different age and gender groups. This makes this network comparable to real life situations for investigating the interaction among users. The history of activities and characteristics of a user in YouTube comprise the elements of a MCES. Experts provide input to the system to weigh and assign the values of these criteria. The system will compile the information collected from different sources (experts) to generate a final user score that indicates the degree of bulliness of that user.

As will be explained in the next section, throughout the stages of this study we use expert knowledge in order to have more understanding about users' behaviour and to select the features that can represent user's characteristics and convey information about their intentions. To our knowledge this is the first time that a MCES is used for cyberbullying detection in social networks. In the next section we describe the dataset and feature space as well as the proposed approach and the procedure for collecting insights from a panel of experts. The results, discussion and future work are presented in sections 3 and 4.

2 Methods and Materials

In this section we will explain the MCES that we have used in our experiment in more depth. The dataset, the feature space as well as the structure of the expert panel and the process of expert knowledge elicitation will also be explained.

2.1 Multi-Criteria Evaluation System (MCES)

MCES is a commonly used technique for decision-making in business, industry, and finance [11]. MCES provides a framework for combining a variety of features and ranking the alternatives based on criteria set by a group of experts. It provides a methodology for experts to compensate for the uncertainty in their knowledge as well as inconsistency between themselves. In this study we used MCES to: (1) rank the importance of the features that we extracted for users in YouTube, (2) standardize and set criteria, and (3) combine the criteria based on the knowledge of experts to evaluate the features for a user. This technique can be applied to combine different sources of knowledge and information to reach a final decision on the bulliness score of a user. Feature values (such as age) reflect the characteristics of a user in YouTube. Absolute values of each feature (F_i i=1…n) correspond to the numerical values of each criterion (C_i). To assign such correspondences we asked m experts (E_j j=1…m) to set criteria (C_i i=1…n) for each feature. For example expert E_j sets criterion C_i by assigning a likelihood value to the feature Fi under specific conditions. For example the criterion C_j is, if the age of the user (F_i) is between 13 and 16 years old, then it is "very likely" that the user is a bully. The experts also assigned weights (W_i i=1…n) to each feature to indicate its relative importance. The score of each user is calculated by taking the weighted average of the criteria.

2.2 Dataset

As the dataset for our study, figures on the user behaviour for YouTube are used. YouTube is the world's largest user-generated content video system [10]; 60 hours of video are uploaded every minute, and over 4 billion videos are viewed every day [12]. Based on YouTube statistics, most of the users are between 18 to 24 years old and its usage among men and women is the same. Obviously also negative behaviour is part of this communication and interaction platform. The videos and the comments trigger bullies to victimize their targets through harassing comments or other disturbing misbehaviours. Despite the fact that the owners of YouTube videos have the possibility to remove offensive comments from the site, most of the comments are not moderated. Therefore, YouTube comments can be used as a source for cyberbullying studies. For our study we sampled 3825 users who had commented on the top 3 videos of the YouTube video categories. YouTube users can carry out different actions, such as posting comments, responding to comments, subscribing to users' channels (i.e. the personal page of the users that shows their activities), uploading videos and, liking and disliking other users' actions. We collected a log of the users' activities for a period of 4 months (April – June 2012). We also captured profile information of the users, such as their age and the date they signed up. We could not access private profiles and private information from the profiles, such as private addresses. In total there are 54,050 comments in our dataset. On average there are 15 comments per user (StDev= 10.7, Median = 14). The average age of the users is 24 with 1.5 years of membership duration. While 38.2% of the users have uploaded less than 10 videos, 1.3% has uploaded more than 100. About one third of the users have no subscriptions while 56% have less than 20, and 1% more than 500 subscriptions.

2.3 Feature Space

Based on a literature review on cyberbullying and social behaviour factors [13] [14], and consultation of domain experts, we compiled a set of 11 features in three categories to identify bullying users. The selection was limited to what is technically possible to extract from YouTube. We grouped the features in the three categories, representing the characteristics, actions and behaviour of the users, respectively (also see Table 1).

Table 1. The summary of feature sets and the units in which they have been presented

Feature Set	Feature Name	Unit	Details
User features	F1 Age	Categorial	Categories; 13-16, 17-19, 20-25, 26-30 and above 30 years old.
	F2 Membership duration	Categorial	Categories; less than 1 year, 1-3 years, More than 3 years.
Content features	F3 Length of the comments	Numerical	Average in Youtube: 12 words
	F4 Profane words in the username	Boolean	True, False
	F5 Profanities and bullying sensitive topics	Numerical	Average per comment in YouTube: 1.2 %
	F6 Second person pronouns	Numerical	Average per comment in YouTube: 2.3 %
	F7 First person pronouns	Numerical	Average per comment in YouTube: 2.2 %
	F8 Non-standard spellings	Numerical	Average per comment in YouTube: 21.5 %
Activity features	F9 Number of uploads	Numerical	Average in YouTube: 4.56
	F10 Number of subscriptions	Numerical	Average in YouTube: 23
	F11 Number of posted comments	Numerical	Average in YouTube: 14.4

- *User features* - This set consists of the personal and demographic information derived from the users' profiles: (F1) The age of the users, divided in 5 age groups: 13-16, 17-19, 20-25, 26-30 and above 30 years old; Cyberbullying differs across different age categories. Frequency of bullying incidents as well as choice of words and language structures change in different age groups. The youngest age at which users can sign up in YouTube is 13 years old and the categories correspond to educational level. We assume that the provided information in the user profile is correct, but we are aware of the fact that this might not be the case. (F2) The membership duration of the users, divided into 3 groups: less than 1 year, 1- 3 years and above 3 years.

- *Content features* - These features are derived from the content of the user comments. This category of features pertains to the writing structure and usage of specific words. (F5) The number of profane words in the comment based on a dictionary of profanities, normalized by the total number of words in the comment. The dictionary consists of 414 profane words including acronyms and abbreviation of the words. The majority of the words are adjectives and nouns. To identify frequent bullying topics such as minority races, religions and physical characteristics we also added a manually compiled set of cyberbullying words to our dictionary [1]. (F3) The length of the comments, which is relevant information as bullying comments are typically short. To detect the comments which are personal and targeting a specific person, we included the (F7) normalized number of first person pronouns and (F6) second person pronouns in the comment. (F4) Usernames containing profanities; YouTube users can choose their username to be their real name and/or surname or can choose any other aliases and combinations of symbols and words. We believe that it is more likely that users with bad intentions would hide their real identity; (F8) Non-standard spelling of the words in the users' comments. This includes misspellings (e.g. 'funy' instead of 'funny'), or informal short forms of the words that are used in online chats and posts (e.g. 'brb' which means 'be right back').

- *Activity features* - This set of features helps to determine how active the user is in the online environment. One of the common activities of the users is to upload videos. These can be home videos provided by users themselves, or videos made by others. A user can also post comments on uploaded videos as well as on other users' comments. Most of the YouTube users have a public channel, in which they upload their videos and in which their activities such as posted comments can be viewed. Users can subscribe to others channels and follow the activities of the owner of the channel if they find it interesting. In this feature set we consider (F9) number of uploads, (F10) number of subscriptions and (F11) number of posted comments.

2.4 Expert Panel

A panel of twelve experts in the area of cyberbullying was convened. The experts have a background in psychology, social studies and communication sciences. The majority of the panel works on cyberbullying causes, effects and solutions from social and psychological perspectives. A smaller number works on social behaviour, psychology and communication studies. During a preliminary meeting the purpose of this survey was explained. The experts completed the survey individually.

2.5 Expert Knowledge Elicitation

Experts were provided with an online questionnaire, consisting of 22 factual questions. To avoid ambiguities, each question also provided a brief definition of the concepts addressed. For each of the features experts were asked to express their opinion on the likelihood that a bully user belongs to a certain category relevant for that feature. For example, "What is the likelihood that a bully user belongs to the following age categories?" where the age categories are given in the question. In this type of questions, experts could express their opinion through a four-point scale answering options; 'Unlikely', 'Less likely', 'Likely' and 'Very likely' corresponding to values 0.125, 0.375, 0.625 and 0.875 respectively [15]. The 'I don't know' option was also available. Experts could provide comments at the end of each question. To understand how informative and helpful the features are in the determination of personality and potential behaviour of a user we also asked experts to weigh the features. The experts' choices were corresponded to values of 1: not informative, 2: partially informative, 3: informative and 4: very informative. The questionnaire required approximately 20 minutes to be completed. It took about three weeks to receive all the responses from the expert panel.

2.6 Evaluation

To measure the consistency of the elicited knowledge from the expert panel, the assignments were analysed in terms of overall disagreement among experts. For each expert, we compared the value that the expert had assigned to each criterion, to the 'median value ±1' of that criterion. If the assigned value was out of this range, it was considered as a different opinion and therefore a disagreement. The final disagreement rate was calculated by taking the ratio of the total number of disagreements to the total number of opinions expressed by each expert on all the criteria.

To evaluate the performance of the proposed MCES approach, we randomly selected one comment per user (n=3825), and manually labelled them. We assumed that any user with at least one bullying comment in our dataset is a bully. Two PhD students independently labelled the comments as bullying and non-bullying based on the definition of cyberbullying in this study. We then compared their labels (inter-annotator agreement = 93%, Kappa = 0.78) and the comments which both students had labelled as bullying, were marked as bullying (9.7% of the posts). The comments for which there was a disagreement were discarded. Using this dataset, we evaluated the discrimination capacity of our model by analysing its receiver operation characteristic (ROC) curves. A ROC curve plots "sensitivity" values (true positive fraction) on the y-axis against "1–specificity" values (false positive fraction) for all thresholds on the x-axis [16]. The area under such a curve (AUC) is a threshold-independent metric and provides a single measure of the performance of the model. AUC scores vary from 0 to 1. AUC values of less than 0.5 indicate discrimination worse than chance; a score of 0.5 implies random predictive discrimination; and a score of 1 indicates perfect discrimination.

3 Results

Figure 1 illustrates the importance of each feature based on the weights that were assigned to them by the experts. Based on the results, profanities and bullying sensitive topics in the history of a user's comments is the most informative feature (average weight equals 3.6). The second and third most informative features are the inclusion of profanities in usernames (average weight equals 3.2) and age (average weight equals 3) respectively. The least informative feature is the number of non-standard spellings in the history of users' comments (average weight equals 1.7). An average likelihood was also assigned to the features and their subcategories in each feature set. How high or low the value of a certain feature is, was measured in comparison to the average value of that feature in YouTube. The experts' choices on the likelihoods, was taken to correspond to values of 0.125 (not likely), 0.375 (less likely), 0.625 (likely) and 0.875 (most likely). The overall rate of disagreement among experts was 0.05.

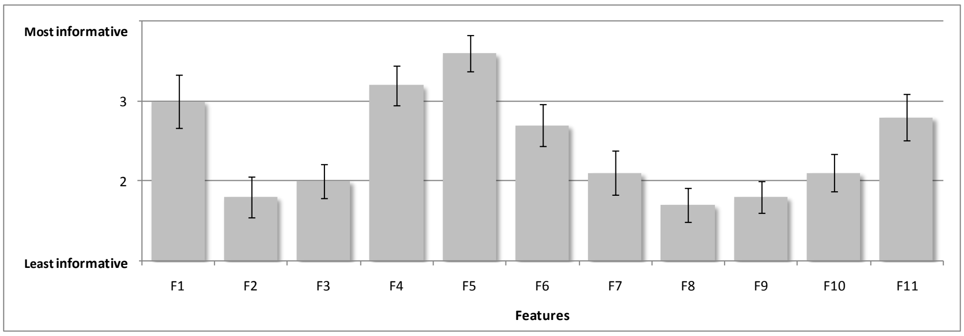

Figure 1. Feature weights indicated by experts

The age range between 13 and 16 years is indicated to be most likely to contain bullies. The age category above 30 years old is ranked as corresponding to the lowest bulliness likelihood. Experts indicate that a typical bully has a membership period shorter than 1 year. The outcome of the questionnaire also indicate that it is more likely that bully users have a high ratio of second person pronouns in their comments as well as profane words in their usernames. These later two features are ranked as the two most likely features in the Content features set. Moreover the likelihood of writing short and right-to-the-point comments is higher than the likelihood of lengthy ones. In the Activity features set, a high number of posting comments has the highest likelihood in comparison to the other features. The number of uploads comes in the second place in this set. See Table 2 for detailed results and the likelihood of each feature and its subcategories.

Table 2. List of features subcategories and their likelihoods based on experts' opinion

Feature	Rule		Likelihood	Std
F1	R1-1	IF 13 < F1 < 16 THEN	0.725	0.242
F1	R1-2	IF 17 < F1 < 19 THEN	0.597	0.232
F1	R1-3	IF 20 < F1 < 25 THEN	0.431	0.167
F1	R1-4	IF 26 < F1 < 30 THEN	0.344	0.088
F1	R1-5	IF F1 > 30 THEN	0.268	0.134
F2	R2-1	IF F2 < 1 THEN	0.525	0.224
F2	R2-2	IF 1 < F2 < 3 THEN	0.475	0.224
F2	R2-3	IF F2 > 3 THEN	0.275	0.137
F3	R3-1	IF F3 < Average THEN	0.688	0.222
F3	R3-2	IF F3 > Average THEN	0.375	0.000
F4	R4-1	IF F4 = True THEN	0.700	0.237
F4	R4-2	IF F4 = False THEN	0.225	0.129
F5	R5-2	IF F5 < Average THEN	0.375	0.000
F5	R5-2	IF F5 > Average THEN	0.688	0.222
F6	R6-1	IF F6 < Average THEN	0.339	0.173
F6	R6-2	IF F6 > Average THEN	0.732	0.197
F7	R7-1	IF F7 < Average THEN	0.375	0.000
F7	R7-2	IF F7 > Average THEN	0.375	0.000
F8	R8-1	IF F8 < Average THEN	0.375	0.000
F8	R8-2	IF F8 > Average THEN	0.486	0.182
F9	R9-1	IF F9 < Average THEN	0.500	0.231
F9	R9-2	IF F9 > Average THEN	0.375	0.125
F10	R10-1	IF F10 < Average THEN	0.458	0.204
F10	R10-2	IF F10 > Average THEN	0.417	0.102
F11	R11-1	IF F11 < Average THEN	0.236	0.132
F11	R11-2	IF F11 > Average THEN	0.725	0.211
F1 / F5	Rx1-5	IF F5 > Average AND F1 > 30 THEN	0.675	NA
F13 / F8	Rx8-2	IF 13 < F1 < 16 AND F8 > Average THEN	0.125	NA

The experts also had the opportunity to give additional comments after each question. Some of the experts combined existing criteria into new criteria. One of the experts indicated, for example, that if the age of the user is above 30, then it is very likely that the profanity use in his/her comments is for harassing purposes. Another expert stated that a high number of non-standard spellings in younger ages, 13 to 16, is not as alarming as the use of slang and misspellings in ages above 30. First we calculated the scores considering all equal weights for criteria; the highest score was 0.65 while the lowest was 0.32. When we applied the weights and the highest score was 0.71, and the lowest was 0.29. In the last step of our experiment, we replaced some of the criteria with their corresponding combined criteria. The highest score increased to 0.75 and the lowest remained the same. As illustrated in Figure 2, the discrimination capacity of the model was 0.71 and improved to 0.72 when the weights of the features were also taken into account. When taking into account the suggested combined criteria, the discrimination capacity was improved to 0.74.

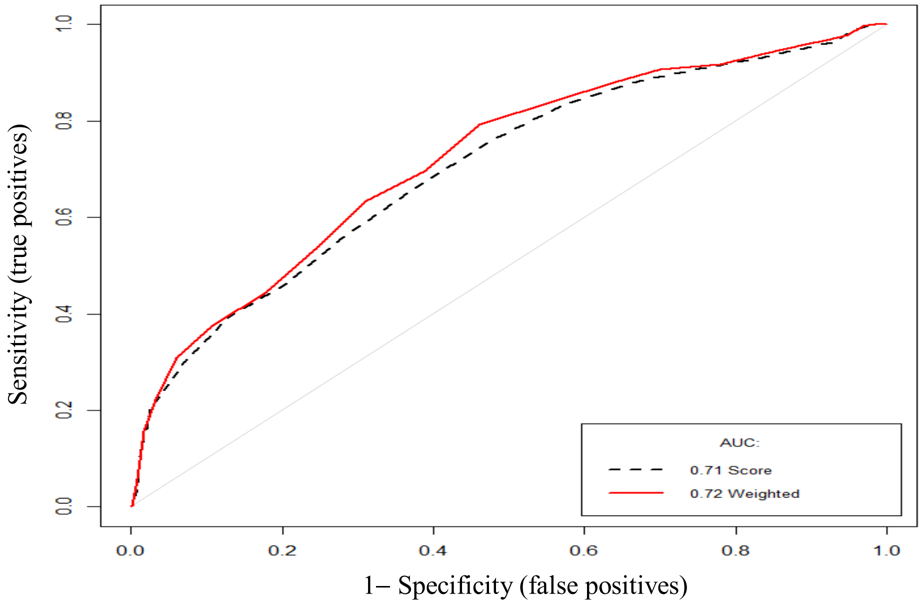

Figure 2. ROC curves of the estimated users' score

4 Discussion

In this experiment we built a MCES to assign a bulliness score to YouTube users. We employed experts' knowledge to define the weights and to set the criteria of the system. The final scores were discriminative for users with a bullying history and those who had not been engaged in hurtful interactions. Our proposed approach has a human touch as it makes use of human knowledge and experts' experience. The MCES approach is flexible towards inconsistencies among different sources of information and can be easily fine-tuned by adding specific criteria to identify forms of human behaviour that are hard to capture by less well understood models. Moreover, the proposed approach easily helps considering context specific criteria in a natural way, rather than using complex formulas which are difficult to interpret. It may contribute to identify social network users who may act in a hurtful way in earlier stages, as systems based on this approach can be used as a preventive tool for network administrators and moderators to stop these users from causing any further harm. Although according to the experts' opinion, the importance of features are different and experts have assigned different weights to features, our results show that the system did not seem to improve significantly after considering the weights of the features. This can be due to the disagreement among experts on the weights. This disagreement may have neutralized the effect that weights should have had on the results.

The proposed approach can be used in other social networks as well. Depending on the network under study and the design of the platform, the activity features may vary. For example if the study is on Twitter, the number of followers can be selected as one of the activity features. Given this variation the advantage of our approach is that the questionnaire can be easily updated by adding the extra features and there is no need to develop any training data to train a new model. Our approach is language-independent and is adaptable to other languages by modifying the dictionaries. We can also outline another advantage of the proposed approach since we can easily carry out the study from the experts' perspective by setting criteria according to their experiences, giving rise to a valuable aid. As suggested in the questionnaire by some experts, having more combined criteria may improve the accuracy of the results. For example, one expert argued that it is not sufficient to only know how active one user is, but we have to study its frequency and changes over time, as the level of activities or harassing behaviour may vary in time. In future work we would like to also compare this method with a machine learning method based on the same feature sets and study how the outcome would differ.

References

[1] M. Dadvar, D. Trieschnigg, R. Ordelman and F.M.G. de Jong, “Improving Cyberbullying Detection with User Context,” Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 7814, pp. 693–696, 2013.

[2] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, “Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying,” ACM Transactions on Interactive Intelligent Systems, vol. 2, no. 3, pp. 1–30, Sep. 2012.

[3] K. Dinakar, R. Reichart, H. Lieberman, “Modeling the detection of textual cyberbullying,” Proceedings of International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain, 2011.

[4] M. Dadvar, and F.M.G. de Jong. “Cyberbullying detection: a step toward a safer Internet yard,” Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012.

[5] K. Reynolds, A. Kontostathis, and L. Edwards, “Using Machine Learning to Detect Cyberbullying,” 2011 10th International Conference on Machine Learning and Applications and Workshops, pp. 241–244, Dec. 2011.

[6] D. Y. Shee and Y.-S. Wang, “Multi-criteria evaluation of the web-based e-learning system: A methodology based on learner satisfaction and its applications,” Computers & Education, vol. 50, no. 3, pp. 894–905, Apr. 2008.

[7] H. Jiang and J. Eastman, “Application of fuzzy measures in multi-criteria evaluation in GIS,” International Journal of Geographical Information Science, no. May 2013, pp. 37–41, 2000.

[8] M. Farah and D. Vanderpooten, “A multiple criteria approach for information retrieval,” String Processing and Information Retrieval, pp. 242–254, 2006.

[9] C. W. Bong, D. W. Holtby, and K. S. Ng, “Fuzzy Multicriteria Decision Analysis for Measurement of Document Content Reliability,” 2012 Fifth International Symposium on Computational Intelligence and Design, pp. 303–306, Oct. 2012.

[10] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, “I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system,” Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 1–13, 2007.

[11] J. Figueira, S. Greco, and M. Ehrgott, “Multiple criteria decision analysis: state of the art surveys,” Vol. 78. Springer, 2005.

[12] YouTube press statistics, http://www.youtube.com/t/press_statistics, [May 2013].

[13] M. Marcelo, J. Almeida, and V. Almeida, “Identifying user behavior in online social networks,” Proceedings of the 1st workshop on Social network systems. pp. 1-6, ACM, 2008.

[14] W. Christo, B. Boe, A. Sala, K. PN Puttaswamy, and B. Y. Zhao, “User interactions in social networks and their implications,” Proceedings of the 4th ACM European conference on Computer systems, pp. 205-218. ACM, 2009.

[15] X. Zhiwei, T. M. Khoshgoftaar, and E. B. Allen, “Application of fuzzy expert systems in assessing operational risk of software,” Information and Software Technology 45.7, pp.373-388, 2003.

[16] H. Fielding, and B. F. John, “A review of methods for the assessment of prediction errors in conservation presence/absence models.” Environmental conservation 24, no. 1, pp. 38-49, 1997.