UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Classification of Sexual Harassment on Facebook Using Term Weighting Schemes

Amer Saeed Ali Al-Katheri and Maheyzah Md Siraj
Faculty of Computing,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
Email: ameralkatheri@outlook.com, maheyzah@utm.my

Submitted: 12/01/2018. Revised edition: 26/03/2018. Accepted: 8/05/2018. Published online: 21/05/2018

***Abstract*—Facebook is the largest Social Network Service, and its users are growing rapidly. Facebook has become one of the main sources of information for individuals and organizations; and this exponential increase of information has raised the issue of information security. In United States alone, 62% of online abuses occurred through Facebook and the most common form of online abuse is sexual harassment with 44%. Victims to online sexual harassment are living under pressure, because the harasser has an ability to propagate messages at any time under any identity. Several content filtering tools for web-based especially Facebook has been proposed. Most of these approaches are not suitable and has limitations when applied to current Social Network Services such as Facebook. As a result, the content-based technique which includes deeper understanding of the semantics of text would probably perform better to forbid illegal post contents. In this project, three terms weighting schemes namely Entropy, TF-IDF, and Modified TF-IDF are used as feature selection process in filtering Facebook posts. The filtering are done by using Support Vector Machine (SVM) classification technique on two dataset based on accuracy and precision. The experimental results show that Modified TF-IDF (M.TFIDF) performed better than Entropy and TFIDF. It is hoped that this study would give other researchers an insight especially who want to work with similar area.**

Keywords — Term weighting, machine learning, text classification, Facebook filtering

I. Introduction

Users on Facebook are growing rapidly as information on these social networks has become one of the main source of information for individuals and organizations. According to Facebook, they have 1.86 billion users, they spent around 20 minutes on each visit, from five web pages browsed in United States there is one view on Facebook, there are 300 million photo upload per day, and each day there are around 10 million websites show buttons related to Facebook as Like button. Also, every 60 seconds on Facebook; more than half million comments, almost 300 thousands statuses update, and more than ten thousand photos added. Moreover, 65% of Facebook' users are under 24 years old, and there are 38% of them under the age of 13 [1].

With the growing of Facebook' users and easy access to low cost internet, users spent more time on services that provided by Facebook. As a result, Facebook is home to 62% of online abuse in the United State [2]. A survey also shows 47% of Americans under the age of 35 have been abused through Facebook or know somebody who has. Furthermore, Sexual harassment is the most common form of online abuse at 44%. Victims to online sexual harassment are living under pressure, because the harasser has an ability to propagate messages at any time under any identity.

II. Problem Background

In United States, 62% of online abuses occurred through Facebook [3]. The most common form of online abuse is Sexual harassment at 44%. Moreover, 65% of Facebook' users are under 24-years-old. As a result, undesirable contents in social media such as sexual harassment must be filtered to protect children from being exposed to one of these crimes.

Undesirable contents on the internet can be filtered and denied. Users are advisable to use a web content filtering to block undesirable contents on the internet. A filtering' technique used is based on Uniform Resource Locator (URL) blocking, which it checks the required URL address with list of URL addresses stored in a database is known as URL dataset. Thus; These software are working by checking each request to the internet, and a content will be either blocked completely or

redirected to another location if it matched a prevented URL. This technique is usually used by organizations to prevent users from accessing and viewing unsuitable content on websites.

On the other hand, another technique can be used which is based on keyword matching. It worked by the basis of the occurrence of prohibited keywords on that specific web page. Both techniques cannot be used to observe contents on a webpage, which make it limited to be used with current social network services such as Facebook. As a result, the content-based technique which includes deeper understanding of the semantics of text and other items of Facebook' content, would probably perform better to forbid illegal posts [4].

III. Problem Statement

Facebook does not provide content-based preferences to control posts showed to a user. Thus, it is not possible to prevent undesired posts, such as harassment' posts, without concerning about the other user who post them. Hence, an investigating for three term weighting schemes on two datasets are used to improve the automated content-based classification techniques which is required to filter out unwanted posts. The objectives of the research are:

i) To pre-process the dataset through data parsing, Stemming, Stopping and represent it into a text document.

ii) To select and classify Facebook' posts based on feature selection; Entropy; TF-IDF, and Modified TF-IDF.

iii)To evaluate the performance of Term Weighting Schemes and find their accuracy, precision, recall and F score.

IV. Research Methodology

The framework of this study consist of three main phases. Each phase will generate the input to the next phase. First phase is the initial step to conduct this research. It starts with data collection, moving to pre-process that data, then represent that data into suitable form. Second phase has the most important steps of the framework, it has the feature selection as well as the classification process. In this phase, term weighting schemes will be implemented to select relevant terms toward sexual harassment, and these terms are used to classify sexual harassment and nonsexual harassment terms. Third phase is to evaluate the performance of term weighting schemes used in the research framework. Figure 1 shows an overview of the research framework. Each phase will be illustrated in detail in the following sections.

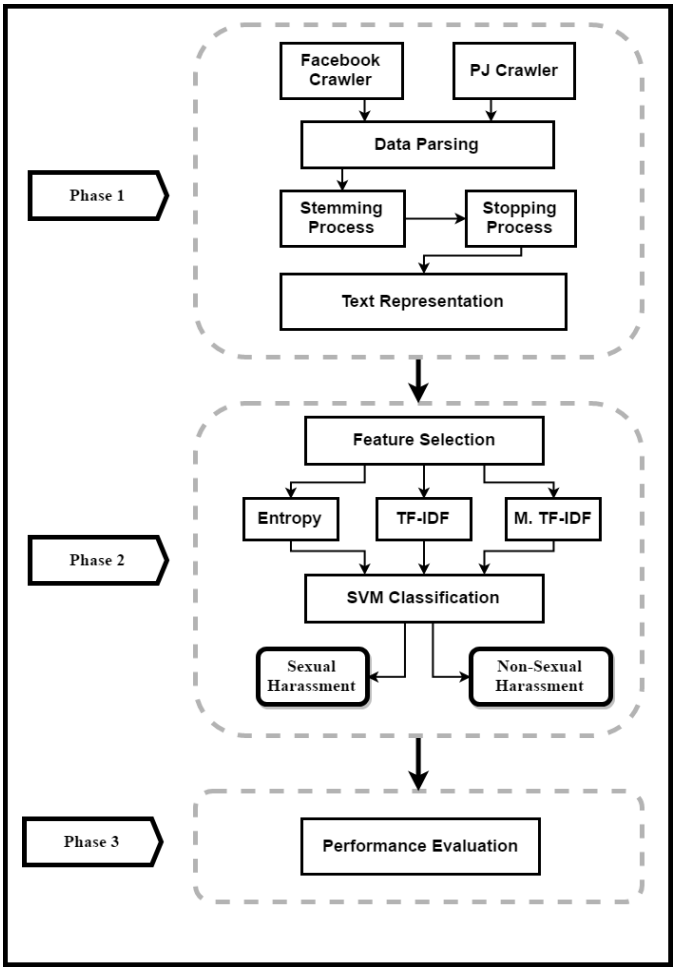

Fig. 1. Research framework.

V. Dataset

This study will use two datasets provided by MyPersonality project and Perverted-Justice Foundation. Firstly, the research team of MyPersonality project are working on various researches related to Facebook, and they released a dataset for researcher that contains around 10 thousands posts from 250 users. Thus; there are 4000 posts are randomly selected from the whole dataset to be used in the proposed research. On the other hand, Perverted-Justive foundation is working with volunteers who carry out sting operations posing as children or teenagers on chat sites and waiting for adults to approach them. They released a dataset containing chat logs with entire conversations between volunteers and predators. For the purpose of this research, 50 chat logs randomly selected, and messages which written by the predators are extracted to represent the positive data. Table I shows a description of the dataset labels.

TABLE I. DATASET LABELS

	Category	Label	Total
1	MyPersonality Dataset	MP	4000 Posts
2	Perverted-Justice Dataset	PJ	50 Chat Logs

TABLE II. DATASET CATEGORIES

Datasets	Category	Training	Testing	**Total**
1	PJ	700 Posts	300 Posts	**1000 Posts**
	MP	15 Chat Logs	5 Chat Logs	**20 Chat Logs**
2	PJ	2000 Posts	1000 Posts	**3000 Posts**
	MP	20 Chat Logs	10 Chat Logs	**30 Chat Logs**

There are two datasets to do the proposed research. The first dataset is nearly balanced dataset and contained 1000 posts and 20 chat logs. The second dataset was an imbalanced dataset, and it contained 3000 posts and 30 chat logs. The reason for creating balanced and imbalanced dataset is to test the performance of aforementioned algorithms with different types of datasets. Imbalanced datasets closely resemble the real life data that algorithms might need to train on. Thus, it is imperative that the learning capacity of feature selection a algorithm is independent of skew in data. Table II shows how the data has divided to two groups.

VI. EXPERIMENTAL SETUP OF TERM WEIGHTING SCHEMES AND FEATURE SELECTION PROCESS

Term feature ranking is a process in which the number of features is reduced by selecting only number of the best features, which represents the most representative terms within the collection. The feature selection would only reduce the dimension of term but not the documents, means word' ranking is not about how frequent single word is repeated in a collection, it is related how important this word to the collection. Thus, the main objective is to get the most relevant features to be used as training and testing of data. Figure 2 shows the detailed term feature ranking process [5].

Therefore after the feature selection process, first set VecA with |β| x N becomes matrix VecB with P´ x N where P´ donates the number of selected features. Prior to this, the term features are ranked according to the sum of their weights. Following this, the sum weight value is sorted from highest to lowest. Finally P´ number of features with the highest sum weight is selected as the best feature, and a new matrix VecB with P´ x N is formed [6].

The processes will be repeated for every weighting schemes; Entropy, TFIDF and M.TFIDF which are used in this study. The formula used to compute the term weights is different for each weighting scheme.

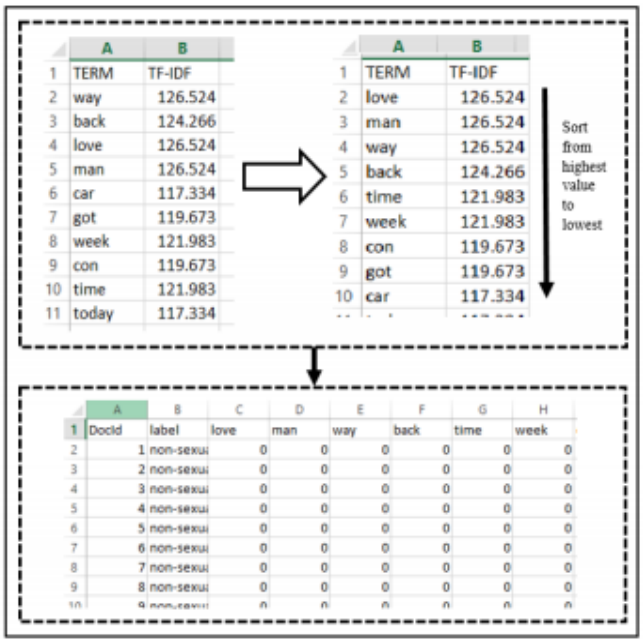

Fig. 2. Term feature ranking process.

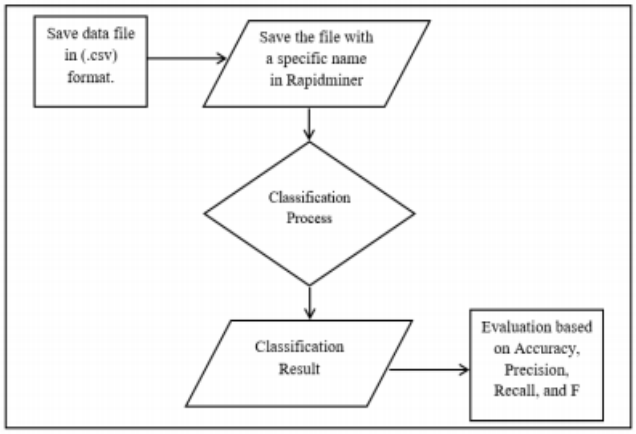

Fig. 3. Steps involved in classification.

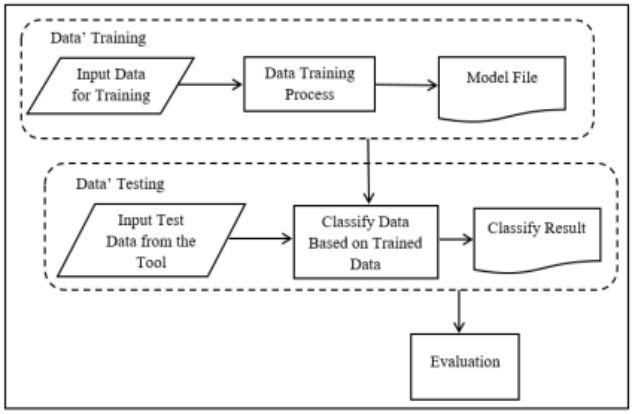

Fig. 4. Overview of classification process.

VII. Implementation of SVM as a Classifier

Rapidminer tool is used in this study, from which the SVM classifier can be easily called. lt has been offered by Rapidminer which is user friendly and reliable too. In order to input data to SVM, several steps are involved such as setting label and id to make the data sets acceptable for Rapidminer and SVM too. Figure 3 displays the steps involved in Classification of the data sets in Rapidminer tool.

Rapidminer accepts several file formats, .csv format is used in this study. Files easily can be called in the tool, then by setting the attributes with labeling then it is saved again in Rapidminer library. Figure 4 displays the overview of classification process using Rapidminer based on SVM. In this study as mentioned earlier in chapter 3, 1300 Facebook posts and 15 chat logs are used for testing the data, and 2700 Facebook posts and 45 chat logs for training the data. It will be repeated for each dataset consecutively.

VIII. Experimental Results And Discussion

In order to examine the performance of each term weighting schemes: Entropy, TFIDF, and M.TFIDF, different number of features are selected, each increment of 30 features is taken as an evaluation benchmark. In first data set, there are 1000 posts from Facebook and 20 chat logs from Preverted-justice. The results for dataset 1 are presented in Table III and Table IV. The second experiment has larger collection compared to the previous experiment. The results for dataset 2 are presented in Table V and Table VI.

In this study, accuracy and precision are analyzed for their ability to classify the sexual harassment content on Facebook. The higher measurement value the better the performance. This clearly illustrates that M.TFIDF achieved the highest measurement with accuracy over Entropy and TFIDF. The result might fluctuate if many features or data used in training and testing in the classification process. The result for these experiments are with only a range of 1-90 features.

Based on observations presented here, the experimental results for two data-sets indicate that they form patterns, the more important are listed below.

Observation-1: M.TFIDF achieves the highest accuracy measurement for both experiments.

Observation-2: M.TFIDF achieves higher value than TFIDF for precision and F measurement when data collection is small. Once the data collection become bigger, Entropy achieves higher than M.TFIDF.

Observation-3: TFIDF always achieves the lowest measurement value among all term weighting schemes.

Observation-4: Recall and F measurement values are improved according to the increase in the number of features, when the selected features are within the number 0-90.

Observation-5: The term weighting schemes achieved their best measurement values when 90 features are selected.

TABLE III. Performance on Accuracy for Dataset 1

Features	Entropy (%)	TFIDF (%)	M.TFIDF (%)
30	84.90	97.12	78.00
60	91.30	92.00	96.50
90	95.80	89.50	96.50
Standard Deviation	5.48	3.88	10.68

TABLE IV. Performance on Precision for Dataset 1

Features	Entropy (%)	TFIDF (%)	M.TFIDF (%)
30	84.30	98.70	75.00
60	95.90	95.00	89.80
90	94.00	96.00	92.00
Standard Deviation	6.22	1.91	9.25

TABLE V. Performance on Accuracy for Dataset 2

Features	Entropy (%)	TFIDF (%)	M.TFIDF (%)
30	82.80	95.80	78.60
60	92.00	92.00	96.00
90	95.00	89.90	96.00
Standard Deviation	5.97	2.99	10.05

TABLE VI. Performance on Precision for Dataset 2

Features	Entropy (%)	TFIDF (%)	M.TFIDF (%)
30	72.50	99.00	78.80
60	98.50	94.50	96.50
90	98.50	89.50	96.50
Standard Deviation	13.28	4.75	10.22

XII Research Findings and Contributions

The aim of this study is to find out which term weighting scheme gives better results for the classification of sexual harassment contents on Facebook; which are namely Entropy, TFIDF and M.TFIDF. The term weighting schemes were tested with two kinds of environments. Each of the term weighting scheme is evaluated using accuracy, precision, recall, and F measurements.

The feature selection methods in this study, namely Entropy, TFIDF and M.TFIDF term weighting schemes. Entropy and M.TFIDF are modifications of TFIDF with a difference that Entropy and M.TFIDF calculates the Local and Global weight. It also takes into consideration document length in the calculations. From the experimental results presented, it is shown that M.TFIDF achieves better performance in terms of accuracy, precision, and F measurement. While TFIDF performs less than Entropy and M.TFIDF in four performance test; accuracy, precision, recall, and F measurements.

Nowadays, several commercial web filtering products are available in the market, but the techniques used by those needs further enhancement which makes them less efficient, especially with today's ever changing web content. This project is accomplished to classify and most importantly performance will be measured based on four different measurements. It will help in understanding the social problems due to online activities adopted by the youth such as social network websites. Furthermore, it may help to highlight

the role of online activities affecting the youth's social life and even it has impacts on parents and community.

XII CONCLUSION

Content analysis is an approach that involves a deeper understanding of the text and other media items by using machine learning methods. The content analysis approach is a trend in web filtering research. This study concerns itself with the similarity of content issues, which is faced sexual harassment content classification The similarity of content issue refers to collections that share similar terminologies for particular class. This study was carried out to distinguish the performance of term weighting schemes Entropy, TFIDF and M.TFIDF. Two different data sets, which represent the two classes which used in this study to examine the term weighting schemes. These are evaluated using standard information retrieval measurements.

REFERENCES

[1] Aparna U.R. and Shaiju Paul. (2016). Feature Selection and Extraction in Data Mining, Green Engineering and Technologies (IC-GET). *2016 Online International Conference.*

[2] Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group Formation in Large Social Networks: Membership, Growth, and Evolution. *Proceedings of 12th International Conference on Knowledge Discovery in Data Mining.*

[3] Barak, A. (2005). Sexual Harrassment on the Internet. *Social Science Computer Review*, 23(1), 77-92.

[4] Breitinger, C., Gipp, B., Langer, S. (2015). Research-paper Recommender Systems: A Literature Survey. *International Journal on Digital Libraries*, 17(4), 305-338.

[5] Cawley, G. and Talbot, N. (2010). Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, *Journal of Machine Learning Research.* 11, 2079-2107.

[6] Celli, F., Pianesi, F., Stillwell, D., Kosinski, M. (2013). Workshop on Computational Personality Recognition (shared task). *Proceedings of WCPR13, in conjunction with ICWSM-13, 2013.*

[7] Danah, M., Nicole, B. (2007). Social Network Sites: Definition, History, and Scholarship, *Journal of Computer-Mediated Communication*, 13.

[8] E. Leopold, J. Kindermann. (2002), Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Mach. Learn,* 46(1-3), 423-444.

[9] F. Sebastiani, (2002), Machine learning in automated text categorization, ACM Comput. Surv. 34 (1) (2002) 1–47.

[10] Goodson, P., McCormick, D., & Evans, A. (2001). Searching for Sexually Explicit Materials on the Internet: An Exploratory Study of College Students' Behavior and Attitudes. *Archives of Sexual Behavior,* 30, 101-118.

[11] G. Eriksson and J. Karlgren. (2012). Features for Modelling Characteristics of Conversations.

[12] G. B. Huang, Q. Y. Zhu, C. K. Siew. (2006), Extreme Learning Machine: Theory and Applications. *Neurocomputing*, 70(1-3), 489-501.

[13] Girish, C., Ferat, S. (2013). A Survey on Feature Selection Methods. *Computers & Electrical Engineering*, 40(1), 16-28.