Smokey: Automatic Recognition of Hostile Messages

Ellen Spertus
Microsoft Research, MIT AI Lab, and University of Washington
Dept. of Computer Science and Engineering
Box 352350
Seattle, WA 98195

ABSTRACT

Abusive messages (flames) can be both a source of frustration and a waste of time for Internet users. This paper describes some approaches to flame recognition, including a prototype system, Smokey. Smokey builds a 47-element feature vector based on the syntax and semantics of each sentence, combining the vectors for the sentences within each message. A training set of 720 messages was used by Quinlan's C4.5 decision-tree generator to determine feature-based rules that were able to correctly categorize 64% of the flames and 98% of the non-flames in a separate test set of 460 messages. Additional techniques for greater accuracy and user customization are also discussed.

Introduction

Flames are one of the current hazards of on-line communication. While some people enjoy exchanging flames, most users consider these abusive and insulting messages to be a nuisance or even upsetting. I describe Smokey, a prototype system to automatically recognize email flames. Smokey combines natural-language processing and sociolinguistic observations to identify messages that not only contain insulting words but use them in an insulting manner. Additional methods, not implemented in Smokey, are also outlined.

There are many different types of flames, occurring in real-time communication, in discussion groups (such as Usenet newsgroups and mailing lists), and in private messages, such as email. Publicly-posted flames tend to be more clever and indirect than private email flames, making them harder to reliably detect. As a first step in this field, Smokey addresses private messages; specifically, comments that are sent via feedback forms on World-Wide Web pages. In order to have enough flames to study, I obtained messages from the webmasters of controversial pages, specifically: NewtWatch (Dorsey and Schnackertz 1997), which criticizes Newt Gingrich; The Right Side of the Web (Donnels 1997), a conservative resource; and Fairness and Accuracy in Reporting (FAIR) (Ernst 1997), a media watch group best known for its criticisms of the veracity of Rush Limbaugh's claims.

The obvious method of identifying flames—looking for obscene expressions—does not work well. Only 12% of the flames contained vulgarities, and over a third of the vulgar messages were not flames. Also, some messages with profanity directed it at someone that both the sender and recipient dislike. For example, the statement "Newt Gingrich is an a------" is a flame if sent to The Right Side of the Web but not if sent to NewtWatch, as, in fact, it was. (All quoted examples, and the typos in them, are genuine. The only change I made is replacing letters of obscene words with dashes.) Smokey looks not only for insulting words and the context in which they are used but also for syntactic constructs that tend to be insulting or condescending, such as imperative statements. Smokey avoids misclassifying friendly messages by looking for praise, requests, and polite speech.

System Architecture

Smokey consists of 5 phases:

1. Messages are converted into a common format with one sentence per line and delimiters between messages. This is done in Emacs Lisp.
2. The text is run through a parser developed by the Microsoft Research Natural Language Processing Group.
3. The output of the parser is converted by sed and awk scripts into Lisp s-expressions.
4. The s-expressions are processed through rules written in Emacs Lisp, producing a feature vector for each message. This phase can also generate output showing where each rule applied.
5. The resulting feature vectors are evaluated with simple rules, produced by the decision tree generator C4.5 (Quinlan 1993).

I removed duplicate, excessively-long, and meaningless messages (someone randomly pressing keys) from the collections; this could be automated. Steps 1 and 3 of Smokey are trivial and will not be further discussed in this paper. The parser, used in step 2, is discussed elsewhere (Richardson 1994) and, as appropriate, in this paper. The s-expressions used by the fourth step to represent each sentence include (1) an ordinary string, (2) a list of words in the sentence, and (3) a tree encoding the grammatical structure of each sentence.

Each rule in step 4 is a Lisp procedure that takes an s-expression representing a sentence as an argument and returns 1 if the rule applies, 0 otherwise. (The only exception was the rule that returns a count of exclamation points.) Mutually exclusive rules are grouped into classes for greater accuracy and efficiency. Each class has a guard procedure which checks whether a rule in the class could apply. If so, the rules are attempted in order until

one succeeds. All regular-expression matching was case-insensitive. Table 1 shows each of the rules, the number of times it is met in the sample data, and the probability that if a sentence satisfies the rule that it is part of a message classified as a flame, maybe, or okay. The reader will want to refer to Table 1 while reading the next section, which describes the rule classes. A feature vector was created for each message by summing the vectors of each sentence.

Most of the rules behave the same regardless of where the data comes from. There are a few variables that get set with site-specific information. For example, the string "Slick Willy" is probably insulting in a message to NewtWatch but not in one to The Right Side of the Web. Site-specific variables will be discussed with the rules that use them.

Rule classes

Noun Phrases used as Appositions

I found that phrases with "you" modified by a noun phrase tend to be insulting. (The technical term is a noun apposition.) Examples are "you bozos", "you flamers", and "you people". Exceptions are "you guys" and sometimes "you folks". Table 1 shows that "you guys" appeared 38 times in the examined messages and that in the sentences it appeared, 66% were in messages classified as okay, 13% as maybe, and 21% as flames. "You folks" was less likely to be part of a flame (13%), and the general case was more likely (53%).

Because the parser marks noun appositions in the grammatical tree of each sentence, the Lisp rule to recognize them is trivial. Unfortunately, the parser sometimes misidentifies noun appositions, in part because of typographical errors in the input, such as: "[T]here are many other fine members of congress deserving of you gentile sympathies also." Here, the sender presumably meant to write "your" instead of "you".

Imperative Statements

Another heuristic based on the syntax of a sentence is that imperative statements (commands) tend to be insulting. Some examples are:

"Have your fun" *"forget about it"*
"Get used to it!" *"Get over it!"*
"Get Lost!!!" *"get a life"*
"f--- you" (but see Gregersen (1977) and Quang (1992))

The guard for the class checks whether the parser marked the statement as being imperative. There are several varieties of imperative statements that are not insulting, including idiomatic expressions, such as "keep up the good work" and "have a nice day"; suggestions and invitations; and sentences that only appear to be imperative because the writer omitted "I", such as "Love your work".

Long imperative statements or those with multiple clauses are less likely to be insulting. Consider: "If you have a candidate, pledge your loyalty to only one, and don't make a mistake and lose yourself in congress." All of these conditions are checked for by rules in the imperative class, as illustrated in Table 1.

A source of miscategorizations is ambiguous statements such as the following:

"Cool page...."
"Just saw our link."

The parser identified these as imperative statements, a reasonable—but amusing—interpretation. "Cool page" was marked as imperative because it can be interpreted as a verb followed by a noun, presumably meaning that the listener should take a page and cool it off. "Just saw our link" was misinterpreted as a command to saw (i.e., with a handsaw) a link. Semantic analysis wouldn't necessarily help, because sawing a link (i.e., of a chain) makes sense. Of course, neither sentence was meant as imperative, so the rule misfires, contributing to its lower-than-expected flame-prediction rate.

Second-Person Rules

Many sentences with a word beginning with "you" (including "your" and "yourself") are insulting; specifically, sentences with "your ilk", "your so-called", and *scare quotes*, such as: "This 'service' of yours reminds me of when I was in college and kids wrote similar comments on the bathroom walls about Reagan."

Profanity Rules

This class is entered when a sentence contains an obscene word. The list did not include "damn" and "hell", since they are used so frequently. A distinction is made depending on whether the sentence also contains the name of a site-specific "villain". For example, for NewtWatch, villains are "Newt", "Gingrich", "Rush", "Limbaugh", and "Helms", so the sentence "Newt Gingrich is an a------" would fall in this category. The names of web browsers, such as lynx, are considered honorary villains, so this rule catches: "Lynx currently s--ts out...the first time you try to page down".

Condescension Rules

Condescending statements recognized through regular expressions were divided into three classes: very, somewhat, and slightly condescending, and are described in Table 1. The only structural rule used for condescending statements is to mark a class of "tag phrases", two-word phrases consisting of a contraction followed by another word and a question mark, such as "It really is a helpless feeling when your side is solidly in the minority, isn't it?" The regular expression for such a tag phrase is: "[,$] ?[a-zA-Z]+'t [a-zA-Z]+\?".

#	Rule	Example	n	ok	maybe	flame
1	"You" followed by "guys"	"I'll be giving you guys a call soon."	38	66%	13%	**21%**
2	"You" followed by "folks"	"Well, you folks have fun poking fun at Newt."	8	63%	25%	**13%**
3	"You" followed by any other noun phrase	"You quivering, socialist, bedwetters must be scared to death to go to all this trouble bashing Newt."	30	33%	13%	**53%**
4	Imperative sentence (Imp.) with "have...day"	"Have a nice day."	2	**100%**	0%	0%
5	Imp. with "keep...work" or "keep...up"	"Keep up the good work."	112	**97%**	2%	1%
6	Imp. containing "look"	"Look forward to hearing from you."	8	**88%**	13%	0%
7	Imp. containing "take"	"Take care."	5	60%	0%	**40%**
8	*Imp. containing "let"*	*"Let's not dilly-dally."*	*31*	***87%***	*13%*	*0%*
9	*Imp. containing "thank"*	*"Thank you."*	*66*	***95%***	*5%*	*0%*
10	Imp. containing "please"	"Please don't judge us all by our 6th district."	47	**85%**	13%	2%
11	Imp. containing "love" or "like"	"Love the artwork!"	9	**100%**	0%	0%
12	Imp. with comma or semicolon or more than 12 words	"If interested, hit my page."	63	62%	21%	**17%**
13	Imp. statement not meeting any of the above rules	"Get used to it!"	159	69%	17%	**14%**
14	Sentence with a word beginning with "you" and with "ilk"	"Your ilk is primarily responsible for most of the ills of this country."	3	0%	67%	**33%**
15	Sentence with "your so called" or "your so-called"	"Read through your so called evidence"	4	25%	25%	**50%**
16	Sentence with "as" or "like" followed by "yourself" or "yourselves"	"Newt is a god untaintable by such pusilanimous vultures as yourselves."	3	0%	33%	**67%**
17	Quoted phrase preceded by "you" or followed by "of yours"	"Why don't you 'fact check' the news media"	2	50%	50%	0%
18	Sentence with obscene word and site-specific villain or browser name	"Newt Gingrich is an a------"	4	**100%**	0%	0%
19	Sentence with obscene word not meeting above rule	"What the f--- is your problem?"	17	29%	12%	**59%**
20	Containing "you...miffed", "we/you...musn't/mustn't", "chicken[?!]" or "what's the matter"	"If you are miffed about this election, just wait till the next."	3	0%	0%	**100%**
21	Containing "your right", "you have a right", or "bash"	"While I respect your right to express opinions I feel that most of what is posted here is..."	28	32%	25%	**43%**
22	Containing "you have got to be"/"you've got to be"/"you must be"	"You have got to be joking!!!"	1	0%	100%	0%
23	Sentence with tag phrase	"It really is a helpless feeling when your side is solidly in the minority, isn't it?"	2	50%	50%	0%

Table 1, part 1: Column n indicates the number of times the construct occurred in the 1222 messages. The okay, maybe, and flame columns indicate the probability that a sentence meeting the rule was in a message of that class; for all messages, the probabilities are 83%, 11%, and 7%, respectively. Bold face is used to indicate probabilities higher than these baselines. Italicized rules were useful predictors of okay messages. *Table 1 continues on the next page.*

#	Rule	Example	n	ok	maybe	flame
24	Contains a negative word near a term for the site	"You will regret that you had anything to do with this crappy home page."	9	22%	33%	**44%**
25	Contains a negative word near "you"	"You Sick idiotic liberals!"	19	16%	11%	**74%**
26	Contains a negative word near "this" used as a pronoun.	"What kind of crap is this?"	3	33%	33%	**33%**
27	Contains a negative word and a site-specific villain	"All the criticism of Newt ... here is quite idiotic"	40	70%	13%	**18%**
28	Contains a negative word but does not meet any of the above rules	"Pardon my lack of tact, but this is the most pathetic thing I believe I have ever seen."	128	52%	27%	**20%**
29	Contains a site-specific insulting phrase, such as "Slick Willy" or "socialis" to liberals.	"GET THE SOCIALISTS OUT OF MY POCKET !"	49	29%	29%	**43%**
30	Insulting epithet, such as "get a life", anywhere in sentence	"maybe you should get A life"	33	21%	39%	**39%**
31	*Sentence with no obscene words (SNOW) that contains "thanks"*	*"Thanks for this service."*	*266*	***92%***	*5%*	*2%*
32	SNOW with "please"	"Please expose this hypocracy..."	128	**86%**	10%	4%
33	*SNOW with "would you"/"I would"/"I'd"*	*"Would you be willing to email me your logo..."*	*157*	***90%***	*8%*	*1%*
34	SNOW with "bless" or "godspeed"	"Godspeed, Good Luck and Stay True!"	5	**100%**	0%	0%
35	SNOW with "congra*" or "kudos"	"This is a very useful page, congrads !"	12	**100%**	0%	0%
36	*SNOW with a positive adjective near a site synonym*	*"You have got a great web site here !"*	*157*	***94%***	*5%*	*1%*
37	*SNOW with a positive verb near a site synonym*	*"...we really enjoy the NEWTWATCH"*	*36*	***92%***	*8%*	*0%*
38	SNOW with "you" near a positive adjective.	"You have a very good thing going, keep it up."	34	**88%**	3%	**9%**
39	SNOW with "I" near a good verb, a good adjective at the beginning of a sentence or at the end of a short sentence	"I was delighted to find 'The Right Side!'" "Great to see a conservative page on the net..."	220	**88%**	9%	3%
40	SNOW with "add" near "link" or "pointer" or with "shall"/ "will" / "recommend" near a site synonym	"I am adding links to your homepage from mine." "I shall use your page as a guide..."	13	**92%**	8%	0%
41	*SNOW containing "link" and not meeting any of the above rules*	*"I invite you to establish links to my 'newt bites - And Children Go Hungry' sticker page..."*	*82*	***99%***	*1%*	*0%*
42	Contains a smiley face, such as ":-)" or ":)"	"I see where you are coming from. Liberal losers who can't get over the loss in '94 ;)"	18	83%	6%	**11%**
43	*Contains a telephone number*	*"I am told that 1-800-768-2221 connects directly to the Congressional switchboard free of charge."*	*24*	***92%***	*4%*	4%
44	*Contains a uniform resource locator (URL), i.e., a web*	*"You can check it out at http://www.sfgate.com"*	*127*	***98%***	*1%*	1%
45	Contains "I" near "help" or "give"	"...if I can help, let me know."	12	83%	17%	0%
46	Contains laughter	"Ha ha ha ha ha ha ha ha ha! Your page is a joke!"	6	50%	17%	**33%**
47	Contains exclamation points	"neWt gInGriCh iS evil!!"	460	81%	10%	**9%**

Table 1, part 2: Column n indicates the number of times the construct occurred in the 1222 messages. The okay, maybe, and flame columns indicate the probability that a sentence meeting the rule was in a message of that class; for all messages, the probabilities are 83%, 11%, and 7%, respectively. Bold face is used to indicate probabilities higher than these baselines. Italicized rules were useful predictors of okay messages.

Insults

The rule class *Insults* is guarded by a check for *bad-words*, which consists of *bad-verbs* ("stink", "suck", etc.), *bad-adjs* ("bad", "lousy", etc.), and *bad-nouns* ("loser", "idiot", etc.). Rules check whether the bad word appears near a name for the page, as in "Your page is a JOKE!", or near the word "you", as in "You Sick idiotic liberals!"

A rule checks for a bad word near "this" used as a pronoun (in place of a noun), such as "What kind of crap is this?!?" Sentences where "this" is used as an adjective (to modify a noun) are not counted, such as "Not only is this country in a bad state..." A separate rule checks for insults containing a site-specific villain.

A separate class of insults is site-specific phrases. These include names ("Watergate" is only mentioned in flames to The Right Side of the Web), derogatory nicknames ("Slick Willy"), and terms that are primarily used when insulting a specific group (e.g., calling liberals "commies" or conservatives "fascists") (Hayakawa and Hayakawa 1990).

Epithets

The final class of insulting statements is epithets, short insulting phrases (Allan and Burridge 1991, Jochnowitz 1987). For example, "get" followed within ten characters by "life", "lost", "real", "clue", "with it", or "used to it". Other epithets are two-word phrases, such as "drop dead". While some of these are caught by the check for imperative statements or vulgarities, others required the epithet rule, such as: "MAYBE YOU SHOULD GET A LIFE AND QUIT TRYING TO USE RUSH AS YOUR WAY TO STARDOM". This rule proved to be one of the most useful: 18% of the flames included epithets, and none of the non-flames included them.

Polite Rules

The politeness rule class is entered if a sentence does not contain an obscene word. A message is considered polite (Brown and Levinson 1987) if it contains "thank" (but not "no thanks"), "please", or constructs with "would", such as "Would you be willing to e-mail me your logo".

Praise Rules

The praise class is entered if a sentence does not contain an obscene word. The simplest rules are that a sentence is considered praise if it contains such word stems as "bless", "godspeed", "kudos", or "congra" (for "congratulations" and related misspellings).

Other rules require vocabulary information. I predefine regular expressions *web-nouns* ("page", "site", etc.), *good-adjectives* ("great", "super", etc.), and *good-verbs* ("enjoy", "agree", etc.). Each site also has a regular expression *page-name* representing the name of the page and common synonyms (such as "NewtWatch" and "Newt Watch"). Rules check whether one of the positive terms occurs near the word "you" or a synonym for a page name (such as "This is my favorite political page"). If the word "like" appears, the rule checks that it is being used as a verb. The sentence "Like your pages" qualifies but not "...cool progressive resources (like Newt Watch)", where "like" is used as a conjunction.

A message may indirectly offer praise in a sentence with the word "I" before a positive verb (such as "i just found your web pages and I love it.") or with a positive adjective at the end of a clause near the beginning of a sentence ("Very interesting!"). Another way to offer indirect praise is to write that one will "add" or "recommend" a "link" or page. Even mentioning the word "link" in a message means it is almost certainly friendly.

Miscellaneous

The remaining classes of rules have a single rule each (i.e., the guard was the rule) and check for smiley faces, phone numbers, uniform resource locators (web addresses), offers, laughter, and exclamation points. All are binary, except for exclamation points, for which a count is returned.

Method

Human Message Ratings

Each of the 1222 messages was rated by four speakers of American English employed by Microsoft who were not otherwise involved in this research. Each message was rated by two men and two women, because gender differences in online (Herring 1995) and offline (Jay 1992) flaming have been observed. No individual rated messages from more than one site, because it was important to remember the intended recipient's political orientation. Volunteers were told to mark a message as being a flame if it contained insulting or abusive language (unless it was directed to someone the sender and recipient both disliked), not merely if the sender expressed disagreement with the recipient. Volunteers could classify a message with "flame", "okay", or "maybe". In 80% of the cases, all four volunteers agreed. In an additional 13% of the cases, exactly three volunteers agreed. I combined the ratings by classifying a message as a flame if at least three individuals considered it one (7.5% of the messages), okay if at least 3 people judged it okay and nobody considered it a flame (80%), and maybe otherwise (13%).

Message classification

The decision tree generator C4.5 (Quinlan 1993) was used with the MLC++ utilities (Kohavi et al 1994) to generate a classifier. Because decision tree generators perform badly when one classification is much more common than the others, it was necessary to weed out messages that were obviously okay to lessen the imbalance. This was done by observing that some features almost always indicated that

a message was okay. For example, only 1 of the 70 messages in the training set with "keep up the good work" was a flame. By making the approximation that messages triggering any of 10 such rules (italicized in Table 1) were okay, the ratio of okay to flame in the remaining messages could be reduced from 10:1 to 4.5:1, as shown in Table 2, allowing effective decision tree generation. The other way we overcame the generator's bias toward the common case was by interpreting its classifications of maybe as flame.

	okay	*maybe*	*flame*	*okay:flame*
original	574	88	58	9.9:1
removed	359	28	10	35.9:1
remaining	215	60	48	4.5:1

Table 2: Messages of each type and okay:flame ratio in original training set, removed messages, and remaining messages.

RESULTS

C4.5 generated the rules shown in Figure 1. The results of the weeding and the rules on the test set are shown in Tables 3 and 4. Interpreting results of maybe as flames, 98% of the okay messages were correctly classified, as were 64% of the flames. The machine classifications for messages that human volunteers disagreed on were considered to be don't cares, reducing the size of the test set from 502 to 460 messages.

```
If (Imperative-short (13)  > 0 ^
   Condescension-somewhat (21) <= 0 ^
   site-specific-insult (29) > 0)
  (Imperative-short (13) <= 1 ^
   Insult-recipient (25) > 0)
  (Insult-other (28) > 0 ^
   epithet (30) > 0)
  (Imperative-short (13) <= 0 ^
   Profanity-no villain (19) > 0)
  (Appositive-guys (1) > 0 ^
   site-specific-insult (29) > 0)
♥class flame

If (Appositive-NP (3) <= 0 ^
   Insult-villain (27) <= 0 ^
   site-specific-insult (29) <= 0 ^
   epithet (30) <= 0 ^
   exclamation-points (47) <= 2)
  (Appositive-NP (3) <= 0 ^
   Imperative-short (13) <= 0 ^
   site-specific-insult (29) <= 0 ^
   epithet (30) <= 0)
♥class ok
```

Figure 1: Ordered rules generated by C4.5. Numbers in parentheses are rule numbers

We also tried using linear regression, which proved less successful. The nominally independent variables were the 47 features plus 3 binary features indicating whether the message came from NewtWatch, FAIR, or The Right Side of the Web. The dependent variable was 1 when the message was rated okay and 0 when rated a flame. Messages for which there was disagreement were not used. When a least squares analysis was performed on a subset of the 720 messages described earlier as the training data, the resulting coefficients proved very accurate for the reserved portion but substantially less accurate for the test set, correctly identifying 97% of the okay messages but only 39% of the flames. We think the reason for the different performance is that the features had been tweaked to be consistent over the first set of 720 messages. If an insulting adjective appeared in a message in this set and was not recognized as insulting, it was manually added to the system. The test set of 502 messages, on the other hand, was entirely out-of-sample, analyzed only after the system had been frozen. Besides its performance, another disadvantage of linear regression is that it requires computing (or at least bounding) all of the features, while the decision tree algorithm requires the computation of only some of the features.

	Human classification		
	okay	*maybe*	*flame*
okay	422(98%)	34 (43%)	10 (36%)
maybe	6 (1%)	11 (14%)	8 (29%)
flame	4 (1%)	34 (13%)	10 (36%)

Table 3: Confusion matrix for test set

	Human classification	
	okay	*flame*
okay	422(98%)	10 (36%)
flame/maybe	10 (2%)	18 (64%)

Table 4: Collapsed confusion matrix for test set

Discussion

Smokey's Limitations

One flame could not be recognized because the typography was unusual: "G E T O V E R I T". The following flame also managed not to trigger any rules:

> "...is a jelly nosed, poodle stomping, candy-brained cow clump. For the champion of American mythology, he sure knows how to knock down a common law tradition which protects the middle class" [ellipsis ("...") in original]

While an additional heuristic suggested by Phil Leone would recognize the cascaded adjectives in the first sentence as an insulting structure, the lack of an explicit subject makes the first sentence hard to interpret.

Other flames pass by using friendly phrases sarcastically, such as: "Keep up your efforts because I see them as truly benign and pointless." Others are not sarcastic but are the exceptions to rules; consider the following message, as annotated by Smokey:

> **Praise (delight)** : I'm glad to see that your incessant name calling and whining hasn't stopped....

> **Praise (delight)** : As long as it continues, I'm glad to say I'll remian on the winning side of politics.

Because the chance that a message that triggers the *Praise-delight* rule is a flame is only 4.5%, Smokey makes a reasonable, but wrong, evaluation.

Limitations to Flame Recognition

While some limitations on automatic flame identification are due to current natural-language recognition technology, others are inherent. Fluent human readers are sometimes unable to tell whether a given message is friendly or sarcastic.

More practical problems are recognizing sarcasm and innuendo and making sense of complex sentences and mistakes in grammar, punctuation, and spelling, which are all too common in email. Here are some examples:

> *Sarcasm:* "Thank you for recognizing the power of Newt.... Keep up the good work!" [This was sent to NewtWatch.]

> *Grammar, etc., mistakes:* "What on earth a BIGGOT like you is doing walking onthe face of earth?"

> *Innuendo:* "Only cowards, cheats, thieves and liars hide behind pseudonyms." [The program cannot infer that the sender is referring to the recipient, who uses a pseudonym.]

Fortunately, statements that are meant to be insulting tend to be near other insults, allowing a message to be correctly labeled even when individual sentences cannot be.

Possibilities for future work include learning from dictionaries and thesauri (Dolan et al 1993), user feedback, or proximity to known insults; morphological analysis; spelling and grammar correction; and analyzing logical parse trees of sentences.

Related Work. Surprisingly little has been written on the grammar of insults in English. Ruwet (1982) has written about the grammar of French insults. Jochnowitz (1987), Quang (1992), and Allan and Burridge (1991) have written about idiomatic epithets in English. There are numerous lists of and articles about dirty words; see the bibliography in (Jay 1992) or the publications of Maledicta Press. Jay (1990, 1992) has had students rate the offensiveness of various taboo words; physiological responses to insults (Dillard and Kinney 1994) have also been measured. Hayakawa and Hayakawa (1990) and Trippett (1986) have written about the emotional content of terms, particularly political ones.

Automatic categorization of texts has been a major area of information retrieval research (Lewis 1992, Lewis and Hayes 1994). Sack has written a system to automatically determine the ideological bias of a text (Sack 1995). Email classification through regular expressions is already in use, such as through the mail program extensions Procmail, Mailagent, and Filter. A different method of filtering unstructured text is through collaboration, such as Tapestry (Goldberg et al 1992), allowing users to rate individual pieces of bulk mail (from mailing lists or news groups) or individual senders; other users can decide whether to read messages based on others' appraisals. The widely used LISTSERV list maintenance program (L-Soft 1995) includes a proprietary algorithm to detect "spam.," inappropriately crossposted messages, usually advertisements. If the software determines that a user has sent spam, the message and subsequent ones from the same user will be sent to the list owner for approval, combining automatic and social filtering.

Implications. One advantage of mailbox filters such as Smokey is that they do not infringe on freedom of speech. People are both free to write what they wish to willing readers and to not read anything they don't want to. Assuming individuals can train their own filters, nobody will be able to control what anybody else can read.

While Smokey isn't perfect, it could be used now, however, to prioritize mail. Maintainers of controversial web sites who are overwhelmed by mail could use it to move suspected non-flames up in priority. They could avoid suspected flames when busy or when already in a bad mood, without delaying much inoffensive email. Similar techniques could be used for other email-related tasks, such as eliminating unsolicited advertisements or routing mail sent to a general company address to the right individual.

A new "arms race" is starting. As these and similar rules get published, flamers will learn how to get around them. Still, there is a net benefit, since the obscene expressions that affect people most emotionally can be eliminated, and most flamers will not be knowledgeable about the defense systems, especially if they are tailored to individuals.

Acknowledgments. I am grateful to Diana Peterson, Simon Corston, Deb Coughlin, Bill Dolan, Karen Jensen, and Lucy Vanderwende for their assistance with natural language processing and linguistics; to David Heckerman, Carl Kadie, John Miller, and Eric Anderson for advice on decision theory and statistics; and to David Lewis for invaluable encouragement and help with information retrieval techniques and the writing of this paper. This

work would not have been possible without the support of Daniel Weise, Dan Ling, and Rick Rashid and the help of Kevin Shields. I received outstanding library support from Pamela Bridgeport and Keith Barland and technical support from Robert Eberl, Ken Martin, and Karen Burks. I am grateful to the individuals and organizations that contributed email: Jeff Donnels, Matt Dorsey and Ted Schnackertz, Michael Ernst, and Mike Silverman. I had valuable discussions with Reinhold Aman, Cathy Ball, Gene Ball, Oren Etzioni, Chris Gual, David Kurlander, Phil Leone, Nate Osgood, and David Perlmutter. Timothy Jay provided me with the most recent version of his excellent bibliography.

References

Allan, Keith and Burridge, Kate. *Euphemism & Dysphemism.* Oxford University Press, 1991.

Brown, Penelope and Levinson, Stephen. *Politeness: some universals in language usage.* Cambridge University Press, 1987.

Dolan, William B.; Vanderwende, L.; and Richardson, S. Automatically Deriving Structured Knowledge Base from On-line Dictionaries, in Proceedings of the Pacific Association for Computational Linguistics, 1993.

Dillard, James Price and Kinney, Terry A. "Experiential and Physiological Responses to Interpersonal Influence." *Human Communication Research*, vol. 20, no. 4 (June 1994), 502-528.

Donnels, Jeff. The Right Side of the Web, 1995-1997. http://www.clark.net/pub/jeffd/index.html.

Dorsey, Matt and Schnackertz, Ted, Jr. NewtWatch, 1995-1997. http://www.cais.com/newtwatch/.

Ernst, Michael. Fairness and Accuracy in Reporting, 1995-1997. http://www.fair.org/fair/.

Goldberg, D.; Nichols, D.; Oki, B.; and Terry, D. Using collaborative filtering to weave an Information Tapestry. *Communications of the ACM* 35, 12 (December 1992), pp. 61-70.

Gregersen, Edgar A. A Note on English Sexual Cursing. *Maledicta: The International Journal of Verbal Aggression*, vol. 1 (1977), pp. 261-268.

Hayakawa, S. I. and Hayakawa, Alan R. *Language in Thought and Action.* Fifth edition. Harcourt Brace Jovanovich, 1990.

Herring, Susan. Posting in a Different Voice: Gender and Ethics in Computer-Mediated Communication, in Ess, Charles, ed., *Philosophical Perspectives in Computer-Mediated Communication.* Albany: SUNY Press, 1995.

Infante, Dominic A. and Wigley, Charles J., III. Verbal Aggressiveness: An Interpersonal Model and Measure. *Communications Monographs,* Volume 53, March 1986.

Jay, Timothy B. What are "Fighting Words"? Presented at Eastern Psychological Association, March, 1990, Philadelphia, PA.

Jay, T. B. *Cursing in America.* Philadelphia: John Benjamins, 1992.

Jochnowitz, George. Acceptable but not Grammatical. *Maledicta: The International Journal of Verbal Aggression*, vol. 9 (1986-1987), pp. 71-74.

Kohavi, Ron; John, George; Long, Richard; Manley, David; and Pfleger, Karl. MLC++: A Machine Learning Library in C++, in *Tools with Artificial Intelligence*, IEEE Computer Society Press, 1994, pp. 740-743. See "http://www.sgi.com/Technology/mlc/"

L-Soft International, Inc. List Owner's Manual for LISTSERV 1.8b. http://www.lsoft.com/manuals/ownerindex.html., 1995.

Lewis, David Dolan. Representation and Learning in Information Retrieval. Doctoral Dissertation, CS Department, Univ. Of Massachusetts, Amherst, 1992.

Lewis, David D. And Hayes, Philip J. Guest Editorial. *ACM Transactions on Information Systems* 12(3) (July 1994):231.

Maledicta Press. P.O. Box 14123, Santa Rosa, CA 95402. 707-523-4761.

Quang Phuc Dong. English Subjects without Overt Grammatical Subject, in Zwicky, A.M., Salus, P.H., Binnick, R.I., & Vanek, A.L., eds., *Studies out in left field: Defamatory essays presented to James D. McCawley.* John Benjamins, Philadelphia, 1992.

Quinlan, J. Ross. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Richardson, Steve. Bootstrapping Statistical Processing into a Rule-based Natural Language Parser, in *The Balancing Act: Combining Symbolic and Statistical Approaches to Language: Proceedings of the Workshop.* Las Cruces, NM, 1994, pp. 96-103.

Ruwet, Nicolas. *Grammaire des Insultes et Autres Etudes.* Editions du Seuil, Paris, 1982.

Sack, Warren. Representing and Recognizing Point of View. *Proceedings of the AAAI-95 Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, 1995.

Trippett, Frank. Watching Out for Loaded Words, in Seyler, Dorothy U. and Boltz, Carol J., eds. *Language Power,* pp. 112-115. Random House, New York, 1986.