Content Filtering of Social Media Sites Using Machine Learning Techniques

U. S. Tambe a, N. R. Kakad a, S. J. Suryawanshi a and S. S. Bhamre a
a Dept of IT, MVPS’s KBT College of Engineering, Nashik, Maharashtra, India

Abstract. To build a social network or social relations between people, we use social networking platforms like Facebook, Twitter, apps, etc. Using this media, users can share their views and opinions about a particular thing. Many people use their media for personal interests, entertainment, the market stocks, or business purposes. Nowadays, user security is the major concern for social networking sites. Online social networks give a little bit of support regarding content filtering. In this article, we proposed a system that provides security regarding malicious content that is posted on their social networking sites. To filter the content that might be unwanted messages, labeled images, or vulgar images, we proposed three level architecture. The user can use the auto-blocking facility as well.

Keywords. Social wall, Machine Learning Techniques, Message Filtering, Labeled Image Filtering, Image Filtering

1. Introduction

In today's life, social media plays a very important role. People spend most of their time on social media for chat and sharing their ideas. Using these media, people can share their information or exchange several types of content such as images, videos, textual messages or audio messages. Many people comment on this shared content. The peoples receive feedback on whatever material he or she has shared on the wall. Sometimes it gives a positive response, a negative response, or suggestions which are very useful for improvement. According to Face book, users make up 90 bits of content each month, though; more than 30 billion bits of content (web links, news stories, blog posts, notes, photo albums, etc.) are shared each month [1].

Users can post any type of content on social media. Unwanted text messages, branded photographs, obscene, pornographic images, personal mischievous comments, and so on are examples. Other users can see these posts and comment on them. The user's social image can be harmed as a result of this message. As a result, the protection of such a user's wall is critical. Up to a point, Face book provides protection. Only a selected group of people on Face book have access to other people's walls, such as friends, friends of friends, or established groups of friends. The user has the option of locking his profile picture. However, because content-based filtering is not supported, such unwanted messages cannot be prevented. The goal of present system is to protect user walls by filtering out unwanted content and user social

1U. S. Tambe, Assistant Professor, Department of IT, MVPS’s KBT College of Engineering, Nashik Email: tambe.ujwala@kbtcoe.org.
media images. Users can modify the filtering rules as per his/her choice. The user has control over who can send messages to their wall. For text filtering, a short text classification technique is used [2]. For training, a vocabulary list is used. The vocabulary list contains words and its class. For labeled image filtering, we used the OCR algorithm. OCR can extract the text from an image and save that text into a text file. Then, we match the extracted data with the dataset. If a match is found, then that image will be filtered by the system. For image filtering, we used the skin sheriff technique, which consists of skin detection and pornography detection algorithm. The skin detection algorithm can extract all the skin areas from the image and label the pixel that classified as skin as grey and non-skin as white. A pornography detection algorithm is used to calculate the maximum area of non-skin pixel. If it exceeds the maximum value, the image will be vulgar.

2. Literature Survey

2.1 Content-based Filtering:

In a Content-based filtering system, a document is approved by comparing the document profile with the user profile using information retrieval techniques like Term Frequency and Inverse Document Frequency (TF-IDF) [3]. As per the user's prior feedback and choices, user characteristics were gathered and profiled. The system needs an item-to-item relationship which helps to advise the document to the user. The system begins by assembling content information about the item, such as behaviors, indications, and so on, for the disease-related item. The system will then put the user to the test in order to rate the item. The system then compares the unrated item to the user profile item and assigns a score to the unrated item, resulting in the user being presented with items that are graded consistent with the score assigned.

2.2 Collaborative Filtering:

Collaborative filtering is a technique which will filter information supporting the user's interest (i.e., history) and therefore the ranking of other users with an equivalent interest. This filtering technique works on an outsized group of individuals to look and find a smaller set of users with related interest. It creates a ranked list of suggestions. It follows the things which are liked by users and combines them to make a ranked list of suggestions. It's widely utilized in many filtering systems or recommender systems, especially in e-commerce applications. Samples of such programs are Amazon.com, YouTube, eBay, etc., where recommendations for new products are given to a user supported by the user's past shopping history, likes and dislikes of comparable users.

2.3 Policy-based Personalization of OSN Contents:

In OSN, there have recently been some operator classification frameworks to personalize entry. Several classification methods for short text messages were proposed to a large number of users of micro blogging services [4]. Twitter, the device focuses on a community of categories and provides content updates. The consumer
would then see those types of tweets that support his or her interests. Kuter and Golbeck, on the other hand, have created a film trust application that uses OSN relationships of trust and origin information to customize location access. However, these systems lack a policy layer of filtering that allows the user to see the results of the classification process and choose how to filter out the unwanted data.

Table 1. Related Work

| Related Work | Objective |
|--------------|-----------|
| J. Golbeck and Kuter[5] | The Film Trust application claims that the OSN title deed relationship and provenance information make it simple to access the website in this post. Such systems lack a filtering policy that allows the user to use the results of the sorting process to decide how to filter and how much unnecessary data to exclude. |
| P. J. Hayes, P. M. Andersen, I. B. Nirenburg and L. M. Schmandt[9]. | The text categorization approach is used in this article. It means that a TCS run-time framework and a set of rules are included in the TCS text sharing programme. The legal structure establishes what classifications you should give documents in the text, as well as the rules that govern how you classify texts into different categories. |
| N. J. Belkin and W. B. Croft[10]. | Unlike database applications, which use highly structured data such as an email, this method of data filtering uses informal or partially structured data. The email format varies depending on where it is sent, but the email body contains random data. |
| P. W. Foltz and S. T. Dumais[11] | The method of analyzing structured data is discussed in this paper. By evaluating the user's response and keywords given by the user, this approach determines the user's interest. Latent Semantic Indexing and keyword matching methods are used. The same choices are considered for this person. |
| Zelikovitz and Hirsh[12] | The author of this paper intends to enhance the separation of short text strings by enhancing sub-reading, a technique that combines labeled training data with a second collection of unwritten but longer linked texts. |
| Bobicevand Sokolova[13] | This paper focuses on using a mathematical learning approach that works well without feature engineering to avoid the problem of error-making functionality. The method used to develop and generate a language model used in classifiers of hard text and not easily integrate soft, multi-membership paradigms is described in this paper. |

3. Proposed System:

In this section, we introduce a Proposed System Architecture to filter unwanted posts on the user wall. There are three levels in the current system, as shown in figure 1. The user interface is provided by the Social Network Manager Layer. External
activities are supported by the Social Network Applications layer. The third layer is graphical user interfaces (GUIs), which are used to display the output. Users communicate with the system through a graphical user interface (GUI) to set up and manage their FRs/BLs (Blacklists).

As shown in figure 2, it describes the layers of the proposed system. The User GUI is used to post messages which can be in any form like text message, labeled image or image. An Auto-blocking facility is provided for an authenticated user.
To post a message on the wall, a user must first navigate to the wall, which is divided into three sections: text filtering, labeled image filtering, and image filtering.

The data of the content is retrieved using a text-based classifier for text filtering. If a match is found, the material will be hidden from view on the wall.

Image Filtering: With the help of the Skin detection algorithm and pornography detection algorithm, we filtered unwanted images from the wall. If an image is found in the dataset, then it is not displayed on the wall.

In labeled image filtering, we used both previous methods. We can extract the text from an image using OCR techniques and apply message filtering rules.

According to this logic, the output will be filtered and published on the wall.

4. Blacklist Management and Filtering Rules

We describe blacklist management and filtering rules here.

4.1 Auto blocking

If a particular user continuously posted unwanted messages on the wall, we can block that person.
4.2 Filtering rules

Users can specify which content should be blocked or displayed on the filtered wall using filter rules. Filter rules are specified according to the user profile and use of social media. An author is a person who defines the rules. OSN users are denoted by Creator Spec, and Content Spec is a Boolean expression.

5. Result

Figure 4 shows the Post Comment Page. The user can select the friend’s wall and post the content, such as text messages, labeled images or images.

![Figure 4. Post Comment Page](image)

Figure 5. Shows the list of unwanted messages and blacklisted users.

![Figure 5. Shows the list of unwanted messages and blacklisted users.](image)
In this paper, we presented a system to filter out unwanted messages, labeled images, and unwanted images from the user wall. The system uses text classifiers and machine learning techniques to filter text messages. For filtering labeled images, we have used OCR techniques. With the help of a pornography detection algorithm, we can detect vulgar images. Our system also provides an auto-blocking system. So only authorized users can post the messages on the user wall.
References

[1]. Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, Moreno Carullo Department of Computer and Communication, University of Insubria "a system to filter out unwanted messages walls OSN user" IEEE Transactions on Knowledge And Engineering Flight Data: 25 the Year 2013

[2]. F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002. J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg, "Predicting Positive and negative links in online social networks," inProc. 19th Int. Conf. World Wide Web, 2010, pp. 641-650.

[3]. Vinaitheerthan Renganathan1*, Ajit N Babu2 and Suptendra Nath Sarbadhikari3 “ A Tutorial on Information Filtering Concepts and Methods for Bio-medical Searching".

[4]. B. Sriram, D. Fuhr, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in Twitter to improve information filtering," in Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, 2010, pp. 841–842.

[5]. J. Golbeck, “Combining provenance with trust in social networks for semantic web content filtering,” in Provenance and Annotation of Data, ser. Lecture Notes in Computer Science, L. Moreau and I. Foster, Eds. Springer Berlin / Heidelberg, 2006, vol. 4145, pp. 101–108.

[6]. Pennock DM, Horvitz E, Lawrence S, Giles CL (2000) Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence (UAI-2000), Morgan Kaufmann Publishers Inc 473-480.

[7]. Boger Z, Kuflik T, Shapira B, Shoval P (2000) Information filtering and automatic keyword identification by artificial neural networks. Proceedings of the 8th European Conference on Information Systems.

[8]. Jennings A, Higuchi H (1993) A user model neural network for a personal news service. User Modeling and User-Adapted Interaction 3: 1-25.

[9]. P. J. Hayes, P. M. Andersen, I. B. Nirenburg, and L. M. Schmandt, "Tcs: a shell for content-based text categorization," in Proceedings of 6th IEEE Conference on Artificial Intelligence Applications (CAIA-90) IEEE Computer Society Press, Los Alamitos, US, 1990, pp. 320–326.

[10]. N. J. Belkin and W. B. Croft, “Information filtering and information retrieval: Two sides of the same coin?” Communications of the ACM, vol. 35, no. 12, pp. 29–38, 1992.

[11]. P. W. Foltz and S. T. Dumais, “Personalized information delivery: An analysis of information filtering methods," Communications of the ACM, vol. 35, no. 12, pp. 51–60, 1992.

[12]. S. Zelikovitz and H. Hirsh, “Improving short text classification using unlabeled background knowledge,” in Proceedings of 17th International Conference on Machine Learning (ICML-00), P. Langley, Ed. Stanford, US: Morgan Kaufmann Publishers, San Francisco, US,2000, pp. 1183–1190.

[13]. V. Bobicev and M. Sokolova, "An effective and robust method for short text classification," in AAAI, D. Fox, and C. P. Gomes, Eds. AAAI Press, 2008, pp. 1444–1445.

[14]. http://www.myvocabulary.com/word-list/violencevocabulary

[15]. https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

[16]. http://www.bannedwordlist.com/lists/swearWords.txt