The Pennsylvania State University

The Graduate School

Department of Computer Science and Engineering

DETECTING OFFENSIVE LANGUAGE IN SOCIAL MEDIAS FOR PROTECTION OF ADOLESCENT ONLINE SAFETY

A Thesis in

Computer Science and Engineering

by

Ying Chen

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2011

The thesis of Ying Chen was reviewed and approved* by the following:

Sencun Zhu
Associate Professor of Computer Science and Engineering & Information Sciences and Technology
Thesis Co-Advisor

Heng Xu
Assistant Professor of Information Sciences and Technology
Thesis Co-Advisor

Wang-Chien Lee
Associate Professor of Computer Science and Engineering

Raj Acharya
Professor of Computer Science and Engineering
Head of the Department of Computer Science and Engineering

*Signatures are on file in the Graduate School

ABSTRACT

Currently adolescents highly rely on social media to interact with other people. Given the complicated environment of social media, it has become very difficult for adolescents to avoid encountering offensive content from time to time. Since the textual content on online social media is highly unstructured, informal, and often misspelled, existing research on message-level offensive language detection cannot accurately detect offensive content, and user-level offensiveness evaluation is still an underresearched area. To bridge this gap, we propose Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive user in social media. We distinguish the contribution of pejoratives/profanities and obscenities in determining offensive content, and introduce hand-authoring syntactic rules in identifying name-calling harassment. In particular, we incorporate users' writing style, structure and specific cyberbullying content as features to predict users' potentiality to send out offensive content. Results from experiments showed that LSF framework achieved significantly better performance than existing methods in offensive content detection. It categorizes 94.34% of offensive sentences and 98.24% of non-offensive sentences, and 90.2% of offensive users and 86.3% of non-offensive users. Meanwhile, processing speed of LSF is approximately 10msec per sentence, suggesting the potential for effective deployment on online social media. We believe such language processing model will greatly help to online offensive language monitoring, eventually to build a better online environment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Currently, people are spending more and more time on social media to connect with others, to share a wide variety of information, and to pursue common interests. Seventy five percent of U.S. households(Diana, 2010) now use social networking sites; 83% of 18-29 year-old(Ries, 2011) Americans are using social media, with 61% doing so every day. Adolescents, especially, devote the majority of their online time to social media engaging with their peers. In 2011, 70% of teens use social networking sites on daily basis(Timothy Johnson et al., 2011); nearly one in four teens hit their favorite social-media sites 10 or more times a day(Gwenn Schurgin O'Keeffe et al., 2011); 75 % of teens own cell phones, and 25% use them for social media. Studies have shown that youth use social media to maintain their existing friendships and as a means to develop interests beyond what they have access to at school or in their local community by online gaming, creative writing, video editing, or other artistic endeavors. Most youth use online networks to associate with people they already know in their offline lives (Ito et al., 2008); 61% of teens communicate with their friends by sending messages through social networking sites, while 42% of them do so on daily basis (Lenhart et al., 2007). Nearly half (49%) of teens use social network sites to make new friends (Smith, 2007). With the heavy traffic social media sites attract, they affect large groups of people.

Adolescents rely on social media to interact with and learn from other people (Ito et al., 2008). Due to the complicated environment of social media, they are at the risk of being exposed to large amounts of offensive content online through offensive content in messaging, wall posts, or comments. While there is no universal agreement as to what is "offensive," for the purpose of

this study, we employ Jay and Janschewitz's (2008) definition of offensive language as vulgar, pornographic, and hateful language. Vulgar language refers to coarse and rude expressions, which includes explicit and offensive reference to sex or bodily functions; pornographic language refers the portrayal of explicit sexual subject matter for the purposes of sexual arousal and erotic satisfaction; hateful language includes any communication outside the law that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, and religion. All of these are generally immoral and harmful for adolescents' mental health. ScanSafe's monthly "Global Threat Report"(Cheng, 2007) for March 2007 found that up to 80 percent of blogs host offensive content. In this number, seventy-four percent contain porn in the format of image, video, and offensive language. In a limited, small-scale analysis of chat transcripts from two of the most popular teen sites, chat participants had 19 percent chance exposing to negative racial or ethnic remarks in monitored chat and 59 percent chance in unmonitored chat (Tynes et al., 2004). In addition, cyber-bullying appears in the way of writing offensive messages via the use of social media. The National Center on Addiction and Substance Abuse at Columbia University reports that 19 percent of teens report that someone has written or posted mean or embarrassing things about them on social networking sites.

Unfortunately, adolescents are more likely to be negatively affected by this biased and harmful content than adults (Lee & Leets, 2002). They will 1) become low self-esteem, afraid of outside world; 2) become "immune" or numb to the horror of abusive language and behavior; 3) gradually accept abusive language and behave as a way to solve arguments; and 4) imitate the offensive language and even the behavior. Furthermore, new press and studies found that children and adolescents were engaged in producing online hate speech (Tynes et al., 2004), 3% of adolescents participated in cyber solicitation in 2008 (Finkelhor et al., 2008), and 13% of adolescents cyber-bullied others in 2010 (Hinduja & Patchin, 2008). Different entities have been aware of the situation and made efforts to eliminate offensive content online for youth protection,

but none of them has fully addressed the problem. There are numbers of online websites such as NetSmarz, Teachtoday, iKeepSafe, WiredSafety, Teenangels, girlscounts and PBS kids that educate adolescents how to act against offensive content on social media. Additionally, the Children's Internet Protection Act (CIPA) was enacted by congress in early 2001 to address concerns about access to visual offensive content over the Internet on school and library computers. CIPA only concerns about image content; unfortunately, there is much more unstructured, despicable textual online content out there than multimedia materials. Currently, administrators of social media manually review the online content to delete offensive materials. However, due to the limited qualified manpower and lack of an automatic system, the review tasks of finding the offensive content are labor intensive, time consuming, and not sustainable in the long term. Some automatic content filtering and parental control software, such as Appen[1] and Internet Security Suite[2], has been developed to detect and filter online offensive content. They aggressively blocked webpages or paragraphs based on dirty word appearance, largely affecting the readability of content on social media. One major problem with the word-based approaches is that it fails to identify a subtle offensive message if none of its terms is strongly offensive. For example, the sentence "you are such a crying pig" is offensive, but none of its words is included in general offensive lexicons. Another problem with these word-based approaches is the high rate of false positives. This is due to the word ambiguity problem—dirty words can be used in conversations between intimate friends, while certain words are mistaken as offensive with these word-based approaches. Moreover, existing methods only detect offensive language on the message level without tracing the source of offensive content. Since none of the current techniques can 100% detect offensive content, adolescents who keep connection with offensive

[1] http://www.appen.com.au/index.cfm?pageid=103

[2] http://shop.ca.com/malware/internet_security_suite.aspx?ggus=36640429&gclid=CJ3LhJmsnZ8CFdA65QodnV5BRQ

users or websites will continually be affected by the remaining harmful content. Therefore, a feasible solution is needed for improving the deficiency of offensive content detection in social media.

To address these challenges, we propose the LSF (Lexical Semantic Feature) language model to effectively detect offensive language in the social media for adolescence protection. LSF provides high accuracy in subtle offensive message detection, and it can eliminate the false positive rate. Besides, LSF not only checks messages, but also the person who posts the messages and his patterns of posting. LSF can be implemented in client side applications for individuals or groups who are concerned for their adolescents' online safety. It is able to detect whether or not online users or websites push recognizable offensive content to adolescence, then trigger applications to alarm the senders to regulate their behavior, and eventually block them if this situation continues. Users are also allowed to adjust the threshold of acceptable level of offensive content based on their own perception of online safety. Our language model may not be able to make adolescents immune to offensive content, because it is hard to fully define what is "offensive." However, we aim to provide a much improved automatic tool to detect most offensive content in social media, so that parents and their adolescents can have better control over the content they are viewing.

The remainder of the paper's organization includes Chapter 2, which reviews existing literature of offensive language detection; Chapter 3, which describes our research gaps and questions; Chapter 4, which introduces, in detail, the proposed LSF approach for detecting offensive content and predicting users' offensiveness in social media; Chapter 5, which presents comparative experiments and the results; and Chapter 6, which examines the limitations of LSF and suggests an outline for future research. The final section concludes the paper with a summary of the contributions of the research.

Chapter 2

Related Works

In this section, we exam the existing offensive content filters on social media, review previous offensive detection research, and analyze the benefit to trace the source.

Existing Offensiveness Content Filters on Social Media

Offensive content has raised alarm in social media for a long time, so there are already several defense mechanisms. Radio programs are heavily monitored by the Federal Communications Commission (FCC) for lewd content and vulgar language. Some radio programs are broadcast with a delay for the purpose of giving the radio station a chance to prevent offensive content from airing. Similarly, when played over the airways, offensive language in music is replaced with silence or bleeped out of existence. As the popularity of radio fades, the primary concern of parents and legislators shifted to television, and more recently, video games. Since television and video games carry mostly multimedia content, and we only focus on textual content in this study, offensive television and video game content will not be discussed here.

As to other social media, blogs encourage their users to avoid posting offensive material and language on their blogs, as well as to contact advertising services to get rid of potentially offensive advertisements. Blog service provider Wordpress monitors its blogs constantly. If it finds that users regularly post material, which is offensive, not safe for work, or not suitable for minors, it will flag their blogs as mature, and exclude them from the global tag lists so that other users cannot search them. Similarly, forums such as indeed2 and ABRSM3 rely on system administrators' observations and users' reports to edit and remove the offensive content; their

members who persistently post offensive messages risk having future postings pre-moderated or user profiles suspended.

Popular online social networking sites have applied several mechanisms to screen offensive content for users' protection. Youtube's safety mode for users, once activated, can hide all comments containing offensive language from users. But if users seek to explore the hidden comments, they can simply click "Text Comments," and pre-screened content will appear—the pejoratives replaced by asterisks. On Facebook, users can add comma-separated keywords to the "Moderation Blacklist." When people include blacklisted keywords in a post and/or a comment on a page, the content will be automatically identified as spam and screened. Twitter client, "Tweetie 1.3," was rejected by Apple Company for allowing foul language to appear in users' tweets. Currently twitter does not pre-screen users' posted contents, claiming that if users encounter offensive content, they can simply block and cease to follow people posting offensive content. Similarly, MySpace encourages users either to move offensive mails to a junk mailbox and report the communications as spam or remove the senders from friend lists. LinkedIn also allows users to report inappropriate photos.

We summarize the features of social media actions against offensive content in terms of lexicon used and techniques used and present them in Table **2-1**. We include two types of lexicons: media predefined lexicon (L1) and user define lexicon (L2). We also include three types of techniques: blocking the keywords (T1), blocking the content containing the keywords (T2), and preventing the source from posting more offensive content (T3).

Table 2-1 Current social media acts against offensive content

Medias	Lexicons		Techniques			Take action in advance
	L1	L2	T1	T2	T3	
Radio	√		√			√
Blog			√(user)	√(user)	√(manually)	
Forum			√(user)	√(manually)	√(manually)	

Facebook		√		√	√(user)[3]	√
Twitter					√(user)	
Youtube	√		√	√		√
Myspace				√(user)	√(user)	
Google+					√(user)	
LinkedIn				√(user)	√(user)	

According to Table 2-1, we find that the majority of popular social media do not define offensive content beforehand, so they fail to detect and filter out offensive content before viewed by users. In addition, they simply rely on the users eliminating offensive content. For youths who lack cognitive awareness of risks, these approaches are hardly effective to block offensive online content. Therefore, parents need additional software and techniques to efficiently detect offensive material in online communities to protect their children under 18 years old from exposure to vulgar, pornographic and hateful language.

Two types of tools that automatically analyze online conversation and detect offensive content are in current use. Content analysis software, such as Appen data stream profiling tools, captures online communications from different channels such as keystrokes, in-browser text editors and instant messaging clients, and generates alerts based on pre-defined anomaly profiles. Parental control software, such as Internet Security Suite, K9 Web Protection and OnGuard Online can record children's online activities and detect harassment based on parent-selected keywords. Both types of software rely on a pattern-matching method, which determine content offensiveness by detecting the appearance of the predefined patterns such as words, phrases and expressions. However, due to word ambiguity problems, the pattern-matching method commonly generates high false positives, eventually overloading parents seeking to protect their children.

[3] It means this one relies on user's report.

Current Techniques for Online Offensive Message Detection

High false positive rates of the pattern-matching method often occur because the text content on online social media is unstructured, informal, and often misspelled. To make computers intelligently identify offensive content, text mining and machine learning approaches were developed to enhance the quick analysis of the text-based data. To understand how the text mining techniques worked in offensive language detection, Table **2-2** shows the taxonomy of the previous studies based on three characteristics: preprocessing methods, feature types, and classification approaches.

Table 2-2 Taxonomy of the previous studies based on three characteristics

Category	**Description**	**Label**
Preprocess methods		
Syntactic Parsers	Natural language parser, Part-of-Speech (POS).	P1
Domain Knowledge	Encoding terms or phrases using offensive word dictionaries, filtering data by pre-defined rules; domain experts manually creating, grouping terms. .	P2
Semantic Feature Types		
Lexical Feature	Bag of Words (BoW), N-gram, TFIDF.	F1
Sentiment Feature	Include pronoun, subjective terms.	F2
Contextual Feature	Similarity feature, Contextual post feature.	F3
User Profiling Feature	User activity feature, Local user activity feature	F4
Classification Approaches		
Rule-based Approach	Keyword/phrase matching, Pattern matching, Rule-based decision table.	C1
Machine Learning Approach	Supervised learning, Unsupervised learning, Support vector machine (SVM), Naïve Bayes classifier, Decision tree, K-nearest neighbor (k-NN) classifier.	C2

Implementing text mining techniques using online text-based data usually includes the following phases: data acquisition and preprocess, feature generation and text representation, classification and evaluation. In the data acquisition and preprocess phase, selected textual data is crawled from online social media and parsed by using either the syntactic parser to grammatically separate the sentences or domain knowledge rules to preprocess sentences into structured formats.

In the feature generation and text representation phase, three kinds of features can be generated from the data: lexical, sentimental and contextual. Lexical features, including Bag-of-Words (BoW), N-gram, etc, present the text as a set of vectors with words and phrases as its elements. Sentimental features capture users' sentiment inside the text by searching for subjective terms and pronouns. Contextual features refer to the semantic and syntactic dependency of sentences or paragraphs. User profiling features capture users' online behaviors, such as their presence, and whether or not their mostly history conversation are offensive. In the classification and evaluation phase, two kinds of approaches are commonly used: rule based approach, and machine learning approach. The rule based approach checks whether the features satisfy some pre-selected rules. For example, Smokey(Spertus, 1997) use second-person rule as one of its criteria to identify online hostile messages because many sentences with a word beginning with "you" (including "your" and "yourself") are insulting. The machine learning approach includes supervised learning and unsupervised learning. Supervised learning refers to the data mining process that firstly textual data is labeled as offensive or inoffensive, and then it will be fed to the classifier to select and weight important semantic and syntactical features, which can later be used to classify unknown textual data as offensive or non-offensive. However, unsupervised learning tries to find hidden structure in unlabeled data, so in unsupervised learning process, training data don't need to be labeled before being fed to the classifier. Table 2-3 presents the summary of important text mining studies with emphasis on offensive message detection.

Table 2-3 Summary of important text mining studies in offensive detection

Authors & Years	**Preprocess Methods**		**Feature Types**				**Classification Approaches**		**Issues**
	P1	P2	F1	F2	F3	F4	C1	C2	
Pazienza & Tudorache, 2011		√	√	√		√		√	Interdisciplinary(psycho/cognitive/linguistic) study on frame modeling in online Italian forums

Razavi et al 2010		√	√				√	√	Offensive language detection using multi-level classification.
Kontostathis & Leatherman, 2009		√	√					√	Tracking and categorization of internet predators.
Yin, D., et al. 2009			√	√	√		√		Detection of harassment on Web 2.0
Mahmud et al 2008		√	√	√			√		Detecting flames and insults in text.
Spertus, 1997	√	√	√	√				√	Automatic recognition of hostile messages.

Early researchers (McEnery et al., 2000a; McEnery et al., 2000b; McEnery & Xiao, 2004; McEnery, 2006) were only using BoW and pattern matching to detect offensive messages. Offensive words may be filled into categories such as religion (e.g. “Jesus”, “heaven”, “hell” and “damn”), sex (e.g. “f***”) (asterisks replace letters of strongly offensive words), racism (e.g. “nigger”), defecation (e.g. “s***”), homophobia (e.g. “queer”) and other matters. However, as discussed in above, only using BoW always brings in high false positive rate because those words can also be used in heat arguments, reaction to others’ offensive posts, and even in conversation between close friends. Later, researchers using N-gram to detect offensive messages (Kontostathis et al., 2010; Pendar, 2007) also encounter problems. N-grams represent subsequences of N items from given sequences. The items can be phonemes, syllables, letters, words or base pairs according to the application. Tri-gram, for example, identify that “you” and “stupid” may related in the sentence, “You don’t notice that you are stupid.” because they are less than three words away. But N-gram suffers from critiques that it has difficulty exploiting related words if they are separated by long-distances within sentences. For example, “you” and “idiot” are seven words apart in the sentence, “You, by any means, are an idiot,”, but they are related. If simply increasing N to solve the problem, the system processing speed may become slow, followed by more false positives. Since lexical features are not sufficient, Yin et al (2009) explored all the lexical, sentiment and contextual features to detect harassing online messages. They use TF-IDF(Wu et al., 2008) (TF means term frequency, the occurrence count of a term in a

document; IDF means inverse document frequency, a measure of the general importance of the term, obtained by dividing the total number of documents by the number of documents containing the term) to select important terms in documents. They also detect whether bad words and pronouns are contained in the same sentences as sentiment features, and further distinguish normal messages from harassing ones by contextual features: 1) posts different from the thread average have the potential to be harassment (similarity feature); 2) the cluster of posts near a harassment post should look different from the cluster of normal posts because of the reaction to harassment posts may affect users' writing styles (contextual post feature). Then, Razavi et al (2010) use machine learning classifier to select the most discriminative words and expressions as features. However, all the above methods only consider semantic features in sentences. Without considering the syntactical structure of messages, they fail to distinguish sentences' offensiveness which contain same words but in different orders, such as "Your mother said that it is a stupid pig" (a normal one) from "It said that your mother is a stupid pig" (an offensive one). Therefore, to consider syntactical features in sentences, natural language parsers are needed to parse sentences on grammatical structures before feature selection (Mahmud, 2008; Razavi et al., 2010; Spertus, 1997; Xu & Zhu, 2010). The parsing results of sentences presented as combinations of a dependency-type and word-pair with the form (governor, dependent). The governor and dependent can be any syntactic elements of sentences. For example, appos (you, idiot) in the sentence "You, by any means, an idiot." means that "idiot" is an appositional modifier of the pronoun "you." Equipped with a parser can help avoid selecting un-related word sets as features in offensive detection.

Rule-based approach is a common used method utilizing predefined rules to classify offensive messages. Smokey (Spertus, 1997), used by Microsoft in commercial applications, uses a rule-based analysis process for hostile messages detection. It can correctly detect 98% of the acceptable messages but only 64% of flame-type messages. The major reason for the high false

positive rate of Smokey is that its semantic rules are too general to be directly useful. Mahmud et al (2008) proposed another rule-based method to detect flames and insults in text. They use domain knowledge to create an extensive set of rules to extract all the possible semantic elements offensive messages could contain. However, their rule-based classifier required such a restricted form of input that raw data needed to go through a lengthy preprocessing phase (which has not yet proved capable of handling all exceptions). Hence, for some excessively long or complicated sentences, this system is very likely to generate errors and is unlikely to be effective for online communities' frequent informal expressions. Similarly, the rule-based approach (D Riffe, 1998; Kontostathis & Leatherman, 2009; Leatherman, 2009; Olson et al., 2007) are used in communication-based approaches to study luring behaviors in online instant messaging platforms. The authors firstly use BoW to catch words connected to certain behaviors in users' messages, and then use pre-defined rules to match the sequence of detected behaviors to determine whether or not users are online offenders. For example, online luring can be defined as a process consisting of three phases: approaching, luring, and asking for personal information. Once users utilize the words indicating they are approaching, then luring, and lastly asking for personal information, they are identified as online offenders. However, the performance of communication-based approaches highly depends on the effectiveness of the selected rules. Since online users' behaviors vary significantly and casual and informal English diction style is common, deployment of communication-based approaches in online communities is problematic.

To overcome the limitation of rule-based methods, researchers switch to machine learning approach to maximum the detection rate of offensive messages. Machine learning approach such as support vector machine, Naïve Bayes classifier, decision tree, k-NN classifier, etc, can help to select the most discriminative words and expressions as features (Mahmud, 2008; Pazienza & Tudorache, 2011; Pazienzaa et al., 2008; Razavi et al., 2010; Sjöbergh & Araki, 2008), and classify messages as positive (in our case is offensive) or negative (inoffensive) when

the importance of different features is unknown. The limitation of using machine learning techniques is that the input dataset should include a balanced number of positive instances and negative instances, and the input instances require careful verification beforehand to avoid over-fitting problem.

Many of contemporary online offensive language researches only focus on sentence-level and message-level constructs, which cannot fully protect adolescents. Since no detection technique is 100% accurate, if users keep connecting with the sources of offensive content—online users or websites—their will continuously being exposed to offensive content. However, studies associated with user level analysis are largely missing. In the existing works, Kontostathis et al (Kontostathis et al., 2010; Kontostathis & Leatherman, 2009; McGhee et al., 2011; Thom et al., 2011) propose rule-based communication model to track and categorize Internet Predators, while Pendar (2007) uses N-gram and TFIDF feature with SVM and K-NN classifier to analyze chat transcripts to differentiate between the victim and the predator. Pazienza & Tudorache (2011) propose to incorporate user profiling features in online frame detection: general impact of user activity features, and local user activity features. General impact of user activity features captures whether or not users have good presence on forums and users' most posts are no flames; while local user activity features detect users' presence and activity on a specific document/topic. The authors have proposed an interesting direction to help detect offensive content by considering users' online behavior. More detailed information such as users' writing styles and structures may further help to increase the detection rate of online offensive content. We also examine several studies on online economic reputation systems on eBay(Dellarocas, 2000; Houser & Wooders, 2006; Resnick et al., 2000; Resnick & Zeckhauser, 2002; Zacharia et al., 2000). They only rely on the rating scores to predict sellers' or buyers' reliability. While rating may work well in reputation systems, it cannot be used to correctly predict online users' offensiveness, because

users' own posts should be the most direct clue for their offensiveness rather than others' feedbacks.

Other Related Text Mining Researches

Other than offensive detection, many researchers in Artificial Intelligence and Natural Language Processing have been working on different kinds of opinion extraction or sentiment analysis (Dave, 2003; Gordon, 2003; Pang et al., 2002; Riloff & Wiebe, 2003; Riloff et al., 2003; Turney & Littman, 2003; Yi et al., 2003; Yu & Hatzivassiloglou, 2003) also adopt the standard text mining procedure as above. In many cases detecting the level of intensity of moods or attitudes could be an effective attribute for offensive language detection. Furthermore, subjective language recognition could also be useful in offensive content detection (Spertus, 1997; Wiebe et al., 2001). Hence, the subjective language detection is a task for which offensive content detection could be considered an offspring. Machine learning algorithms are largely used in this area for classifying texts based on some of their constituent words and expressions (Bruce & Wiebe, 1999; Wiebe et al., 2001; Wiebe et al., 1999).

In summary, current research of offensive language detection has not fully addressed the problems of detecting offensive sentences in social media, and few existing studies consider user offensiveness evaluation. While controlling online content for adolescents' safety is increasingly important, no known methods provide high precision and accuracy rates, high processing speeds, and tolerance for misspelling and grammar errors. A new approach to satisfying these deficits is necessary to effectively protect youths from offensive content in social media.

Chapter 3

Research Gaps and Questions

Based on our review in the previous section, we have identified several important research gaps.

Firstly, existing defense mechanisms on social media against offensive content highly rely on users' self-reports, and there are few contemporary researches in offensive detection, which can provide high precision and accuracy rates, high processing speeds, and tolerance for misspelling and grammar errors .

Secondly, studies associated with user level analysis are largely missing.

From the research gaps identified above, this study aims to answer the following research questions:

1. How to develop automatic tools to detect offensive content in social media, which can provide high precision and accuracy rate, high processing speed, and easily adapt to any format of English writing style?
2. How to predict users' potentiality to send out offensive content?

Thus, given online users' conversations, the research goal for this task is to develop an efficient language model. The language model should be able to accurately and quickly determine whether or not there is problematic content which is harmful to minors inside these conversations, and predict undetected problematic words based on their context. Besides, given online users' conversations, we also want to predict their potential to send out offensive messages. If they have high chance to become or they already are offensive, youth should have the right to know that and disconnect with them as soon as possible.

Chapter 4

System Design

In order to tackle challenges above, we propose a cascaded system for text mining on social media content. Figure 4-1 illustrates the proposed architecture LSF (Lexical Syntactic Feature), which can detect offensive content and identify potential offensive user in social media. The system consists of two major components: sentence offensiveness prediction and user offensiveness aggregation. We firstly chunk users' conversation history into posts, then sentences. For each sentence, its offensiveness can be referred by the fact that whether or not it contains offensive phrases, and whether or not the offensive phrases are used to describe other users. Later the sentence offensiveness can be synthesized to compute the overall offensiveness of users.

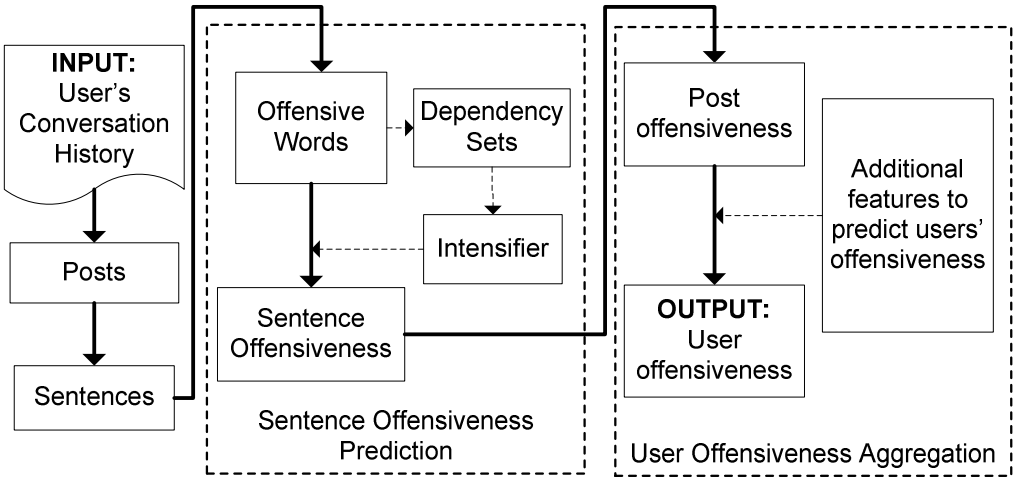

Figure 4-1 Overall architecture of LSF

Sentence Offensiveness Prediction

In this paper we propose a new method of sentence offensive analysis based on offensive word lexicons and sentence syntactic structures to bridge the gap of the previous methods for sentence offensive detection.(Mahmud, 2008; Razavi *et al.*, 2010; Spertus, 1997; Yin, Xue, Hong, Davison*, et al.*, 2009).

A user's offensiveness can be inferred from the offensiveness of his/her previous posts, which can be further chunked into sentences. Not all the sentences are offensive. Offensive sentences always associate with the occurrence of pejoratives, profanities, or obscenities (later pejoratives are used to denote all the three). Strong obscenities, such as "f***" and "s***", are conventionally and generally offensive; but other weaker pejoratives and profanities, such as "stupid" and "liar," are less offensive. This research differentiates between these two levels of offensiveness. The offensive word lexicon used by this research consists of the lexicon used in study (Xu & Zhu, 2010) and a lexicon, based on Urban Dictionary[4], established during the coding process. Pejoratives receive the label of strongly offensive word if more than 80% of its use is offensive ways. Otherwise, known pejoratives receive the label of weakly offensive word.

The definitions of offensiveness level of each offensive word, w, in sentence, s, is:

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

Where $a_2 < a_1 \leq 1$.

Once a pejorative describes or targets an online user, or semantically associates with another pejorative, it becomes more offensive from users' perceptions. For example, "you stupid" and "f***ing stupid," are much more insulting than "This game is stupid." In addition, the dataset of Content Analysis for the Web 2.0 Workshop[5] shows that most of the offensive

[4] http://www.urbandictionary.com/
[5] http://caw2.barcelonamedia.org/

sentences include not only pejoratives but also user identifiers, i.e. second person pronouns, victim's screen names, and other terms referring to people. Table **4-1** lists some examples.

Table 4-1 Language features of offensive sentences

Language Features	Example
Second person pronoun (victim's screen name) + pejorative(i.e. JK, gay, wtf, emo, fag, loner, loser)	<You, gay>
Offensive adjective (i.e. stupid, foolish, sissy) + people referring terms (i.e. emo, bitch, whore, boy, girl)	<stupid, bitch> <sissy, boy>

Thus, when pejoratives grammatically relate to user identifiers or other pejoratives in sentences, the offensiveness level requires redefining. This study equipped with a nature language parser, proposed by Stanford Natural Language Processing Group[6], to capture the grammatical dependencies within a sentence. The parsing results of sentences become combinations of a dependency-type and word-pair with the form (governor, dependent). Some selected dependency types capture the possible grammatical relations between a pejorative and a user-identifier (or another offensive word) in a sentence. The study also proposes syntactical intensifier detection rules listed in Table **4-2**. A represents a user identifier, and B represents a pejorative.

Table 4-2 Syntactical intensifier detection rules

Rules	**Meanings**	**Examples**	**Dependency Types**
Descriptive Modifiers and complements: A(noun, verb, adj) ←B(adj, adv, noun)	B is used to define or modify A.	you f***ing; you who f***ing; you…the one…f***ing.	abbrev(abbreviation modifier), acomp(adjectival complement), amod(adjectival modifier), appos(appositional modifier), nn(noun compound modifier), partmod(participial modifier)
Object: B(noun, verb) ←A(noun)	A is B's direct or indirect object.	F*** yourselves; shut the f** up; f*** you idiot; you are an idiot; you say that f***...	dobj(direct object), iobj(indirect object), nsubj(nominal subject)

[6] http://nlp.stanford.edu/

Subject: A(noun)→B(noun, verb)	A is B's subject or passive subject.	you f***…; you are **ed… …f***ed by you…	nsubj(nominal subject), nsubjpass (passive nominal subject), xsubj(controlling subject), agent(passive verb's subject).
Close phrase, coordinating conjunction: A and B; …A, B…; …B, B…	A and B or two Bs are close to each other in a sentence, but be separated by comma or semicolon.	F** and stupid; you, idiot.	conj (conjunct), parataxis(from Greek for "place side by side")
Possession modifiers: A(noun)→B(noun)	A is a possessive determiner of B.	your f***; s*** falls out of your mouth.	poss(holds between the user and its possessive determiner)
Rhetorical questions: A(noun)←B(noun)	B is used to describe clause with A as root (main object).	Do you have a point, f***?	rcmod(relative clause modifier)

The offensiveness levels of pejoratives and other inappropriate words receive refinement by multiplying their prior offensiveness levels by an intensifier(Zhang et al., 2009). If in sentence, s, all words syntactically related to a pejorative, w, are categorized as set, $D_{w,s} = \{d_1, \dots, d_l\}$, the intensifier, I_w, of pejorative, w, is:

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

where $b_1 > b_2 \geq 1$.

Consequently, the offensiveness value of sentence, s, becomes a determined linear combination of words' offensiveness, $O_s = \sum p_w I_w$. The following algorithm illustrates the whole process of sentence offensiveness prediction.

Procedure SentenceOffensiveness(*s*)
TDset: dependency_type (*governor*, *dependent*)←*TDgenerator*(*s*); //Parse s to get typed dependency relations
O_s=0
FOR each word *w* in *s*
  IF it is an pejorative
    Search *TDset*
    IF (*w* equals to either *governor* or *dependent*)
      IF (its *dependency type* appears in Table **4-2**)
        IF (Its corresponding *dependent* and *governor* is a user identifier)
          $O_s = O_s + p_w * b_1$ //p_w: prior offensiveness of word *w*
        ELSE IF (Its corresponding *dependent* and *governor* is an offensive word)

$O_s = O_s + p_w * b_2$
END IF
END IF
ELSE $O_s = O_s + p_w$
END IF
END IF
END FOR
RETURN O_s

User Offensiveness Aggregation

Since the previous researches in user-level offensive analysis are lacking, we examine several studies on document level sentiment analysis (Pang et al., 2002; Tsou, 2005; Turney, 2002; Zhang et al., 2009). They predict the overall polarity of a document by aggregating polarity scores of individual sentences. Noting that sentences vary in their importance in a document, the authors also assigned different weights to the sentences to adjust their contribution to the overall polarity. Assuming users' writing styles are consistent, the offensiveness of their conversation history will reflect the potentiality for them to send out offensive content in the future. But we cannot simply sum up the offensive values of all sentences to compute users' offensiveness, because sentence offensive can be affected by the nearby ones. For example, for the post "Stupid guys need more care. You are one of them." If you calculate sentence offensiveness without considering the context, the offensiveness of the above post will not be detected even using natural language parsers. To bypass the limitation of the current parsers, we transmit the posts to new large sentences by replacing the periods with commas before feeding them to parsers. Then the parsers will generate different phrase sets to further calculate the offensiveness of modified posts. By noticing that the modified posts may sometimes miss the original meanings, we have to make a tradeoff between using the sums of sentence offensiveness to represent post offensiveness and using the offensiveness of the modified posts. In this case, we choose to use the maximum value of them to finally represent post offensiveness.

In addition, 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. Considering the following cases:

- *Punctuations and uppercase words.* Users may use punctuations and words with all uppercase letters to indicate feelings or speaking volume. Punctuation, such as exclamation marks can emphasize offensiveness of posts. (i.e. Both "You stupid!" and "You STUPID." are stronger than "You stupid.").
- *Intensive use of pejoratives.* Some users tend to post short insulting comments, such as "Holy s***. You idiot." Consequently, compared to those who post the same number of pejoratives but in longer sentences, the latter users appear more offensive for intense use of pejoratives and obscenities.
- *Behaviors in different time periods.* Social network platforms list the most recent posts at the top of users' conversation histories to attract readers' attention. To better discover whether or not a user is offensive, the offensiveness of the most recent posts needs to be checked. In addition, apparently, users may use offensive words to defend themselves when they are arguing with others who are offensive. Thus, to make sure users' offensiveness values are unaffected by offensive messages posted in short time periods is a good way to differentiate general offensive users from occasional ones.
- *Imperative sentences.* Users who frequently use imperative sentences tend to be more insulting, because imperative sentences deliver stronger sentiments. For example, a user who posts "Stupid u." can gain the perception of being a more offensive and aggressive human than the ones posting "You are stupid."
- *Cyberbullying related content.* O'Neill and Zinga(O'Neill & Zinga, 2008) described seven types of children who, due to differences from peer, may be easy targets for

online bullies. Those children may have unusual races, have religious beliefs, or just appear to have non-typical sexual orientations. Detecting online conversations related to these contents also provides clues for identifying online offensive users.

Three types of feature are largely used in authorship analysis (Hansen et al., 2007; Ma et al., 2011; Orebaugh & Allnutt, 2010; Symonenko et al., 2004; Zheng et al., 2006; Zheng et al., 2010) on cybercrime investigation: style feature, structural feature, and content-specific feature. Style feature and structural feature capture users' language patterns, while content-specific features helping to identify offensive content in users' conversations. Inspired by these ideas, we propose a new method for user-level offensive analysis as following:

Given a user, u, we retrieve his/her conversation history and chunk it into several posts, $\{p_1, \dots, p_m\}$, and for each, $p_i (i = 1, \dots, m)$, containing sentences, $\{s_1, \dots, s_n\}$. Sentence offensiveness are denoted as, $\{O_{s_1}, \dots, O_{s_n}\}$. The original offensiveness, O_p, of post p is, $O_p = \sum O_s$. The offensiveness of modified post can be presented as, $O_{p \to s}$. So the final post offensiveness $O_p{}'$ of post p can be calculated as, $O_p{}' = max(\sum O_s, O_{p \to s})$. Hence, the offensiveness value, O_u, of user ,u, can be presented as, $O_u = \frac{1}{m} \sum O_p$. We use average instead of sum to calculate user offensiveness is because users who have more posts are not necessary to be more offensive than others. O_u, should be no less than 0.

The calculated user offensiveness value is one of the features to determine users' potentiality to offend others. In addition, we also add three types of features to better classify potential offensive users: style, structural, and content-specific. The style features infer users' offensiveness from their language pattern, including whether or not they are frequently/recently using pejoratives and intensifiers such as uppercase letters and punctuations. The structural features capture the way users construct their posts. They check whether or not users are frequently using imperative sentences. They also try to infer users' writing style by check

pejoratives are often used as nouns, verbs, adjs, or advs. The content-specific features check whether or not users post suspicious content which probably will be identified as cyberbullying messages. The details of features are summarized in Table 4-3.

Table 4-3 Additional feature selection for user offensiveness analysis

Style features	Structural features	Content-specific features
-Frequency of strong/weak/general offensive words used in users' conversation -Average sentence length -Ratio of short sentences -Appearance of punctuations -Appearance of words with all uppercase letters -Appearance of offensive words in most posts over the whole time periods -Appearance of offensive words in recent posts	-Ratio of imperative sentences -Appearance of using offensive words as nouns, verbs, adjs and advs.	-Race -Religion -Violence -Sexual orientation -Clothes -Accent -Appearance -Intelligence related -Having special needs or disabilities

As the size of the feature sets become large, we use machine learning to perform the evaluation.

Chapter 5

Experiment

This section we conducted several experiments to examine LSF on offensive detection in social media.

Dataset Description

The experimental dataset, retrieved from Youtube comment boards, is a selection of text comments from postings in reaction the top 18 most popular videos. Classification of the videos includes eight categories, such as Religion, Music, and Sports. Each text comment associates with a user ID and text content. The User ID identifies the author who generates the comment, and the text content contains the user's opinion. The dataset includes comments from 2,175,474 distinct users. We randomly select a uniform distributed sample from the complete dataset, which includes 636 users. The features of sample data comparing to original data are similar as listed in Table **5**-**1**:

Table 5-1 Features of sample dataset vs. complete dataset

Average	Sample Dataset(SD)	Complete Dataset(CD)
No. of posts per user	1.91	2.80
No. of sentences per post	3.04	3.08

Pre-processing

Before feeding the dataset to the proposed classifier, an automatic pre-processing of the data assembles the comments by users and then chunks them into sentences. For each sentence in the sample dataset, an automatic spelling and grammar correction process precedes introduction

of the sample dataset to the classifier. With the help of WordNet[7] corpus and spell-correction algorithm[8], correction of spelling and grammar mistakes in the raw sentences occurs by tasks such as deleting repeat letters in words, deleting meaningless symbols, splitting long words, transposing substituted letters, and replacing the incorrect and missing letters in words. As a result, words missing letters, such as "speling," are corrected to "spelling"; misspelled words, such as "korrect," change to "correct."

Baseline Methods in Sentence Offensive Prediction

The experiment uses three learning-based approaches as baselines for detecting offensive sentences:

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 certain offensive words are used to describe users in a given sentence.

[7] WordNet, at http://wordnet.princeton.edu/
[8] Spell-Correction Algorithm, at http://norvig.com/spell-correct.html

Techniques in User Offensive Aggregation

This study adopts three machine learning classifiers implemented in Weka (Witten & Frank, 2005)(the standard machine learning software developed at the University of Waikato)—Naïve Bayes, J48, and DTNB(Decision Table/Naïve Bayes hybrid classifier (Hall & Frank, 2008))—to detect offensive users.

The Naive Bayes classifier's basis is application of Bayes' theorem with strong (naive) independence assumptions. Despite the fact that the independence assumptions are often inaccurate, the naive Bayes classifier has several properties that allow being trained efficiently in supervised learning settings. Besides, it only requires a small amount of training data to estimate the parameters necessary for classification.

J48 implemented C4.5 algorithm—a decision-tree generating algorithm developed by Quinlan (Quinlan, 1986). It adopts a divide-and-conquer strategy to generate classification results in accurate and a higher rate.

Evaluation Metrics

In this study's experiments, standard evaluation metrics for classification (Pang et al., 2002; Turney, 2002; Ye et al., 2006) (i.e., accuracy, precision, recall, and F-measures) are used to evaluate the performance of LSF. In particular, accuracy (recall) measures the overall classification correctness, which represents the percent of actually real offensive messages posts that are correctly identified. The false positive (FP) rate represents the percent of identified posts that are not truly offensive messages. The false negative (FN) rate represents the percent of actually real offensive messages posts that are unidentified. Precision presents the percent of

identified posts that are truly offensive messages, and f-score (Yin, Xue, Hong, & Davison, 2009) represents the weighted harmonic mean of precision and recall, which is defined as:

$$f - score = \frac{2(precision \times recall)}{precision + recall}$$

Experimental Results

As explained in above, the offensive word lexicon in this research consists of the lexicon used in study (Xu & Zhu, 2010) and a lexicon, based on Urban Dictionary, established during the coding process. Pejoratives receive the label of strongly offensive word if more than 80% of its use is offensive ways. Otherwise, known pejoratives receive the label of weakly offensive word. In total, we select 538 strongly offensive words and 181 weakly offensive words when feeding the sample dataset to LSF classifier. We define "1" to be the threshold for offensive sentence classification, that is, sentences with offensiveness values more than (inclusive) "1" receive labels of offensive sentences, because by our definition, offensive sentence means a sentence containing strongly offensive words, or containing weakly offensive words used to describe another user. The experimental parameters are set as: $a_1 = 1; a_2 = 0.5; b_1 = 2; b_2 = 1.5$.

Accuracy

Subsequently, a manual check on the classifier's output produced the results as shown in Table **5**-**2**.

Table 5-2 Accuracies of sentence level offensiveness detection

Relations	Accuracy	FP rate	FN rate	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%
LSF	94.34%	1.76%	5.66%	98.24%	96.25%

significant at $\alpha=0.05$.

According to Table **5-2**, the bag-of-words approach can identify the 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. Once N equals to the length of sentences, N-gram is equivalent to the bag-of-words approach. To further apply N-gram in the classification, application of different values of N are necessary to balance, perfectly, the trade-off between accuracy and false positive rate.

Moreover, none of the baseline approaches provides false negative rates less than 33%, because many of the obviously offensive sentences are imperatives, which omit all user identifiers. However, simply using an offensive word as the only detection feature produces an even higher false positive rate. LSF obtains its highest F-score because it sufficiently balances the precision-accuracy tradeoff. Unfortunately, the parser sometimes misidentifies noun appositions, in part because of typographical errors in the input, such as: “[T]here are many people deserving of you stupid sympathies also.” Here, the sender presumably meant to write “your” instead of “you.” This is the major reason for false negative rates. The false positive rate arises mainly from multiple appearances of weakly offensive words, for example, “fake and stupid,” which can only represent a negative opinion for a video clip.

When titles of some movies and songs contain offensive words, such as the song “Jizz in My Pants” on the Youtube, a conclusion as to whether or not a user is offensive will be difficult to reach without sufficient contextual information, because users can be just discussing online materials. In this case, a false positive may occur. However, based on the assessment that someone who prefers online content with offensive language is generally more likely to be offensive, the proposed classifier still provides accurate classification results.

Speed

In addition to accuracy measurement, assessment of processing speed on masses of text messages is necessary, because speed is a critical attribute for offensive detection in real-time online communities. The processing time in each case appear in Figure 5-1.

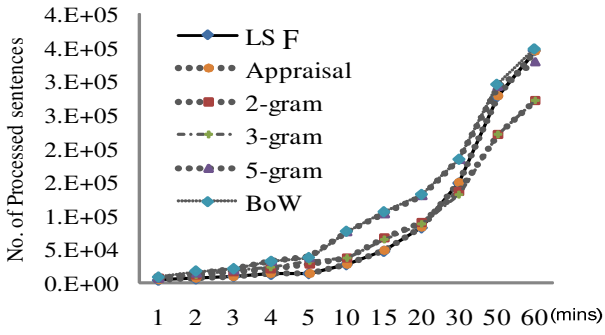

Figure 5-1 Data Processing Time

From Figure 5-1, the processing rate of the proposed LSF is at least equal to other approaches. Thus, the proposed method is practical for application to OSNs and other real-time online communities.

User Level Aggregation Evaluation

Offensiveness values of users are calculated from synthesizing the offensiveness values of all sentences. After excluding the obviously non-offensive users (whose offensiveness values are 0), this study randomly selects 249 users with uniformly distributed offensiveness values from the dataset to represent an experimental sample. (The selected users have 32 posts on average.) Then, three experts, with Cronbach's α value 0.73 (agreement rate), manually label the users in the sample. A valid user label is generated when all of the experts put the same label on that user. As a result, 128 users receive labels of "offensive," and 78 users receive labels of "non-offensive." The labels and selected feature sets are fed into the three machine learning classifiers. The results (f-scores) appear in Table 5-3. F1, F2 and F3 represent style features, structural features and content-specific features, respectively.

Table 5-3 F-score for different feature sets using NaiveBayes, J48 and DTNB classifiers

Feature Sets	NaïveBayes	J48	DTNB
O-value[13]	0.781	0.805	0.785
O-value+F1	0.832	0.823	0.876
O-value +F1+F2	0.840	0.839	0.864
O-value +F1+F2+F3	0.840	0.843	**0.882**

According to Table 5-3, the performance of the three machine learning classifiers is better when adding new feature sets. The improvement indicates the additional feature sets are all useful for helping to detect offensive online users. The accuracy of detecting offensive users by offensiveness values is only 80.5% (Table **5**-**4**). However, after feeding new feature sets into the classifiers, the accuracy increases to 88.2%, indicating a significant improvement.

Table 5-4 Classification result

Techniques	Accuracy	FP	Precision	Recall	F-score
O-Value	0.789	0.277	0.82	0.789	0.805
NaiveBayes	0.789	0.145	0.897	0.789	0.840
J48	0.827	0.217	0.859	0.827	0.843
DTNB	0.902	0.229	0.863	0.902	**0.882**

Discussion

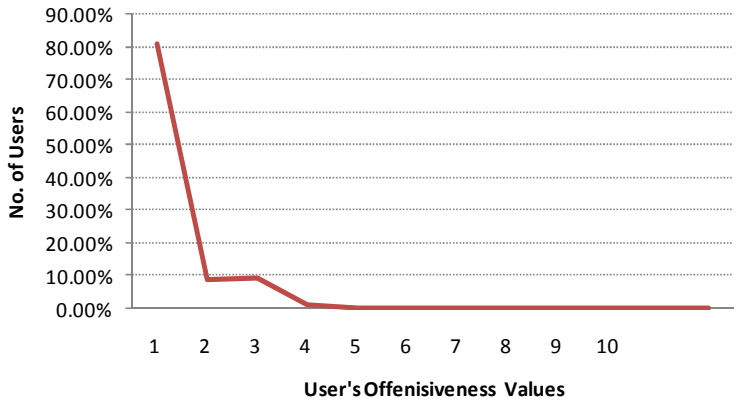

Figure 5-2 Users' offensiveness distribution on Youtube

When we check users' offensiveness distribution on Youtube(Figure **5-2**), we find that 81% of Youtube users do not use any offensive words in their posts, and the portion of users with offensiveness value over four is less than 1%. It indicates users' offensiveness values on the Youtube website satisfy the power law, and the number of users and their offensiveness values negatively correlate.

Chapter 6

Limitations and Future Work

This research still has some limitations associated with the research's lexical semantic approach.

First, the language used in online communities is casual and in an informal English writing style, compared to the language used for journals and newspapers. When applying NLP techniques, errors in texts create inaccurate assessments, as mentioned in Chapter 5. However, since most text messages in online communities have very simple and neat grammatical structures, highly accurate textual analysis should be relatively easy. Thus, if the parser has difficulty analyzing the comment, it will probably cause difficulty to human readers as well.

Secondly, social networking systems or online communities normally generate new words every single day. Since the study's sentence offensiveness prediction heavily relies on an offensive word lexicon, the lexicon requires constant updating; therefore, future effort will focus on implementing machine learning techniques or using users' feedback for addition of new offensive language to the lexicon.

Moreover, in user offensiveness aggregation, currently, consideration only includes the effect of exclamation marks on a sentence's sentiment, but other punctuation such as quotation marks can change (dilute) a sentence's sentiment. Future research will consider assigning different weights to different user-level intensifiers.

Chapter 7

Conclusion

In this study, we investigate existing text-mining methods in offensive content detection for adolescence protection. Then we propose the Lexical Syntactical Feature (LSF) to identify offensive content in social media, and further predict users' potentiality to send out offensive content. Our research has several contributions.

Conceptually, we give a practical definition on online offensive content, and further distinguish the contribution of pejoratives/ profanities and obscenities in determining offensive content, and introduce hand-authoring syntactic rules in identifying name-calling harassment.

On the technical side, we revised the traditional machine learning which only using lexical, sentiment, and contextual features in offensiveness detection, and also incorporated style features (capture users' writing style), structure features (users' writing structure) and context-specific features (cyberbullying related content) to better predict users' potentiality to send out offensive content in online social media.

Experimental result shows that the LSF sentence offensive prediction algorithm outperforms traditional learning-based approach in terms of precision, recall and f-score. It achieves high processing speed for effective deployment on social media too. Besides, LSF tolerates informal and misspelling content, so it can easily adapt to any formats of English writing styles. We believe that such language processing model will greatly help to online offensive language monitoring, and eventually to build a better online environment.

In the future we will test our LSF framework with other language and compare performance with different languages. We also plan to apply the framework to analyze user-generated data in other social media.

Reference

Bruce, R. F., & Wiebe, J. M. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering, 5*(2), 187-205.

Cheng, J. (2007). Report: 80 percent of blogs contain "offensive" content. *ars technica* Retrieved 11/7, 2011, from http://arstechnica.com/security/news/2007/04/report-80-percent-of-blogs-contain-offensive-content.ars

D Riffe, S. L., F Fico (1998). Analyzing Media Messages: Using Quantitative Content Analysis in Research. *Lawrence Erlbaum Associates*.

Dave, K., Lawrence,S. , Pennock, D. M. . (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proc. of the 12th International Conference on the World Wide Web*, 519-528.

Dellarocas, C. (2000). *Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior.* Paper presented at the Proceedings of the 2nd ACM conference on Electronic commerce EC '00, New York.

Diana, A. (2010). 75% Of U.S. Households Use Social Networking. *InformationWeek.* Retrieved 11/7, 2011, from http://www.informationweek.com/news/225700459

Finkelhor, D., Hargittai, E., Hinduja, S., Lenhart, A., Mitchell, K., Patchin, J., Rosen, L., Wolak, J., & Ybarra, M. (2008). Online Threats to Youth. *Literature Review Prepared for the Internet Safety Technical Task Force*.

Gordon, A., Kazemzadeh,A., Nair,A., Petrova,M. . (2003). Recognizing Expressions of Commonsense Psychology in English Text. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, & MEDIA, C. O. C. A. (2011). Clinical Report--The Impact of Social Media on Children, Adolescents, and Families. *Pediatrics*.

Hall, M., & Frank, E. (2008). *Combining naive Bayes and decision tables.* Paper presented at the Proceedings of the Twenty-First International FLAIRS Conference.

Hansen, J. V., Lowry, P. B., Meservy, R. D., & McDonald, D. M. (2007). Genetic programming for prevention of cyberterrorism through dynamic and evolving intrusion detection. *Decision Support Systems, 43*(4), 1362-1374.

Hinduja, S., & Patchin, J. W. (2008). Bullying beyond the schoolyard: Preventing and responding to cyberbullying. In (pp. 13): Corwin Pr.

Houser, D., & Wooders, J. (2006). Reputation in auctions: Theory, and evidence from eBay. *Journal of Economics & Management Strategy, 15*(2), 353-369.

Ito, M., Horst, H., Bittanti, M., Boyd, D., Herr-Stephenson, B., Lange, P. G., Pascoe, C., & Robinson, L. (2008). Living and learning with new media: Summary of findings from the Digital Youth Project. *The John D. and Catherine T. MacArthur Foundation Reports on Digital Media and Learning*.

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture, 4*(2), 267-288.

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. *Text Mining*, 149-164.

Kontostathis, A., & Leatherman, L. E. A. (2009). Chatcoder: Toward the tracking and categorization of internet predators. *In Proc. Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining (SDM 2009).*

Leatherman, A. (2009). *Luring language and virtual victims: Coding cyber-predators online communicative behavior*.

Lee, E., & Leets, L. (2002). Persuasive storytelling by hate groups online. *American Behavioral Scientist, 45*(6), 927.

Lenhart, A., Madden, M., Smith, A., & Macgill, A. (2007). Teens and social media: An overview. *Pew Internet and American Life Project). Washington, DC*.
Ma, J., Teng, G., Chang, S., Zhang, X., & Xiao, K. (2011). Social Network Analysis Based on Authorship Identification for Cybercrime Investigation. *Intelligence and Security Informatics*, 27-35.
Mahmud, A., Ahmed, Kazi Zubair, and Khan, Mumit (2008). *Detecting flames and insults in text.* Paper presented at the Proc. of 6th International Conference on Natural Language Processing (ICON' 08).
McEnery, A., Baker, J., & Hardie, A. (2000a). Swearing and abuse in modern British English.
McEnery, A., Baker, J. P., & Hardie, A. (2000b). Assessing claims about language use with corpus data–swearing and abuse. *Corpora galore, 30*, 45.
McEnery, A., & Xiao, Z. (2004). Swearing in modern British English: the case of fuck in the BNC. *Language and Literature, 13*(3), 235.
McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present* (Vol. 1): Psychology Press.
McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., & Jakubowski, E. (2011). Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce, 15*(3), 103-122.
O'Neill, T., & Zinga, D. (2008). *Children's rights: multidisciplinary approaches to participation and protection*: Univ of Toronto Pr.
Olson, L., Daggs, J., Ellevold, B., & Rogers, T. (2007). Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory 17(3)*, 231–251.
Orebaugh, A., & Allnutt, D. J. (2010). Data Mining Instant Messaging Communications to Perform Author Identification for Cybercrime Investigations. *Digital Forensics and Cyber Crime*, 99-110.
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *In EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 79-86.
Pazienza, M., & Tudorache, A. (2011). Interdisciplinary contributions to flame modeling. *AI* IA 2011: Artificial Intelligence Around Man and Beyond*, 213-224.
Pazienzaa, M. T., Stellatoa, A., & Tudoracheab, A. (2008). Flames, Risky Discussions, No Flames Recognition in Forums: Citeseer.
Pendar, N. (2007). *Toward Spotting the Pedophile Telling victim from predator in text chats.* Paper presented at the Proceedings of the First IEEE International Conference on Semantic Computing.
Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106.
Razavi, A., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive Language Detection Using Multi-level Classification. *Advances in Artificial Intelligence, 6085/2010*, 16-27.
Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM, 43*(12), 45-48.
Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *Advances in Applied Microeconomics: A Research Annual, 11*, 127-157.
Ries, T. (2011). 65% of Online Americans Use Social Networking Sites; Young Adult Women Are The Power Users. *The Realtime Report* Retrieved 11/7, 2011, from http://therealtimereport.com/2011/08/29/65-of-online-americans-use-social-networking-sites-young-adult-women-are-the-power-users/
Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *EMNLP*.
Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning Subjective Nouns using Extraction Pattern Bootstrapping. *Proceeings of the Seventh CoNLL conference*.
Sjöbergh, J., & Araki, K. (2008). *A multi-lingual dictionary of dirty words.* Paper presented at the LREC.
Smith, A. (2007). Teens and Online Stranger Contact. *Pew Internet & American Life Project*.
Spertus, E. (1997). Smokey: Automatic Recognition of Hostile Messages. *Innovative Applications of Artificial Intelligence (IAAI) '97*.
Symonenko, S., Liddy, E. D., Yilmazel, O., Del Zoppo, R., Brown, E., & Downey, M. (2004). Semantic analysis for monitoring insider threats. *Intelligence and Security Informatics*, 492-500.
Thom, B., Kontostathis, A., & Edwards, L. (2011). *SafeChat: Using Open Source Software to Protect Minors from Internet Predation*. Paper presented at the Proceedings of the ACM WebSci'11.

Timothy Johnson, Robert Shapiro, & Tourangeau, R. (2011). National Survey of American Attitudes on Substance Abuse XVI: Teens and Parents. *The National Center on Addiction and Substance Abuse.* Retrieved 11/7, 2011, from http://www.casacolumbia.org/templates/NewsRoom.aspx?articleid=648&zoneid=51

Tsou, B. K. Y., Yuen,R.W.M.,Kwong, O.Y., Lai,T.B.Y.,&Wong,W.L. (2005). Polarity classification of celebrity coverage in the Chinese press. *Paper presented at the International Conference on Intelligence Analysis*.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the Association for Computational Linguistics (ACL)*, 417-424.

Turney, P., & Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*.

Tynes, B., Reynolds, L., & Greenfield, P. M. (2004). Adolescence, race, and ethnicity on the Internet: A comparison of discourse in monitored vs. unmonitored chat rooms. *Journal of Applied Developmental Psychology, 25*(6), 667-684.

Wiebe, J., Bruce, R., Bell, M., Martin, M., & Wilson, T. (2001). *A Corpus Study of Evaluative and Speculative Language.* Paper presented at the Proceedings of 2nd ACL SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). *Development and use of a gold-standard data set for subjectivity classifications*.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann Pub.

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS), 26*(3), 1-37.

Xu, Z., & Zhu, S. (2010). Filtering Offensive Language in Online Communities using Grammatical Relations. *CEAS, Collaboration, Electronic messaging, AntiAbuse and Spam Conference*.

Ye, Q., Shi, W., & Li, Y. (2006). *Sentiment classification for movie reviews in Chinese by improved semantic oriented approach.* Paper presented at the HICSS '06. Proceedings of the 39th Annual Hawaii international Conference on System Sciences.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. *Third IEEE International Conference on Data Mining (ICDM'03)*.

Yin, D., Xue, Z., Hong, L., & Davison, B. (2009). *Kontostathis A and Edwards L 2009 Detection of harassment on Web 2.0.* Paper presented at the the Content Analysis in the Web 2.0 (CAW2.0) Workshop.

Yin, D., Xue, Z., Hong, L., Davison, B., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on Web 2.0. *Proceedings of the Content Analysis in the Web 2.0 (CAW2.0) Workshop at WWW2009*.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *Paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zacharia, G., Moukas, A., & Maes, P. (2000). Collaborative reputation mechanisms for electronic marketplaces. *Decision Support Systems, 29*(4), 371-388.

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

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 57*(3), 378-393.

Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2010). Authorship analysis in cybercrime investigation. *Intelligence and Security Informatics*, 959-959.