A Survey on Different Text Categorization Techniques for Text Filtration

Shashank H. Yadav
Computer Technology Department
Yeshwantrao Chavan College of Engineering
Nagpur, India

Balu L. Parne
Computer Technology Department
Yeshwantrao Chavan College of Engineering
Nagpur, India

Abstract: Social Network has become an integral part of 21st Century Generation. It has touched every individual’s life in one or the other way. Daily tons of information is being shared on different social network platform. These sites provide various means for uploading and publishing the contents. These contents are being shared in various forms such as Text, Images, Videos, etc. These uploaded content may contains abusing words, explicit images which may be unsuitable for social platforms. As such there is no proper mechanism for restricting offensive contents from publishing it on these sites. These problems gave rise to an idea for our proposed approach. In our approach we are developing a prototype of social network platform in which we will apply some filtering mechanism to the uploaded contents and prevent them from publishing on social network platform. It basically focuses on text filtration and image filtration for any abusive and explicit content respectively. In this paper, we have done a survey on different text categorization techniques and algorithms used for text categorization. We started our survey by studying various text categorization techniques used for recognizing of offensive texts. Further, we studied various algorithms for implementing these techniques into our prototype.

Keywords: Social Network; Content Publishing; Offensive Words; Filtration; Prototype; Categorization

I. INTRODUCTION

Social networking sites (SNSs) have gained considerable popularity, moving peoples real-world social circles to the Internet and thus changing millions of user’s communication habits. In these SNSs, users can share material with other users. SNSs must check this new material before or after publication to detect, and thus withdraw inadequate or illegal content. Social network users can report other user’s content as inappropriate by arguing that it intrudes on their privacy rights. The large number of users who flag content as offensive makes moderation difficult for human administrators.

For this purpose in our approach we will try to use automatic filtering based approach that automatically filter offensive and explicit contents before publishing it on social networking sites. Our approach will be applicable to contents in the form of texts, images. In this project we are developing a prototype of SNSs for adolescent online users which will be beneficial for their online safety. In this prototype we are providing security to the social network by applying some filtration mechanism for the content shared on social networking platform. We focus basically on text and image contents which are frequently being uploaded on such sites. We divided our project mainly in two parts i.e. Text Module and Image Module.

For implementing this, we first focused on Text Filtration Module. This module will detect the offensive words from the uploaded contents that include comments, post, blogs, etc. and filter it out before publishing it on the social platform. For doing this we have done a survey on various text categorization techniques used for text filtration in NLP. After doing a three study of various techniques we will use the techniques which is best suited for our approach. The following section will give the detail information of different text categorization techniques.

II. RELATED WORK

G. Xiang et al [3] proposes a novel semi-supervised approach for detecting profanity-related offensive content in Twitter. His approach exploits linguistic regularities in profane language via statistical topic modeling on a huge Twitter corpus, and detects offensive tweets using these automatically generated features.

Weiming Hu, et al [4] proposes a novel framework in which Web pages are divided, using the C4.5 decision tree algorithm. The skin detection algorithm combines multidimensional histograms with the EM algorithm to speed up Gaussian mixture skin modeling without compromising the accuracy of the resulting model.

Félix Gómez Mármol, et al [5] proposes that SNSs could adopt reputation-based mechanisms to assess accusers’ behavior when reporting any content.

Ho and Watters et al [6] use the Bayes classifier to recognize pornographic texts. They not only consider the influence of different words on the weights of the Bayes network but also assign different weights to the same words when they appear in different Web page components such as title, metadata and body.

Lee et al [7], [8] count the frequencies with which keywords appear in a text. The frequencies, together with the relevant Web page features, are used as the input to the Kohonen self-organizing neural network (KSOM). After the learning stage is completed, the KSOM is used to determine whether a text can be classified as pornographic.

Du et al [9] extract feature vectors from pornographic and normal texts and save them in a database. To test a new text, its feature vector is extracted and matched to each of the saved feature vectors.

R. Du et al [9] uses decision making algorithm for classifying the offensive words from the texts.

Ying Chen et al [10] propose Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive users in social media.
III. DIFFERENT TECHNIQUES FOR TEXT CATEGORIZATION

A. KSOM (Kohonen Self-Organizing Neural Network)

In this for recognizing texts the approach mainly based on keyword matching and statistics. Frequencies are counted with which keywords appear in a text. Frequencies are used as i/p to the KSOM. After learning, KSOM (Kohonen Self-Organizing Neural Network) used to determine offensive texts [4]. Lee et al count the frequencies with which keywords appear in a text. The frequencies, together with the relevant Web page features, are used as the input to the Kohonen self-organizing neural network (KSOM) [4]. After the learning stage is completed, the KSOM is used to determine whether a text can be classified as pornographic.

B. CNN (Cellular Neural Network)

It represents the semantic connections between keywords. It is formed by connecting a number of cells, where each cell is a processor with multiple inputs and a single output. Each Cell is connected to a few adjacent cells. To construct a CNN-like word net, Hu et al. define a cell for each keyword. The associations between keywords are represented by connections between the corresponding cells. When a keyword is read, the cell corresponding to this keyword accepts it as an input. Then, the cell’s state is changed according to its previous state and the states of its adjacent cells.

C. Bayes Classifier for Discrete Texts

The semantic associations in discrete texts can be ignored, so only the frequencies with which the keywords appear in discrete texts are used as the features for recognition. Hu at al use the Bayes classifier to classify discrete texts. As the Bayes classifier outputs probabilities that a text belongs to the categories rather than a certain result that the text belongs to a category, and the probabilities are used for the fusion of texts and images.

D. BOW (Bag-of-Words) Model

The bag-of-words model is is used in NLP and IR [11]. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity [11]. The bag-of-words model is commonly used in methods of document categorization, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

E. Text Mining

The implementing text mining techniques to analyze online data requires the following phases [10]:
- Data acquisition and preprocess,
- Feature extraction, and
- Classification.

The major challenges of using text mining to detect offensive contents lie on the feature selection phrase, as follows:

1) Message-level Feature Extraction - Most offensive content detection research extracts two kinds of features: lexical and syntactic features [10].
   a) Lexical Features - Lexical features treat each word and phrase as an entity. Word patterns such as appearance of certain keywords and their frequencies are often used to represent the language model. Techniques used in this are Bags-of-Words (BOW) & N-gram approach.

b) Syntactic Features - To consider syntactical features in sentences, natural language parsers are introduced to parse sentences on grammatical structures before feature selection. Equipping with a parser can help avoid selecting un-related word sets as features in offensiveness detection.

2) User-level Offensiveness Detection - There are some limited efforts at the user level.

For example, Kontostathis et al [10] propose a rule-based communication model to track and categorize online predators. Pendar uses lexical features with machine learning classifiers to differentiate victims from predators in online chatting environment. Pazienza and Tudorache propose utilizing user profiling features to detect aggressive discussions. They use users’ online behavior histories (e.g., presence and conversations) to predict whether or not users’ future posts will be offensive.

F. Aho-Corasick String Matching Algorithm

It is an efficient algorithm for pattern matching. This algorithm is used for searching a text of specified keywords. Aho-Corasick is useful when you have a set of keywords and you want to find all occurrences of keywords in the text or check if any of the keywords is present in the given chunks of text [12]. It is much more efficient as compare to other string matching algorithm if you have keywords that don't change often. You should use this algorithm if you have a large set of keywords that remains unchanged.

G. N-Grams

N-Grams are the basic method for text categorization. It has an ability to classify the unknown text with highest certainty. N-grams are substrings generated from larger text [1]. In this approach the input text is divided into small parts with n maximum length, then the frequency of occurrence are counted. After the count of all n-grams, the least frequent are discarded and the rest is written to a profile. The profile of unknown text is then compared to each and every language (category) listed.

Here the number of n-grams stored may not be the same for input and language profiles [2]. Usually it is a good idea to put more n-grams in language profile.
IV. OUR PROPOSED WORK

The contributions of this research work that we carried out are:

- It will help us in our Text Module for filtrating offensive and abusive contents.
- In text module we will use keyword matching algorithm for detecting abusive keyword from comments, text messages, text contents uploaded in the form of paragraphs, etc.
- After detecting the offensive content by using one of the algorithms and techniques which we have discussed above then we will apply our filtration process and security measures for restricting them from publishing on the social platform.

Finally, this module will be used with other modules of our project for implementing our proposed architecture for building a social network prototype for adolescent online safety.

A. Text Filtration Module

- Content Upload Module
- Filtration Technique Implementation
- Abusive Keyword Detection

B. Text Restriction Module

- Pop up alert messages for users to give warnings.
- Replacement of abusive words with other suitable word as per the context.
- Restrict offensive content from publishing on social platform.

V. CONCLUSION

In this survey paper, we have done literature survey on different text categorization techniques and gone through the comparative study of different techniques. Considering the ability and ease of use of different text categorization we will try to implement it in our proposed work. Further, we need to work on the datasets which we are going to use for our project and also focus on creating a database of possible abusive keywords from online sources. Also a lot of work is to be done for the actual implementation of these techniques in social networks.

REFERENCES