Poster: oFBI: Detecting Offensive Language in Social Networks for Protection of Youth Online Safety

Ying Chen[1]
[1]Department of Computer Science and Engineering,
The Pennsylvania State University,
University Park, PA 16802
yxc242@cse.psu.edu

Yilu Zhou[2]
[2] Department of Information Systems and Technology Management,
George Washington University,
Washington, DC 20052
yzhou@gwu.edu

Sencun Zhu[1,3]
szhu@cse.psu.edu

Heng Xu[3]
[3]College of Information Sciences and Technology,
The Pennsylvania State University,
University Park, PA 16802
hxu@ist.psu.edu

ABSTRACT
Highly developed online social networks promote flagrant use of profanities, obscenities, and insults, potentially harmful to youths' mental health. This project proposes a framework, Online Friend Background Investigator (oFBI) to provide warning information to youth when they receive new friend requests, thereby avoiding potential cyber offenders. The study's methodology explores lexical semantic features in users' conversation histories to predict offensiveness. The study demonstrates an approach to provide highly accurate and precise performance, with an acceptable data processing rate for deployment on online social networks and other real-time online communities.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information filtering, Evaluation/methodology; H.5.2 [Information Interfaces and Representation]: User Interface—Graphical user interface.

General Terms
Algorithms, Measurement, Design, Security, Languages.

Keywords
Cyberbullying, Youth, Offensive Language, Social Network.

1. INTRODUCTION
Online social networks (OSN) create new ways to socialize and interact; however, highly developed OSNs encourage flagrant use of profanities, obscenities, and insults, potentially harmful for adolescents' mental health. Since teenagers by nature sensation seeking, they are more likely take greater risks than younger children and adults. Cyber-offenders can easily access youth's profiles and harass them once they become "friends" online. Thus, an investigation on *friend requesters*[1] is critical to guard against and isolate potential cyber-offenders.

This study proposes a lexical semantic approach (LSA) to predicting online users' offensiveness levels. The experimental results show that the proposed approach can correctly categorize 94.34% of offensive sentences and 98.24% of non-offensive sentences. LSA can achieve classification performance better than that obtained from learning-based methods, at a processing speed around 10msec per sentence. These results suggest our effective deployment of LSA on the server side of OSNs.

2. METHODOLOGY
A lexical semantic approach (LSA) is proposed to tackle the challenges in user offensiveness estimation. Figure 1 shows the overall architecture of LSA, which includes two major steps: sentence offensiveness analysis and user offensiveness aggregation. In sentence offensiveness analysis step, we search for lexical units (pejoratives) in sentences, and refine the offensiveness of each lexical unit by looking at its related words in sentences. Then the offensiveness values of all sentences are synthesized to compute the overall offensiveness of the user.

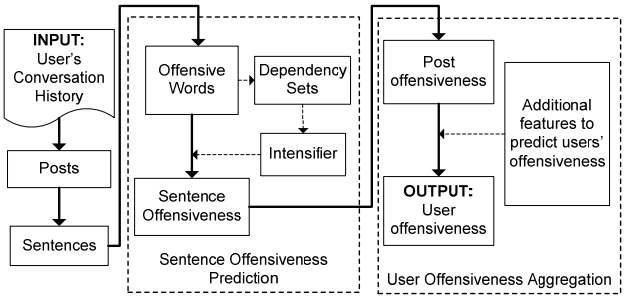

Figure 1 Framework of LSA

2.1 Sentence Offensiveness Prediction
Offensive messages always include offensive words. Strongly offensive words, such as "f***" and "s***", are conventionally and generally offensive; but other weaker offensive words, such as "stupid" and "liar", are less identified. The study differentiates between these two types of offensive words, and assigns offensiveness levels accordingly. Hence, the definitions of offensiveness level of each pejorative, *w*, in sentence, *s*, is:

$$p_w = \begin{cases} a_1 & \text{if } w \text{ is a strongly offensive word} \\ a_2 & \text{if } w \text{ is a weakly offensive word} \end{cases}$$

where $a_2 < a_1 \leq 1$.

Once a pejorative describes an online user, or semantically associates with another pejorative, it becomes more offensive from users' perceptions. Thus, an intensifier[1] is needed to scale the offensiveness value of words. We use typed dependency parser, proposed by *Stanford Natural Language Processing Group*[2], to capture the grammatical dependencies within a sentence. If

[1] Friend requesters refer to users who send friend requests to others.

Symposium On Usable Privacy and Security (SOUPS) 2011, July 20-22, 2011, Pittsburgh, PA, USA.

[2] http://nlp.stanford.edu/software/lex-parser.shtml

sentence, *s*, is parsed and all words semantically related to pejorative, *w*, are categorized as set $D_{w,s} = \{d_1, \cdots d_l\}$, the intensifier, I_w, of word, *w*, can be defined as:

$$I_w = \begin{cases} b_1 & if\ \exists d_i \in D_{w,s},\ d_i\ is\ a\ user\ identifier \\ b_2 & if\ \exists d_i \in D_{w,s},\ d_i\ is\ an\ offensive\ word \\ 1 & otherwise \end{cases}$$

where $b_1 > b_2 \geq 1$.

Consequently, the offensiveness value of sentence, *s*, becomes a determined linear combination of words' offensiveness, $o_s = \sum p_w I_w$.

2.2 User Offensiveness Aggregation

Synthesizing the offensiveness values of all sentences allows computing the overall offensiveness of the user. Thus, given a post, p, containing sentences, $\{s_1, \ldots s_n\}$, and the offensiveness of the sentences, $\{o_{s_1}, \cdots o_{s_n}\}$, the offensiveness, O_p, of p should be, $O_p = \sum o_s$. Hence, the offensiveness value, O_u, of a user who has m posts is, $O_u = 1/m \sum O_p$, because users who frequently post offensive messages are more offensive than occasional offenders. If $O_u > 0$, then the conversation history of user, u, indicates offensiveness to some extent.

However, other features such as the punctuations used, the constructed manner of sentences, and the organization of sentences within posts could also affect others' perception of the poster's offensiveness. Thus, the study adds three types of features[2]—style features, structural features, and content-specific features—to help identify online offenders, and utilizes machine learning techniques to perform the evaluations.

2.3 User Interface Design

One way to implement oFBI is to calculate and store every user's offensiveness value on the server-side of social networks, and update the values when users have any additions to their user-spaces. When a user sends a friend request to an adolescent, the senders' offensiveness value, as complementary information, is transmitting along with the friend request. Figure 2 depicts a sample user interface of oFBI.

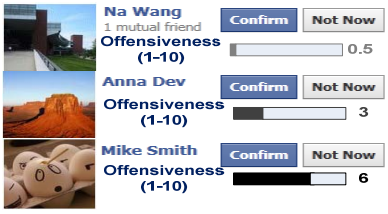

Figure 2 A sample display of oFBI on OSNs

3. RESULTS

The experimental dataset, retrieved from Youtube comment boards, is a selection from the top 18 most popular videos, which received 6,737,683 comments from 2,175,474 distinct users. The experiment uses three learning-based approaches as baselines for sentence offensiveness prediction:

- Bag-of-words (BoW) approach: The BoW approach disregards grammar and word order and detects offensive sentences by checking whether or not they contain user identifiers and offensive words.
- N-gram approach: The N-gram approach detects offensive sentences by selecting all sequences of n-words in a given sentence and checks whether or not the sequences include user identifiers and offensive words.
- Appraisal approach: The Appraisal approach detects offensive sentences by checking whether or not phrases in a given sentence direct certain offensive words towards an online user.

We randomly select a uniform distributed sample from the complete dataset. The sample includes total 3400 sentences and our classifier detects 333 offensive sentences. The experimental parameters are a_1=1; a_2=0.5; b_1=2, and b_2=1.5. The accuracy and speed of the proposed sentence-level offensiveness prediction appear in Table 1 and Figure 3.

Table 1 Accuracies of Sentence Level Offensiveness Detection

Methods	Accuracy	FP	FN	Precision	F-score
BoW	66.88%	9.32%	33.13%	90.68%	76.98%
2-gram	33.75%	3.57%	66.25%	96.43%	50.00%
3-gram	46.25%	3.90%	53.75%	96.10%	62.45%
5-gram	61.88%	5.71%	38.13%	94.29%	74.72%
Appraisal	66.25%	0.93%	33.75%	99.07%	79.40%
LSA	94.34%	1.76%	5.66%	98.24%	96.25%

Significant at $\alpha = 0.05$

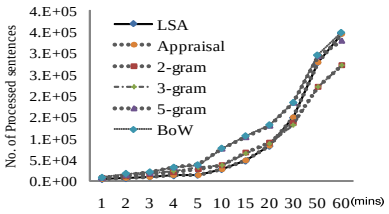

Figure 3 Data processing time

4. DISCUSSION

According to Table 1, the bag-of-words approach can identify most obviously offensive sentences. However, the technique generates a high false positive rate because it captures numbers of unrelated <user identifier, offensive word> sets. The accuracy of n-gram is low when n is small. However, as n increases, the false positive rate increases as well. Moreover, none of the baseline approaches provides false negative rates less than 33%, because many of the obviously offensive sentences are imperative sentences, which omit all user identifiers. However, simply using an offensive word as the only detection feature produces an even higher false positive rate. LSA obtains its highest F-score because it balances well the precision-accuracy tradeoff. From Figure 3, the processing rate of the proposed LSA is at least equal to other approaches. Thus, the proposed method is practical for application to OSNs and other real-time online communities.

5. REFERENCES

[1] Zhang, C., D. Zeng, J. Li, F.Y. Wang, and W. Zuo, 2009, Sentiment analysis of Chinese documents: from sentence to document level. Journal of the American Society for Information Science and Technology. 60(12): p. 2474-2487.

[2] Zheng, R., Y. Qin, Z. Huang, and H. Chen, 2010, Authorship analysis in cybercrime investigation. Intelligence and Security Informatics: p. 959-959.