International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 7985-7998
© Research India Publications
http://www.ripublication.com

Spontaneous Message Detection of Annoying situation in Community Networks Using Mining Algorithm

Senthil kumari[1], Dr.G.Tamil pavai[2] , Thirumani thangam[3]
Rupavathy[4] , Baby D. Dayana[5]

Government College of engineering Tirunelveli
Email: - psk045@gmail.com

Abstract

Main concerns in data mining investigation is social controls of data mining for handling ambiguity, noise, or incompleteness on text data. We describe an innovative approach for unplanned text data detection of community networks achieved by classification mechanism. In a tangible do main claim with humbles secrecy backgrounds provided by community network for evading annoying content is presented on consumer message partition. To avoid this, mining methodology provides the capability to unswervingly switch the messages and similarly recover the superiority of ordering. Here we designated learning centered mining approaches with pre-processing technique for complete this effort. Our involvement of work compact with rule -based personalization for automatic text categorization which was appropriate in many dissimilar frameworks and offers tolerance value for permits the background of comments conferring to a variety of conditions associated with the policy or rule arrangements processed by learning algorithm. Remarkably we find that the choice of classifier has predicted the class labels for control the inadequate documents on community network with great value of effect.

Introduction

Data mining is the process to discover interesting knowledge from large amounts of data. Recent research projects in two closely related areas of computer science is machine learning and data mining that have been developed methods for constructing statistical models of network data. Examples of such data include social networks, networks of web pages, complex relational Databases and data on interrelated people, places, things, and events Extracted from text documents. Web content Mining is used to discover useful and relevant information from a large amount of Data. In community network, information can be used for a different purpose but there is the possibility of posting (or) commenting other posts on particular public (or) private areas. Automatic text classification is mainly used to give user the ability to control the message written on their own walls by filtering out unwanted messages the course of action of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both.

Automatic text classification has always been an important application and research topic since the inception of digital documents. Today, text classification is a necessity due to the very large amount of text documents that we have to deal with daily. In general, text classification includes topic based text classification and text genre-based classification. Topic-based text categorization classifies documents according to their topics. Texts can also be written in many genres, for Instance, scientific articles, news reports, movie reviews, and advertisements. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. The accuracy of modern text classification systems rivals that of trained human professionals, thanks to a combination of information retrieval (IR) technology and machine learning (ML) technology.

Text mining or knowledge discovery from text (KDT) deals with the machine supported analysis of text. It uses methods from information retrieval, natural language processing (NLP) information extraction and also connects them with the algorithms and methods of Knowledge discovery of data, data mining, machine learning and statistics. Current research in the area of text mining tackles problems of text representation, classification, clustering, or the search and modeling of hidden patterns. Text mining usually involves the process of structuring the input text.

Some of the technologies that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization classification, clustering, concept linkage, information conception, and question answering. In this new and current area of technology, developments and techniques, efficient and effective, document classification is becoming a challenging and highly required area to capably categorize text documents into mutually exclusive categories.

Text categorization is an upcoming and vital field in today’s world which is most importantly required and demanded to efficiently categorize various text documents into different categories.

The aim of the present work is experimentally evaluate an automated system with text categorization techniques for automatically assign text message. The major efforts in building a rule based text classifier are concentrated with selection of discriminate features. The solution is we inherit the learning model and the elicitation procedure for generating pre-classified data.

Overview

- **Chapter 1** Discusses about the basis of data mining and text mining and issues related to the social impact on community network
- **Chapter 2**Express the technologies related to the proposed system
- **Chapter 3** Deals with related work and models explanations.
- **Chapter4** System constructions and implementation outline are explained in this section.

- **Chapter 5** Deals with the system output design and classification result
- **Chapter 6** Results of the proposed model in term of error rate values for each class model

Literature Overview

A literature review is a description of the literature relevant to a particular field or topic. It gives an overview of what has been said, who the key writers are, what are the prevailing theories and hypotheses, what questions are being asked, and what methods and methodologies are appropriate and useful.

Techniques for text data classification

This chapter explains about the various papers that are published related to the text classification and ML techniques with different classification technique.

1. BoosTexter: A Boosting-based System for Text Categorization Robert E Schapiro, Yoram Singer [2000] suggested the goal of the learning algorithm is to predict all and only all of the correct labels. Thus, the learned classifier is evaluated in terms of its ability to predict a good approximation of the set of labels associated with a given document. In the second extension and the goal is to design a classifier that ranks the labels so that the correct labels will receive the highest ranks. Suffer from over fitting problem.

2. **Content-Based Book Recommending Using Learning for Text Categorization Raymond J. Mooney, Loriene Roy [2011]** suggested Recommender systems improve access to relevant products and information by making personalized suggestions based on previous examples of a user's likes and dislikes. By contrast, content-based methods use information about an item itself to make suggestions. This approach has the advantage of being able to recommend previously unrated items to users with unique interests and to provide explanations for its recommendations. Availability of data inconsistency.

3. Content-based Filtering in On-line Social Networks, M. Vanetti, E. Binaghi, B. Carminati, M. Carullo and E. Ferrari[2010] proposed work is early encouraging results we have obtained on the classification procedure prompt us to continue with other work that will aim to improve the quality of classification. Additionally, we plan to enhance our filtering rule system, with a more sophisticated approach to manage those messages caught just for the tolerance and to decide when a user should be inserted into a BL.

4. Inductive Learning Algorithms and Representations for Text Categorization, Susan Dumais, John Platt, David Heckerman, Mehran Sahami[2012]It describe results from experiments using a collection of hand-tagged financial newswire stories from Reuters. We use supervised learning methods to build our classifiers, and evaluate the resulting models on new test cases. Low Performance for Highly Co-related Features

5. A System to Filter Unwanted Messages from OSN User Walls, Marco vanity, elisabetta binaghi, Elena Ferrari [2013] suggested the fundamental issue in today On-line Social Networks (OSNs) is to give users the ability to control the messages posted on their own private space to avoid that unwanted content is displayed. Up to now OSNs provide little support to this requirement. To fill the gap, a system allowing OSN users to have a direct control on the messages posted on their walls. This is achieved through a user to customize the filtering criteria to be applied to their walls, and a Machine Learning based soft classifier automatically labeling messages in support of content-based filtering. Availability of data inconsistency and noisy data.

6. Maximum Likelihood Estimation for Filtering Thresholds. Zhang and J. Colane [2010] suggested the filtering system examines a document flow to find documents that match the information needs described by profiles consisting of queries and related context or history. Filtering systems based on statistical models use a numeric score to indicate how well a document matches a profile, and only disseminate a document when its score is above some threshold and the system may be provided with relevance judgments for the document classification. Low Classification Accuracy

7. Document Categorization by Term Association Narmada Mahindra [2012] A good text classifier is a classifier that efficiently categorizes Large sets of text

documents in a reasonable time Frame and with an acceptable accuracy, and that provides Classification rules that are human readable for possible Fine-tuning. We present a Novel approach for automatic text categorization that borrows from market basket analysis techniques using association Rule mining in the data-mining field. Number of term increase & required large physical memory

8. Collaborative filtering for People to People Recommendation in Social NetworksM.J. Pazzani and D. Billson [2011] states predicting people other people May like has recently become an important task in many online social networks. Collaborative filtering methods to enable people to people recommendation. Users can be similar to other users in two ways – either having similar “taste” for the users they contact, or having similar “attractiveness” for the users. Demographic Filtering and Multi-criteria Ratings is required for classification process.

9. Machine learning text categorization in osn to filter unwanted messages, Chou and H. Chen [2008] suggested Major issue in OSN (Online Social Network) is to preventing security in posting unwanted messages to avoid this issue, BL mechanism is proposed in this paper, which avoids undesired creators messages. BL is used to determine which user should be inserted in BL and decide when the retention of the user is finished. Conditional independence is violated by real world data.

Methodology

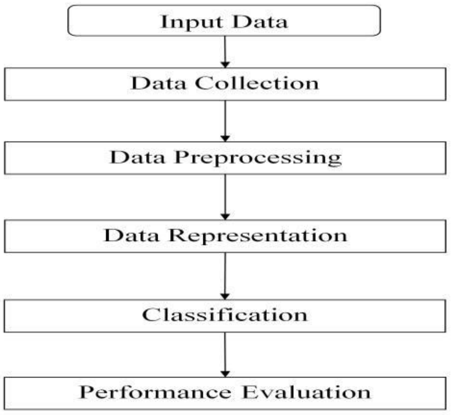


Figure 1: System Flow

The methodologies of the research work shows in figure 1. It begins with data collection and performs the data preprocessing for making data consistence. After pre-process the data indexing used to assign the weight for each class label based on statistical methods. Classification provide a set of classification rules that can be used later to evaluate a new case and classify in a predefined set of classes. The accuracy of predictions made about an instance after classifier evaluation.

Data Collection

Document gathering is major step of classification method and then Data Pre-processing is a step which used to presents the documents into clear arrangement. In a representation phase the document has to be converted from the full text version to a document vector. The main idea of Feature collection (FC) is to select subsection of features from the original documents. FC is performed by keeping the words with highest score according to predetermined measure of the importance of the word. By matching the different classifiers, based on the performance results best classifier should be preferred by the following measures such as True positive, true negative, false positive, false negative should denoted as TP, TN, FP, FN. Here the DC process may consist of following steps to formulate the input data as...

- ➢ DC may be formalized as the assignment of resembling the unidentified objective function
- ➢ $\Phi : D \times C \rightarrow \{T,F\}$ (that describes how documents should to be classified , conferring to a supposedly authoritative expert) by means of a function

$\Phi: D \times C \rightarrow \{T, F\}$ terms the classifier

Where $C = \{c1. . . c|C|\}$ is a predefined set of groups
D is a (feasibly infinite) set of documents.
If Φ (dj, ci) = T,

Then dj is positive instance (or a member) of ci,
Else if Φ (dj, ci) = F is a negative instance of ci.

Pre-Processing

After the collection stage the Data may give poor results due to class imbalance problem so we need to identify the problem in an initial phase and then only the documents prepared for classification that may represented by a great amount of features. Commonly the steps in pre-processing taken here as follows,

Tokenization: A document is treated as a string, and then partitioned into a list of tokens.

Removing stop words: Stop words such as "the", "a", "and", etc. are frequently occurring, so the insignificant words need to be removed.

Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute.

Data Indexing

The document representation or indexing is one of the pre-processing techniques that are used to reduce the complication of the documents and make them easier to switch. The document has to be transformed from the full text version to a document vector. The most commonly used document representation is called vector space model (SMART).Documents are represented by vectors of words. Usually, one has a collection of documents which is represented by word by word document environment.VSM representation scheme has its own limitations. Some of them are high dimensionality of the demonstration, loss of correlation with adjacent words and loss of semantic relationship that exist among the expressions in a document. After pre-processing and indexing the important step of text classification, is feature selection. It constructs vector space, which improves the scalability, effectiveness and accurateness of a text classifier. The main idea of Feature collection (FC) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word.

Rule Based Classification

The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Rules based classification method uses the rule-based inference to classify documents to their annotated categories a popular format for interpretable solutions is the conjunctive normal form model. It is recommended to reduce the size of rules set without affecting the performance of the classification. The commercial value being able to classify documents automatically by content. Reasonable amount of labeled data with automatic classification can do to predict the classification accuracy.

Experimental Setup

Here we have carried out the entire module of the classification and start by relating the dataset. Here we have three main steps are Data collection, Data preprocessing and Classification process, evaluation. In the figure 1, we have explained the entire experimental arrangement of classification procedure. Here in the primary step web data are collected from the different categories of community networks. We included two datasets for test and training model. These records are given to WEKA. From the key input files, we execute the removal of stop words and stemming function. Behind that we have generated text to term matrix whose value gives the count of each word in each web data. Then each term is assigned TF-IDF function for modeling the data from the web. Finally we apply the classification algorithms to produce the classification outcome. The Task is to Building a model for classifying the messages posted in community network. Our data source was selected only message where the user was expressed with posted messages represented in numerical value (other conventions varied too widely to allow for automatic processing). Collected Messages were automatically into the following categories are hate, love, natural, sex, violence, casual, profile, contact, like and unlike, post, status & command messages. For the work described in this paper, we concentrated only on discriminating between the above message type for identify the messages under the form of bad and good text data in data source. It was achieved by Weka is a collection of machine learning algorithms for data mining tasks. At first we have to Generating datasets for experiments then Convert the data into an ARFF, CSV, and C4.5, XRFF format and load data to perform the pre-processing and classification.

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of supports many different standard data mining tasks.

1. Preprocess - used to choose the data file to be used by the application.
2. Classification - used to test and train different learning schemes on the preprocessed data file under experimentation.
3. Evaluation - Evaluation on training data and test data.

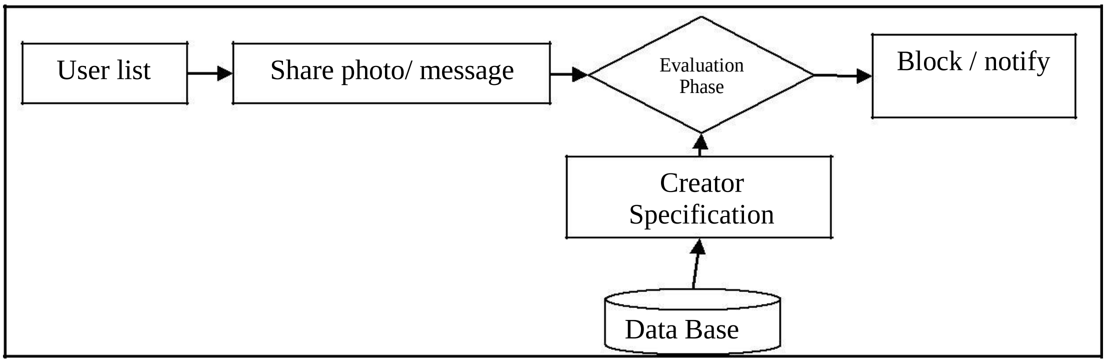
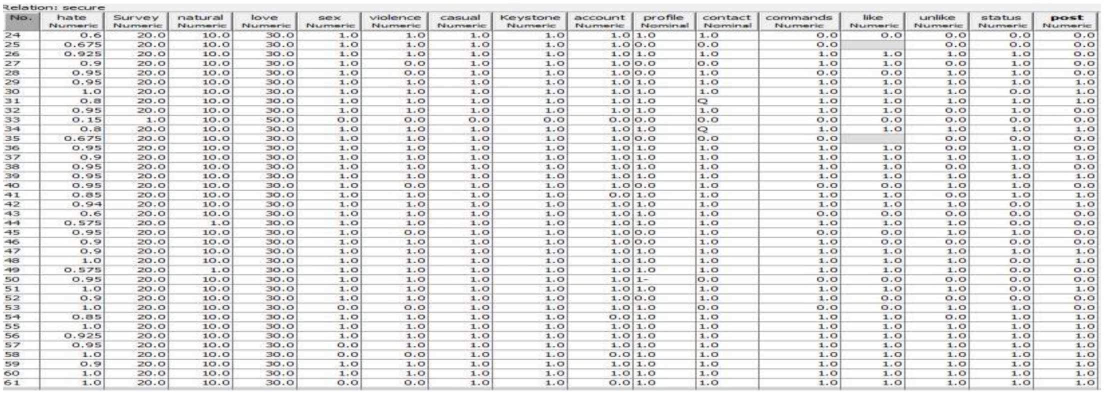


Figure 2: Network management task

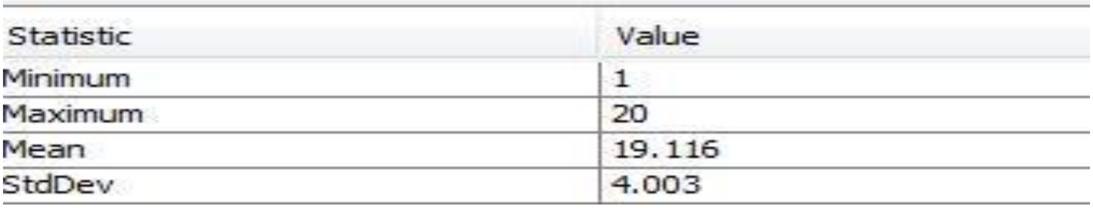
Implementation Outline



Relation: secure

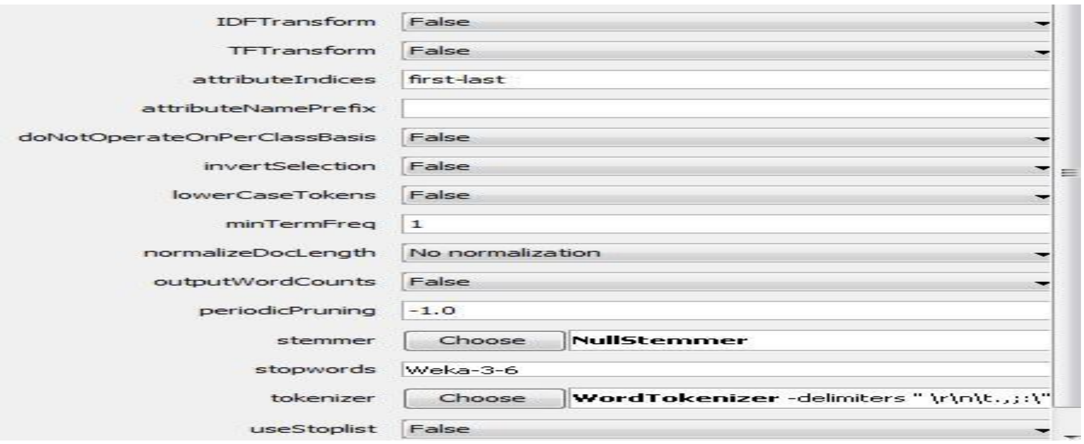
No.	hate Numeric	Survey Numeric	natural Numeric	love Numeric	sex Numeric	violence Numeric	casual Numeric	Keystone Numeric	account Numeric	profile Nominal	contact Nominal	commands Numeric	like Numeric	unlike Numeric	status Numeric	post Numeric
24	0.6	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
25	0.675	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0		0.0	0.0	0.0
26	0.925	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0
27	0.9	20.0	10.0	30.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0
28	0.95	20.0	10.0	30.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0
29	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
30	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
31	0.8	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	Q	1.0	1.0	1.0	1.0	1.0
32	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0
33	0.15	1.0	10.0	50.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	0.8	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	Q	1.0	1.0	1.0	1.0	1.0
35	0.675	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0		0.0	0.0	0.0
36	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0
37	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
38	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0
39	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
40	0.95	20.0	10.0	30.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0
41	0.85	20.0	10.0	30.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0
42	0.94	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
43	0.6	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
44	0.575	20.0	1.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
45	0.95	20.0	10.0	30.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0
46	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
47	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
48	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
49	0.575	20.0	1.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
50	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1-	0.0	0.0	0.0	0.0	0.0	0.0
51	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
52	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
53	1.0	20.0	10.0	30.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0
54	0.85	20.0	10.0	30.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0
55	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
56	0.925	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
57	0.95	20.0	10.0	30.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
58	1.0	20.0	10.0	30.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
59	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
60	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
61	1.0	20.0	10.0	30.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(A) DATA SOURCE



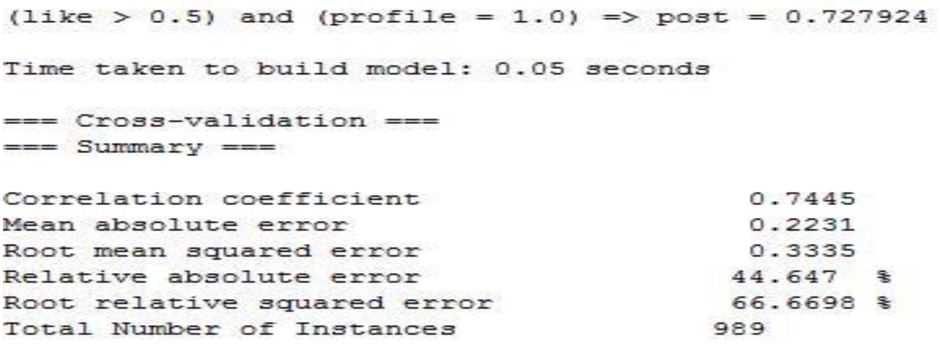
Statistic	Value
Minimum	1
Maximum	20
Mean	19.116
StdDev	4.003

(B) DATA INDEXING VALUES



IDFTransform	False
TFTransform	False
attributeIndices	first-last
attributeNamePrefix	
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	False
minTermFreq	1
normalizeDocLength	No normalization
outputWordCounts	False
periodicPruning	-1.0
stemmer	Choose **NullStemmer**
stopwords	Weka-3-6
tokenizer	Choose **WordTokenizer** -delimiters " \r\n\t.,;:\'"
useStoplist	False

(C) Loading the input data and perform the DATA PRE-PROCESSING



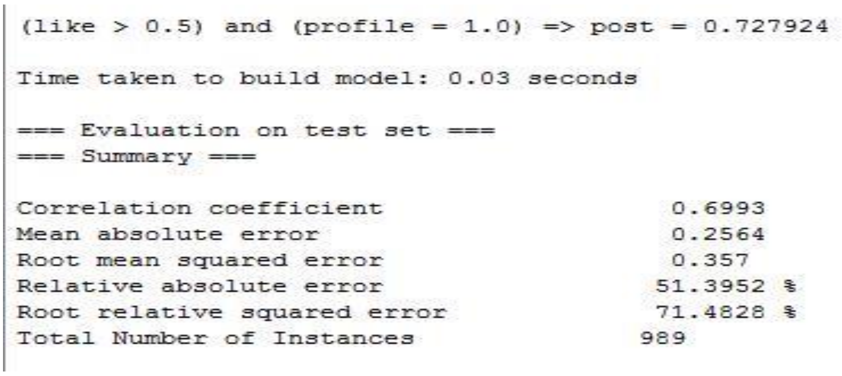
```
(like > 0.5) and (profile = 1.0) => post = 0.727924

Time taken to build model: 0.05 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                    0.7445
Mean absolute error                        0.2231
Root mean squared error                    0.3335
Relative absolute error                   44.647   %
Root relative squared error               66.6698 %
Total Number of Instances                989
```

(D) Perform classification using RULE model for training data



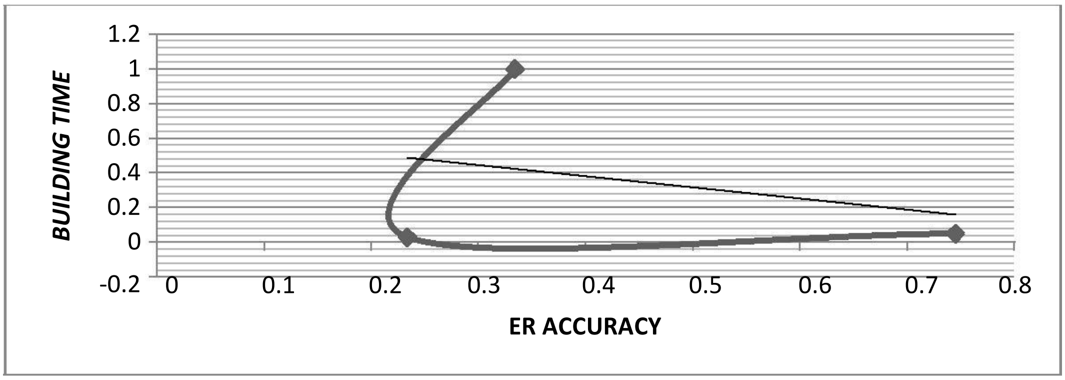
```
(like > 0.5) and (profile = 1.0) => post = 0.727924

Time taken to build model: 0.03 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                    0.6993
Mean absolute error                        0.2564
Root mean squared error                    0.357
Relative absolute error                   51.3952 %
Root relative squared error               71.4828 %
Total Number of Instances                989
```

(E) Classification Result for test data



(F) Concurrence relation between building time and ER accuracy

Results of The Proposed Model In Term of Error Rate Values For Each Class Model

Table 1: Evaluation on training data

Methods	Building time (sec)	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Multi layer perceptions	5.19	0.8413	0.1812	0.2785	36.266 %	55.6776 %
bagging	0.13	0.9067	0.879	0.2107	17.5922%	42.1219%
M5 rules	0.31	0.9058	0.0982	0.2118	19.6558%	42.334%
MSP	0.13	0.9019	0.1201	0.2715	24.0416%	43.48883%
RBFNet work	0.09	0.6114	0.3186	0.3953	63.7571%	79.0279%
Rule based model	0.06	0.7445	0.2231	0.3335	44.647%	66.6698%
Linear regression	0.19	0.8475	0.1808	0.2652	36.1913%	53.0095%

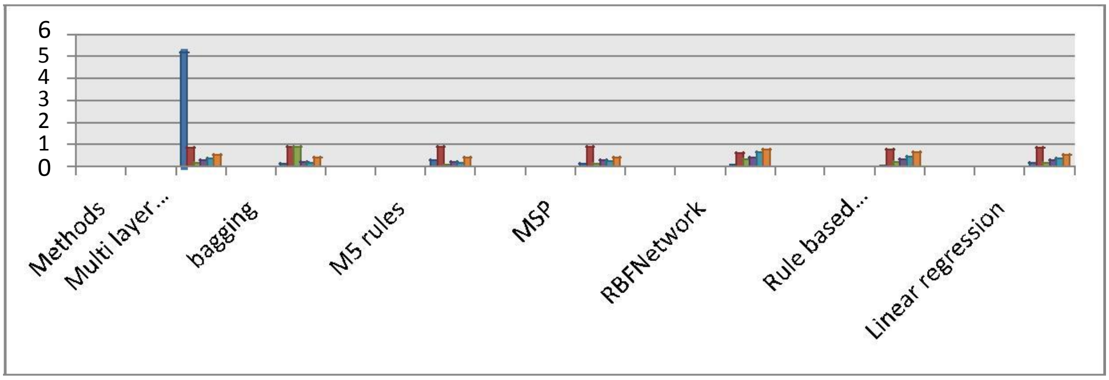


Figure 3: The proposed model analysis for training data

Table 2: Evaluation on test data

Methods	**Building time (sec)**	**Correlation coefficient**	**Mean absolute error**	**Root mean squared error**	**Relative absolute error**	**Root relative squared error**
Multi layer perceptions	5.79	0.8872	0.1373	0.2337	27.5191%	46.7973%
bagging	0.17	0.9086	0.087	0.2087	17.4274%	41.7816%
M5 rules	0.36	0.9079	0.1001	0.2098	20.0656%	41.9968%
MSP	0.09	0.9045	0.1158	0.2145	23.2101%	42.9381%
RBFNetwork	0.08	0.6984	0.2556	0.3575	51.2191%	71.5675%
Rule based model	0.06	0.6993	0.2564	0.357	51.3952%	71.4828%
Linear regression	0.19	0.8489	0.1792	0.264	35.9158%	52.8531%

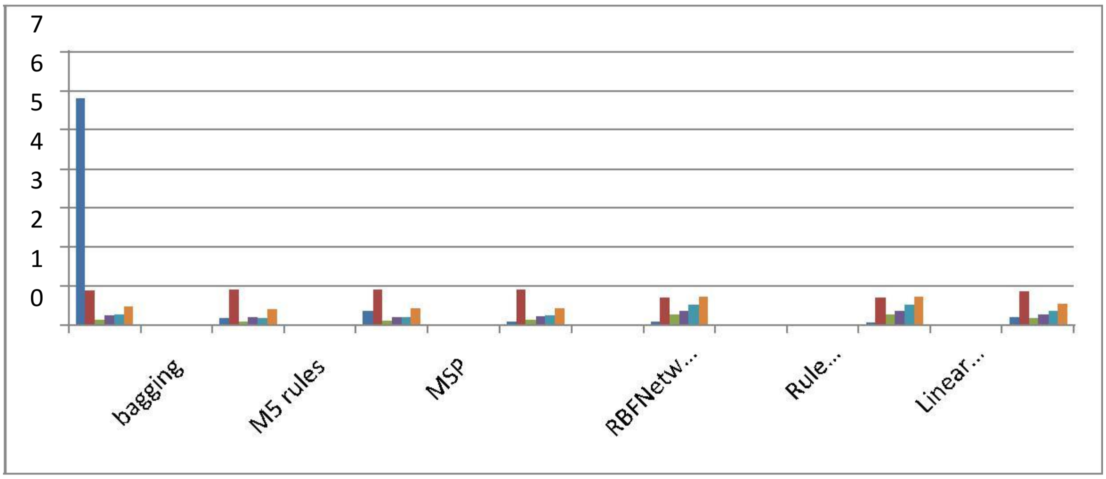
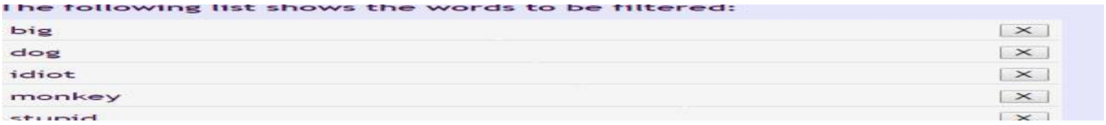


Figure 4: The proposed model analysis for test data

Application Development



The following list shows the words to be filtered:

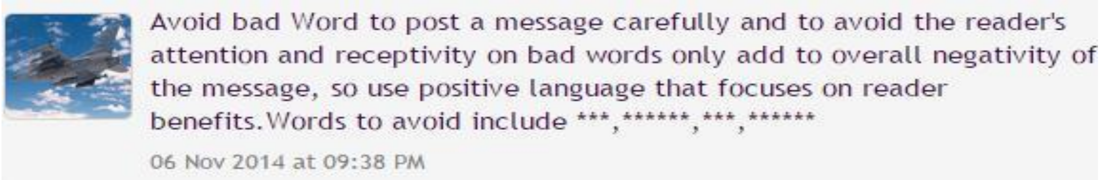
big

dog

idiot

monkey

stupid



Avoid bad Word to post a message carefully and to avoid the reader's attention and receptivity on bad words only add to overall negativity of the message, so use positive language that focuses on reader benefits.Words to avoid include ***,******,***,******

06 Nov 2014 at 09:38 PM

Page for word categorization on posted messages and the Result for avoiding unwanted messages post message on user wall.

Conclusion

Finally the rule based mining system was prevent the licentious messages from the social network and the usage of machine learning has given higher results to the system to trace the messages and the users to distinguish between the good and bad messages and the authorized and unauthorized users in the social networks can find the user profiles automatically.

Thus the machine learning technique plays a vital role in order to generate the blacklist of the bad words and the unauthorized users. The user has to update his privacy setting in his account by using this method to prevent the obscenity in their public profile. In this context a rule based analysis has been conducted and to provide the usage of the good and bad words by the persons in the sites. Overall, the obscenity of the users has been prevented.

Future Work

This paper elaborates few systems which detect offensive content and identify potential offensive users in social media. Enforcing message level classification is conceived as a key service for osn for that the system allows osn users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system.

User-level offensiveness evaluation is still an under researched area and i plan to study strategies and techniques limiting the inferences that a user can do on the enforced filtering rules with the aim of bypassing the filtering system, such as for instance randomly notifying a message that should instead be blocked, or detecting modifications to profile attributes that have been made for the only purpose of defeating the filtering system based on the user level classification.

References

[1] A. Adomavicius and g. Tuzhilin, “toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” ieee trans. Knowledge and data eng., vol. 17, no. 6, pp. 734-749, june 2005.

[2] M. Chau and h. Chen, “a machine learning approach to web page filtering using content and structure analysis,” decision support systems, vol. 44, no. 2, pp. 482-494, 2008.

[3] R.j.mooney and l. Roy, “content-based book recommending using learning for text categorization,” proc. Fifth acm conf. Digital libraries, pp. 195-204, 2000.

[4] M. Vanetti, e. Binaghi, b. Carminati, m. Carullo, and e. Ferrari,“content-based filtering in on-line social networks,” in proceedings of ecml/pkdd workshop on privacy and security issues in data mining and machine learning (psdml 2010)

[5] Marco vanetti, elisabetta binaghi, elena ferrari, barbara carminati, moreno carullo, “a system to filter unwantedmessages from osn user walls,”ieee transaction on knowledge and data engineering, vol. 25, 2013.

[6] Fong, p.w.l., anwar, m.m., zhao, z.: a privacy preservation model for face book-style social network systems. In: proceedings of 14th european symposium on research in computer security (esorics). Pp. 303–320 (2009).

[7] K. Nirmala, s. Satheesh kumar, “a survey on text categorization in online social networks,” in proceedings ofinternational journal of emerging technology and advanced engineering, volume 3, issue 9, september 2013**.**

[8] N.j. belkin and w.b. croft, “information filtering and information retrieval: two sides of the same coin?” Comm. Acm, vol. 35, no.12, pp. 29-38,1992

[9] P.w.foltz and s.t.dumais, “personalized information delivery: an analysis of information filtering methods,”comm. Acm, vol. 35, no. 12, pp. 51-60, 1992

[10] F.sebastiani, "machine learning automated text categorization", acm computing surveys, vol.34, no.1, pp.1-47, 2002.

[11] G. Amati and f. Crestani, “probabilistic learning for selective dissemination of information,” information processing and management, vol. 35, no. 5, pp. 633-654, 1999

[12] Su, x., khoshgoftaar, t.: a survey of collaborative filtering techniques. Advances in artificial intelligence 2009.

[13] pavlov, d., pennock, d.m.: a maximum entropy approach to collaborative filtering in dynamic, sparse, high dimensional domains. In: neural information processing systems, pp. 1441–1448 (2002).

[14] international journal of artificial intelligence & applications vol.3, no.2, march 2012 10.5121/ijaia.2012.3208 85text classification and classifiers:a survey

[15] sujapriya. S , g. Immanuel gnana durai , dr.c.kumar charlie paul “filtering unwanted messages from online social networks (osn) using rule based technique**”,** iosr journal of computer engineering (iosr-jce) e-issn: 2278-0661, p- issn: 2278-8727volume 16, issue 1, ver. I(jan. 2014), pp 66-70

[16] p. W. Foltz and s. T. Dumais, “personalized information delivery:an analysis of information filtering methods,” communications ofthe acm, vol. 35, no. 12, pp. 51–60, 1992

[17] s. Pollock, “a rule-based message filtering system,” acm transactionson office information systems, vol. 6, no. 3, pp. 232–254,1988.

[18] p. E. Baclace, “competitive agents for information filtering,” communicationsof the acm, vol. 35, no. 12, p. 50, 1992.m. J. Pazzani and d. Billsus, “learning and revising user profiles:the identification of interesting web sites,” machine learning,vol. 27, no. 3, pp. 313–331, 1997.

[19] y. Zhang and j. Callan, “maximum likelihood estimation for filteringthresholds,” in proceedings of the 24th annual international acmsigir conference on research and development in informationretrieval, 2001, pp. 294–302.

[20] c. Apte, f. Damerau, s. M. Weiss, d. Sholom, and m. Weiss,“automated learning of decision rules for text categorization,” transactionson information systems, vol. 12, no. 3, pp. 233–251, 1994.

[21] s. Dumais, j. Platt, d. Heckerman, and m. Sahami, “inductivelearning algorithms and representations for text categorization,” inproceedings of seventh international conference on information andknowledge management (cikm98), 1998, pp. 148–155.

[22] m. Carullo, e. Binaghi, and i. Gallo, “an online document clusteringtechnique for short web contents,” pattern recognition letters,vol. 30, pp. 870–876, july 2009.

[23] rashmi r.atkare and p.d.soni, “survey of filtering system for osn (online social networks),” int.j.computertechnology & applications,vol 4 (6), pp. 969-972, nov-dec 2013

[24] p.w.foltz and s.t.dumais, “personalized information delivery: an analysis of information filtering methods,”comm. Acm, vol. 35, no. 12, pp. 51-60, 1992..

[25] aijun an and xiangji huang, "feature selection with rough sets for web page categorization", york university, toronto, ontario,canada.

Author Details

Senthil Kumari P received her Bachelor degree in CSE from University College of Engineering Nagercoil, Tamilnadu in 2013. In 2015, she is pursuing her Master degree in Computer Science and Engineering from Government college of Engineering, Tirunelveli .Her research interest includes data mining and cloud security.

Dr. Tamil Pavai G assistant professor (sr/gr) GCE Tirunelveli, she received her Bachelor degree in CSE from Thiyagaraja College of Engineering Madurai, Tamilnadu in 199 and she was pursuing her Master degree in Computer Science and Engineering from Government college of Engineering, Tirunelveli .Her research interest area is medical image processing. She honored her doctoral degree from Anna University, Chennai

Thirumani Thangam V received her Bachelor degree in CSE from Sethu institute of technology, Madurai Tamilnadu in 2012. In 2015, she is pursuing her Master degree in Computer Science and Engineering from Government college of Engineering, Tirunelveli .Her research interest includes data mining and cloud computing

N.Rupavathy received her Bachelor degree in Information Technology from Sri Sivasubramaniya Nadar College of Engineering, Tamilnadu in 2013. In 2015, she is pursuing her Master degree in Computer Science and Engineering from Government college of Engineering, Tirunelveli .Her research interest includes wireless sensor networks and cloud computing

Baby.D.Dayana received her Bachelor degree in Computer Science and Engineering from Maamallan Institute of Technology, Sriperumpudur Chennai,TamilNadu 2013. In 2015, She is pursuing her Master degree in Computer Science and Engineering at Government College Of Engineering, Tirunelveli. Her research interest includes Data mining and cloud computing.