**Detecting Offensive Language in Social Media to Protect Adolescent Online Safety**

Ying Chen1

1Department of Computer Science and

Engineering

The Pennsylvania State University

University Park, PA, USA

yxc242@psu.edu

Yilu Zhou2

2Department of Information Systems and

Technology Management

George Washington University

Washington, DC, USA

yzhou@gwu.edu

Sencun Zhu1,3, Heng Xu3

3College of Information Sciences and

Technology

The Pennsylvania State University*,*

University Park, PA, USA

{sxz16, [hxx4}@psu.edu](mailto:hxx4%7d@psu.edu)

*Dataset Description*

The experimental dataset, retrieved from Youtube comment boards, is a selection of text comments from postings in reaction to the top 18 videos.. Each text comment includes a user id, a timestamp and text content. The user id identifies the author who posted the comment, the timestamp records when the comment was posted and the text content contained a user’s comments. The dataset includes comments from 2,175,474 distinct users.

*Pre-processing*

an automatic pre-processing procedure assembles the comments for each user and chunks them into sentences., an automatic spelling and grammar correction process precedes introduction of the sample dataset to the classifier. With the help of WordNet corpus and spell-correction algorithm2, correction of spelling and grammar mistakes in the raw sentences occurs by tasks such as deleting repeated letters in words, deleting meaningless symbols, splitting long words, transposing substituted letters,and replacing the incorrect and missing letters in words.

*Experiment Settings in Sentence Offensive Prediction*

The experiment compares six approaches in sentence

offensive prediction:

*a) Bag-of-words (BoW)*: The BoW approach disregards grammar and word order and detects offensive sentences by checking whether or not they contain both user identifiers and offensive words. This approach also acts as a benchmark.

*b) 2-gram*: The N-gram approach detects offensive sentences by selecting all sequences of n words in a given sentence and checking whether or not the sequences include both user identifiers and offensive words. In this approach, N equals to 2, it also acts as a benchmark.

*c) 3-gram*: N-gram approach, selecting all sequences of 3 words in a given sentence. It also acts as a benchmark.

*d) 5-gram*: N-gram approach, selecting all sequences of 5 words in a given sentence. It also acts as a benchmark.

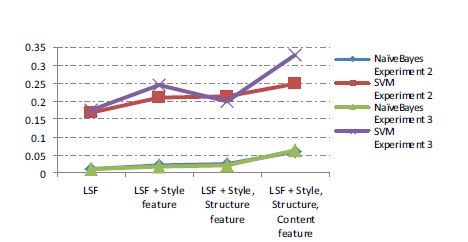
*e) Appraisal approach:* The appraisal approach was proposed for sentiment analysis [26], here we use it on

sentence offenisve detection for comparision. It can detect offensive sentences by going through all types of dependency sets and checking whether or not certain offensive words and user identifiers are grammatically

*LSF:* The sentence offensive prediction method proposed in this study Classifier

We further examined the classification rates for different feature sets using NaiveBayes and SVM classifiers. Since the rates vary from time to time, we run each instance 5 times and take the average., we find that the calculation rate of machine learning techniques is much faster than feature extraction time in Fig.6, the longest running time for machine learning classifiers is only 0.33s to predict users’ offensiveness. And the classification rate is independent on the number of users and the number of sentences. Generally, NaiveBayes works much faster than SVM in classification, but SVM produces more accurate classification results. it achieves precision of 77.9% and recall of 77.8% in user offensive detection using SVM,

\_\_\_\_ \_\_\_\_ \_\_\_\_ \_\_\_\_



Draw backs

Because of their use of simple lexicon-based automatic filtering approach to block the offensive words and sentences, these systems have low accuracy and may generate many false positive alerts. In addition, when these systems depend on users and administrators to detect and report offensive contents, they often fail to take actions in a timely fashion

BoW

approach alone not only yields low accuracy in subtle

offensive language detection, but also brings in a high false

positive rate especially during heated arguments, defensive

reactions to others’ offensive posts, and even conversations

between close friends. N-gram approach is considered as an

improved approach in that it brings words’ nearby context

information into consideration to detect offensive contents