Фильтрация оскорблений в текстовых сообщениях

Содержание

Задача “Detecting Insults in Social Commentary”

Обзор решений

- Bag of words
- Прочие методы

Статистический парсинг

- POS tagging
- PCFG. Consistency tree
- Dependency tree
- Smokey
- Фильтр для социальных сетей

Detecting Insults in Social Commentary

Данные:

- Список сообщений - {время, текст}
- Обучающая выборка ~ 4000 сообщений

Требуется классифицировать сообщения на два класса: оскорбительние и не оскорбительные.

Функционал качества – AUC

Два набора выборок: промежуточная и окончательная

Detecting Insults in Social Commentary

Линейная регрессия

Основные признаки:

- количество слов “you” и его производных
- количество слов из badword list
- количество пар вида “you … <badword>”
- kNN по времени сообщения

Результат:

- промежуточные данные – 0.861
- окончательные данные - 0.666

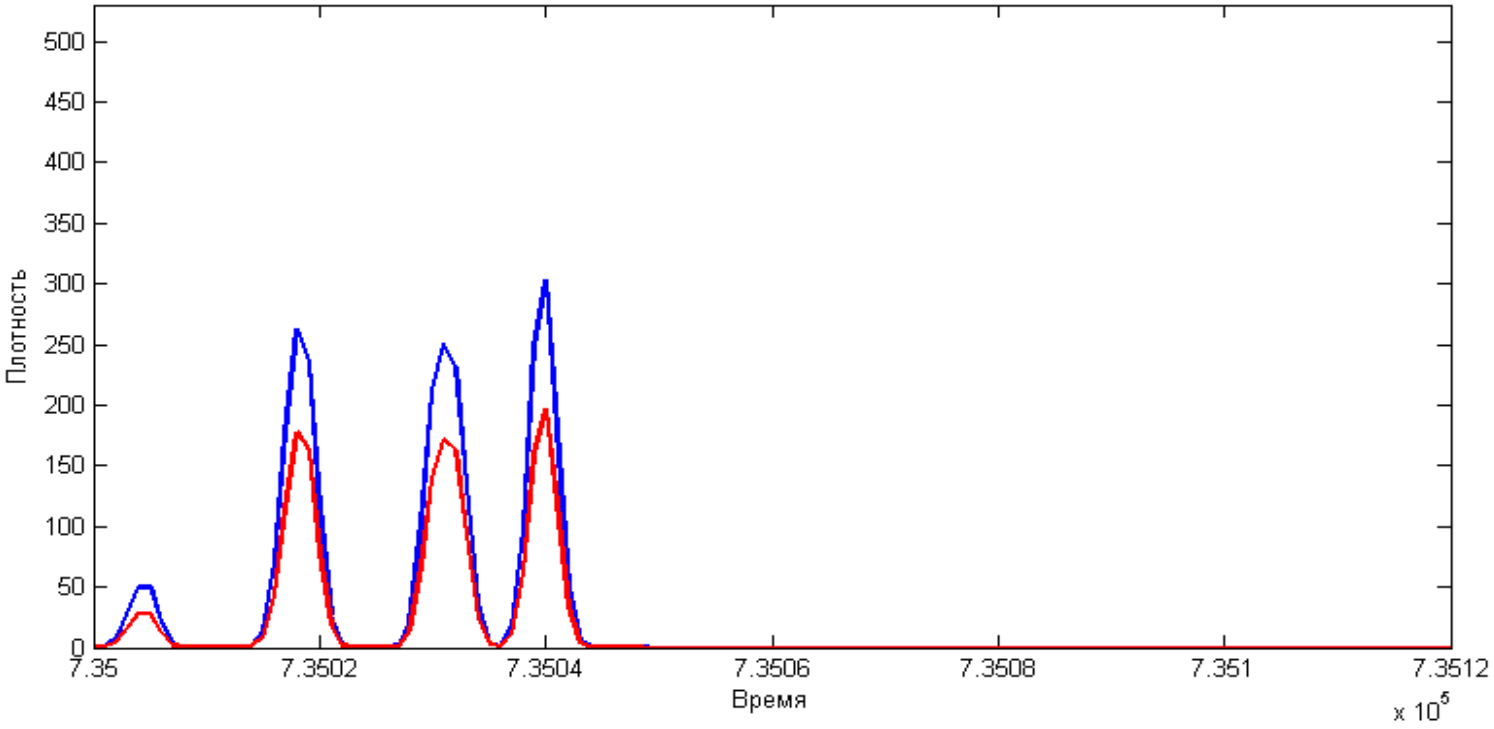
Detecting Insults in Social Commentary

Проблемы:

- плохое качество данных
- разное распределение сообщений по времени в промежуточных и окончательных данных

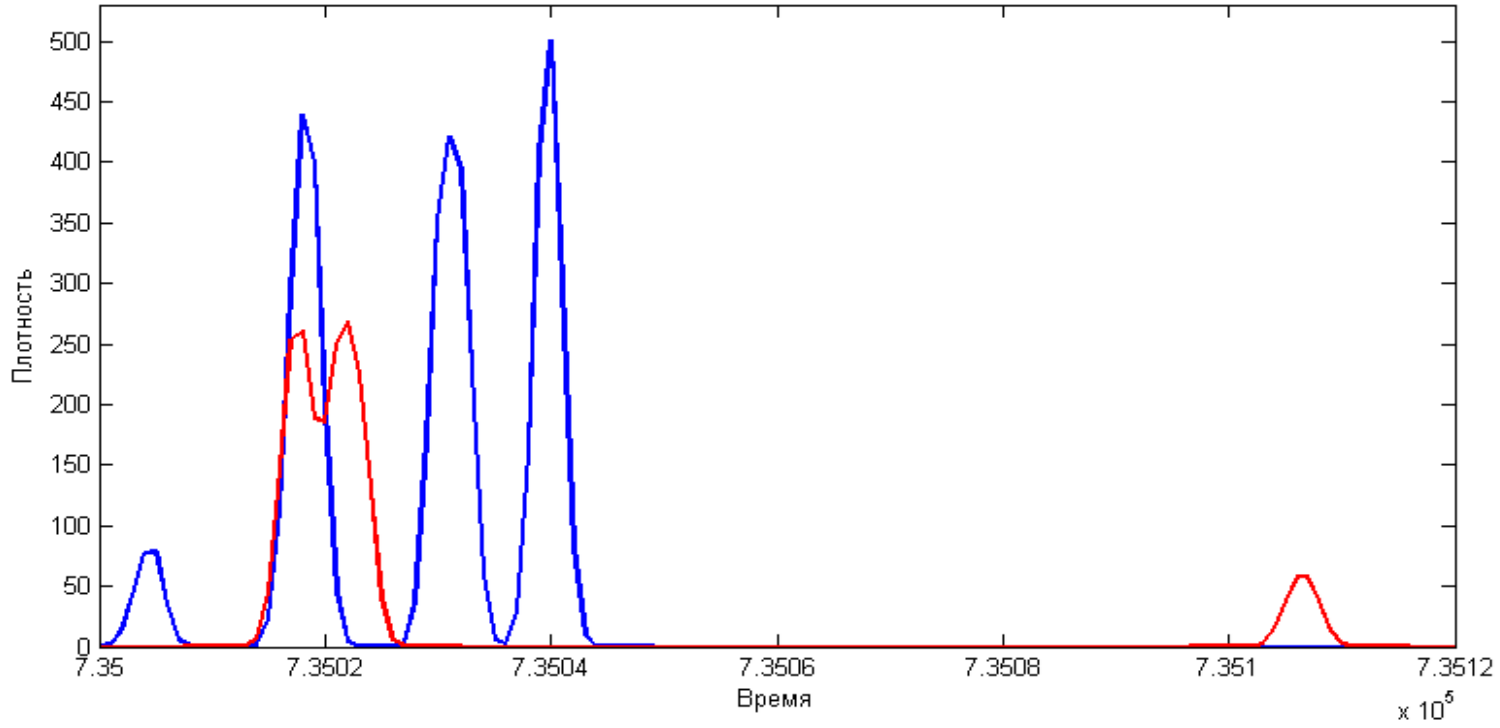
Detecting Insults in Social Commentary

Плотность сообщений по времени (промежуточные данные)



Detecting Insults in Social Commentary

Плотность сообщений по времени (конечные данные)



Обзор решений.
Bag of words

1. Все слова нумеруются
2. Генерируется мартица признаков: m_{ij} – количество слов с номером j в i-м сообщении
3. Применяется нормировка TF-IDF (term frequency – inverse document frequency)
4. Затем логистическая регрессия

Помимо отдельных слов можно рассматривать буквенные и словесные n-граммы

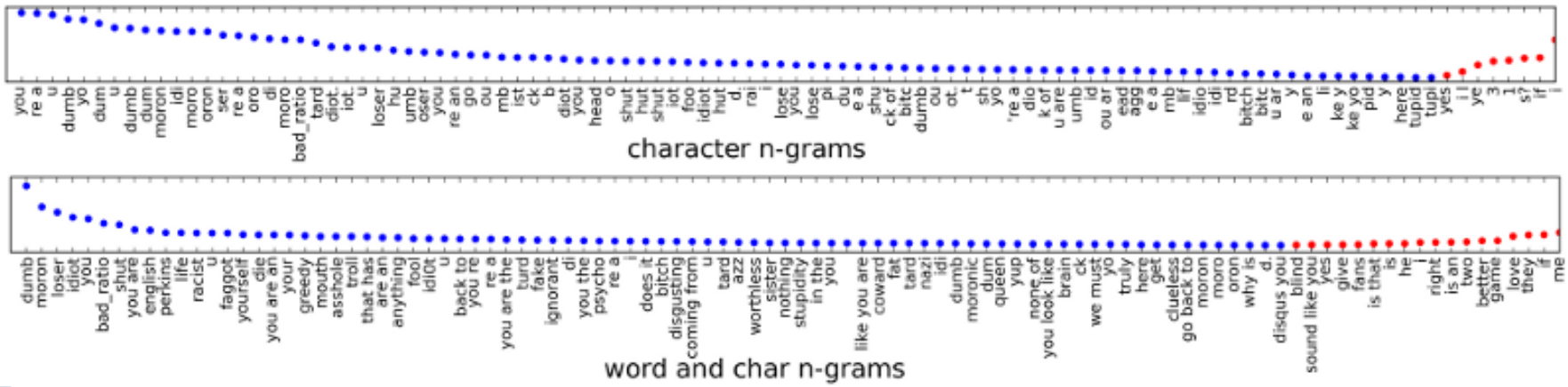
Обзор решений.
Bag of words

$$TF_{ij} = \frac{m_{ij}}{\sum_k m_{ik}}$$

$$IDF_{ij} = -\log \frac{\sum_k [m_{kj} > 0]}{l}$$

$$TF_{ij} * IDF_{ij} = -\frac{m_{ij}}{\sum_k m_{ik}} \log \frac{\sum_k [m_{kj} > 0]}{l}$$

Обзор решений.
Bag of words



- наибольший положительный вес
- наибольший по модулю отрицательный вес

Обзор решений. Прочие методы

Также использовались:

- L1-регуляризация
- Stochastic gradient
- SVM
- kNN
- Badword list
- Information gain
- Maximum entropy
- Chi-squared

Статистический парсинг. POS-tagging

Part-Of-Speech (POS) tagging – задача определения частей речи у слов в тексте.

- Скрытые марковские модели
- Статистические методы
- SVM
- Maximum entropy classifier
- Nearest neighbour

Качество – около 95%

Статистический парсинг. PCFG

Стохастическая контекстно-свободная грамматика (Porbabilistic context-free grammar, PCFG) – контекстно-свободная грамматика, в которой для каждого нетерминала на множестве всех правил, раскрывающих его, задано вероятностное распределение.

PCFG – это четверка $(\sum, N, S, P)$, где $\sum$ - множество терминалов, N – множество нетерминалов, S – начальный символ, $P = \{E \rightarrow \xi(p)\}$ – множество правил вывода.

Статистический парсинг. PCFG

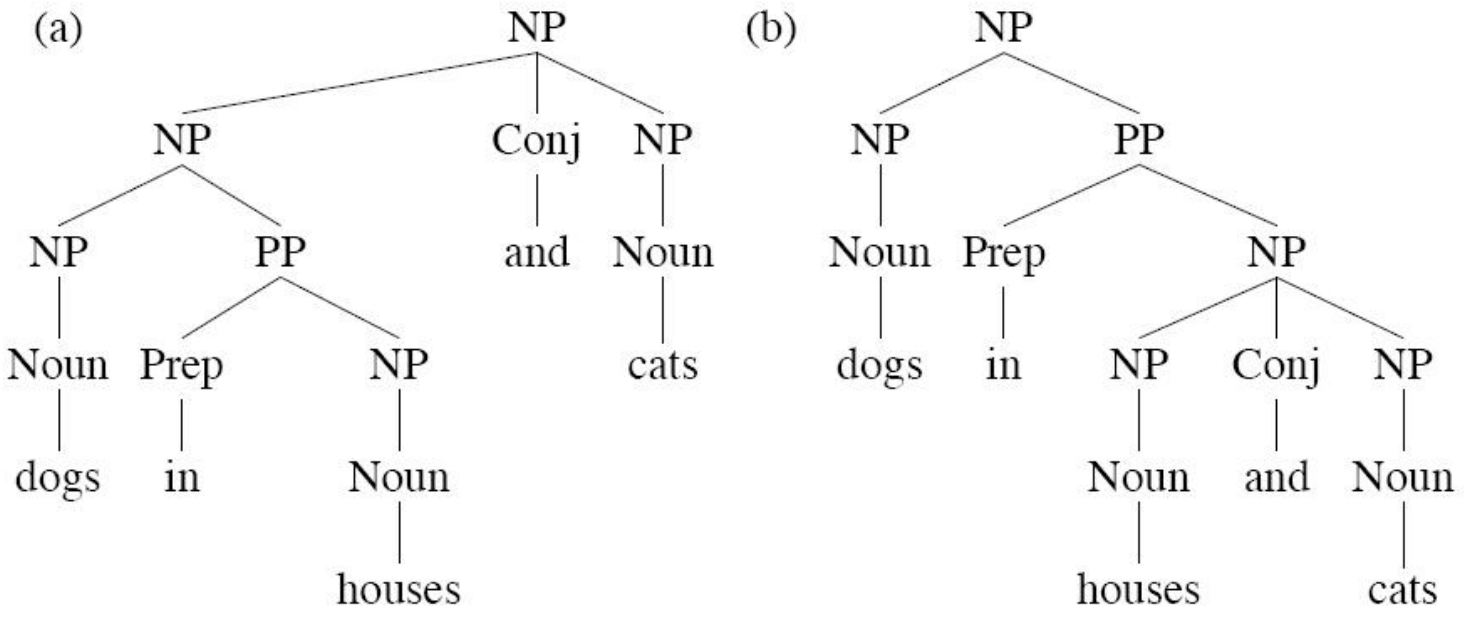
Терминалы – все части речи.

Нетерминалы – грамматические конструкции (словосочетания, предложения) разных типов.

Penn treebank – корпус документов из ~1.2 миллиона слов, составленный из статей Wall Street Journal. Содержит 36 нетерминальных символов (Penn treebank tags)

Статистический парсинг. PCFG

Пример дерева разбора (parse tree, consistency tree)



Статистический парсинг. PCFG

Две задачи:

- Задача настройки вероятностей
- Задача разбора строки по грамматике

Статистический парсинг. PCFG

Настройка вероятностей правил

Для обучения используется корпус документов с построенным разбором.

$$P(E \to \xi | E) = \frac{count(E \to \xi)}{count(E \to \eta)}$$

Другой способ – максимизация правдоподобия (Inside-Outside algorithm)

Статистический парсинг. PCFG

Разбор строки.

Probabilistic Cocke-Younger-Kasami (CYK) algorithm

Грамматика приводится к нормальной форме Хомского (Chomsky Normal Form, CNF), которая допускает только правила трех видов:

$$E \to AB$$

$$E \to \alpha$$

$$E \to \varepsilon$$

Статистический парсинг. PCFG

Разбор строки.

Probabilistic Cocke-Younger-Kasami (CYK) algorithm

$f(i, j, E)$ — максимум правдоводобия вывода из E подпоследовательности слов с i-го по j-е.

Правдоподобие вывода предложения равно $f(1, n, S)$, где n – длина предложения.

Если E – терминал, то $f(i, j, E)$ равно 0 или 1.

Статистический парсинг. PCFG

Разбор строки.

Probabilistic Cocke-Younger-Kasami (CYK) algorithm

Если E нетерминал:

$$f(i,j,E) = \max[\max_{\alpha \epsilon \Sigma} f(i,j,\alpha) P(E \to \alpha | E)) ,$$

$$\max_{A,B \epsilon N;\, k \epsilon [i,j]} \big(f(i,k,A) * f(k,j,B)\big) P(E \to AB | E)]$$

Применяется правило, на котором достигается максимум. Таким образом строится дерево разбора.

Статистический парсинг. PCFG

Разбор строки.

Probabilistic Cocke-Younger-Kasami (CYK) algorithm

Реализация с помощью метода динамического программирования.

Сложность алгоритма – $O(n^3 * |P| * |N|)$.

Если брать не максимум, а суммировать, то получим правдоподобие вывода строки.

Статистический парсинг. Dependency structure

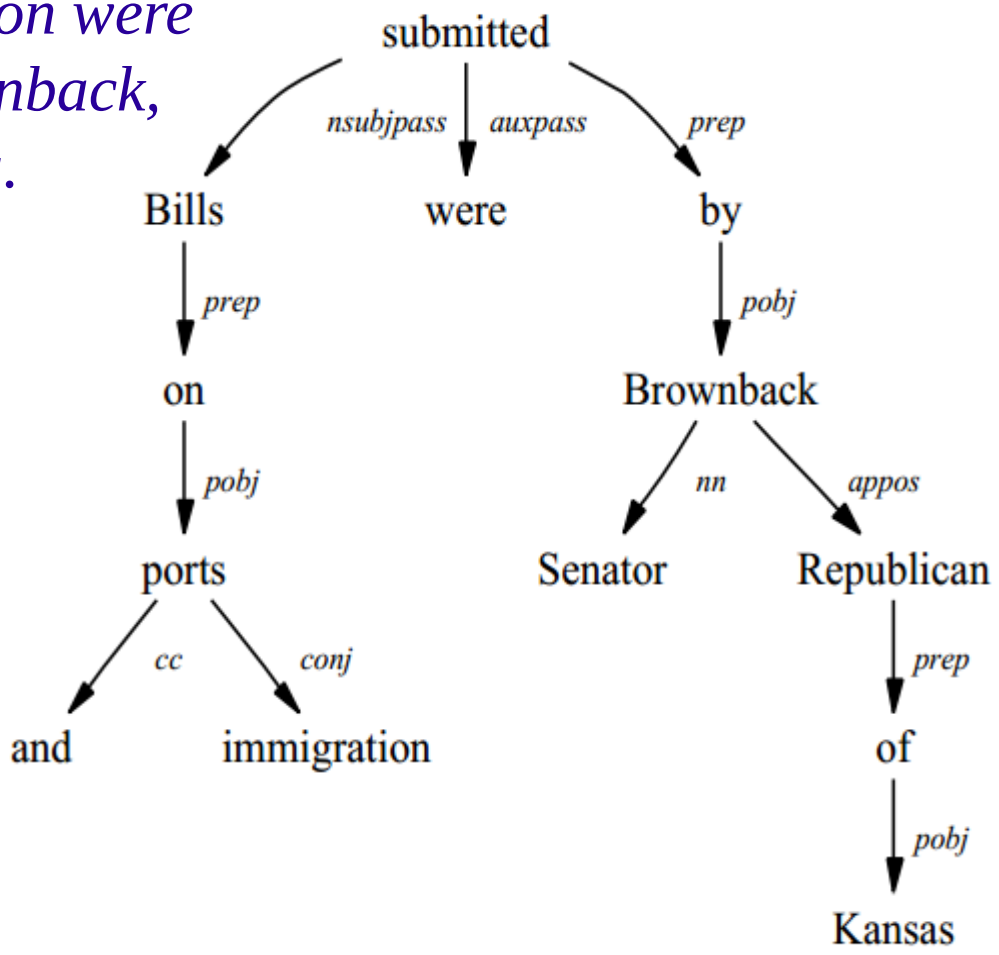
Dependency structure – дерево зависимостей слов в предложении.

Некоторые виды зависимостей:

- Семантические
- Морфологические
- Синтаксические (наиболее используемые)

Статистический парсинг. Dependency structure

Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas.



Smokey

Три класса сообщений: offensive, ok, maybe.

1. Разбиение сообщений на предложения.
2. Анализ предложения и построение дерева разбора с помощью Microsoft Research Natural Language Processing Group parser.
3. Поиск шаблонов в предложении.
4. Построение вектора признаков предложений.
5. Построение вектора признаков сообщений путем суммирования векторов признаков предложений.
6. Построение дерева решений C4.5

Smokey.
Шаблоны

1. Noun phrase, начинающиеся с “you”.
2. Imperative (повелительные) statements.
3. Second-Person Rules
4. Profanity (site-specific)
5. Condescending (“снисходительные”) statements
6. Insults (слова из badword list рядом с you или this как местоимением)

Smokey.
Шаблоны

7. Оскорбительные эпитеты.
8. Вежливые выражения (please, thanks …)
9. Praise (одобрительные) statements (kudos, bless, …)
10. Другие

#	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	"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	"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%

#	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! Your page is a joke!"	6	50%	17%	**33%**
47	Contains exclamation points	"neWt gInGriCh iS evil!!"	460	81%	10%	**9%**

Smokey

Тренировочная выборка: 720 сообщений.

Результат:

	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%)

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

Статистический парсинг. Фильтр для социальных сетей

Фильтрация текста в соответствии со следующими правилами:

- Если в словосочетании зависимое слово содержит оскорбление, то его можно убрать.
- Если оскорбление содержится в *basic pattern*, то необходимо убрать всё предложение целиком.

Статистический парсинг. Фильтр для социальных сетей

1. Анализ текста (с помощью Stanford NLP group parser): POS-tagging, typed dependency generation, parse tree.

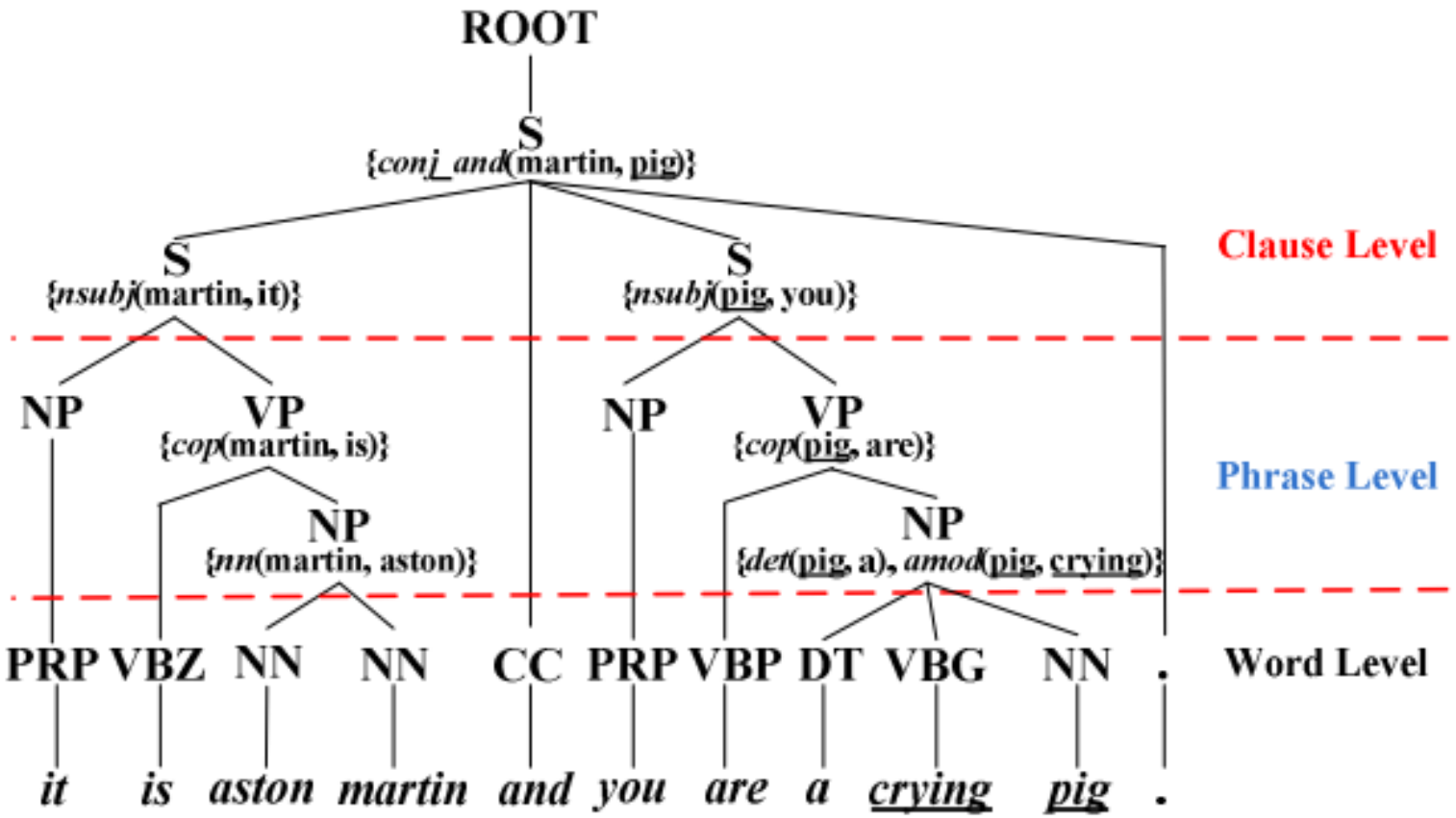
2. Объединение dependency structure и consistency structure. Результат – relations tree (RelTree).

input : a parse tree $PTree$,
a set of typed dependency relations $TDset$
output: a RelTree $RelTree$

```
1  RelTree ← PTree;
2  Remove all word nodes in RelTree;
3  Traverse RelTree in postorder  foreach node n
   visited do
4      if n is a leaf node then
5          n.wordset ← {n};/*create word nodes*/
6      end
7      if n is not a leaf node then
8          n.wordset ← ∅;
9          foreach direct child node c_i do
10             n.wordset ← n.wordset ∪ c_i.wordset;
11             n.rel ← ∅;
12             foreach relation T_i(G_i, D_i) in TDset do
13                 if G_i ∈ n.wordset and
                   D_i ∈ n.wordset then
14                     n.rel ← n.rel ∪ T_i(G_i, D_i);
15                     TDset ← TDset − T_i(G_i, D_i);
16                 end
17             end
18         end
19     end
20 end
21 Return RelTree;
```

Статистический парсинг. Фильтр для социальных сетей

RelTree:



Статистический парсинг. Фильтр для социальных сетей

3. Разметка узлов и зависимостей.

Задача – пометить узлы, поджелащие удалению.

Рассматривается зависимая пара T = (G, D). Обозначим:

- P(G) и C(G) – соответственно словосочетание и простое предложение с главным словом G.

  Пример:

  P(aston) = “aston martin”, C(aston) = “it is aston martin”

- H(T) – метка зависимости T, H(n) – метка узла n.

Статистический парсинг. Фильтр для социальных сетей

H(T)	Зависимости
H(P(G))	cop, expl, measure, partmod, poss, possessive, preconj , prep/prepc, purpcl , quantmod, rcmod, ref , tmod
H(P(D))	pcomp, pobj , predet
H(C(G))	complm, mark, rel
H(P(G)) OR H(C(D))	xcomp
H(C(G)) OR H(P(D))	xsubj
H(G) OR H(P(D))	nsubj, nsubjpass
H(G) AND H(D)	conj, nn, number, dep
H(G)	aomp, advcl, advmod, agent, amod, appos, attr, aux, auxpass, cc, ccomp, det, neg, num, parataxis, punct
H(G) OR H(C(D))	csubj, csubjpass
H(P(G)) OR H(P(D))	abbrev, dobj, infmod, iobj, prt

Статистический парсинг. Фильтр для социальных сетей

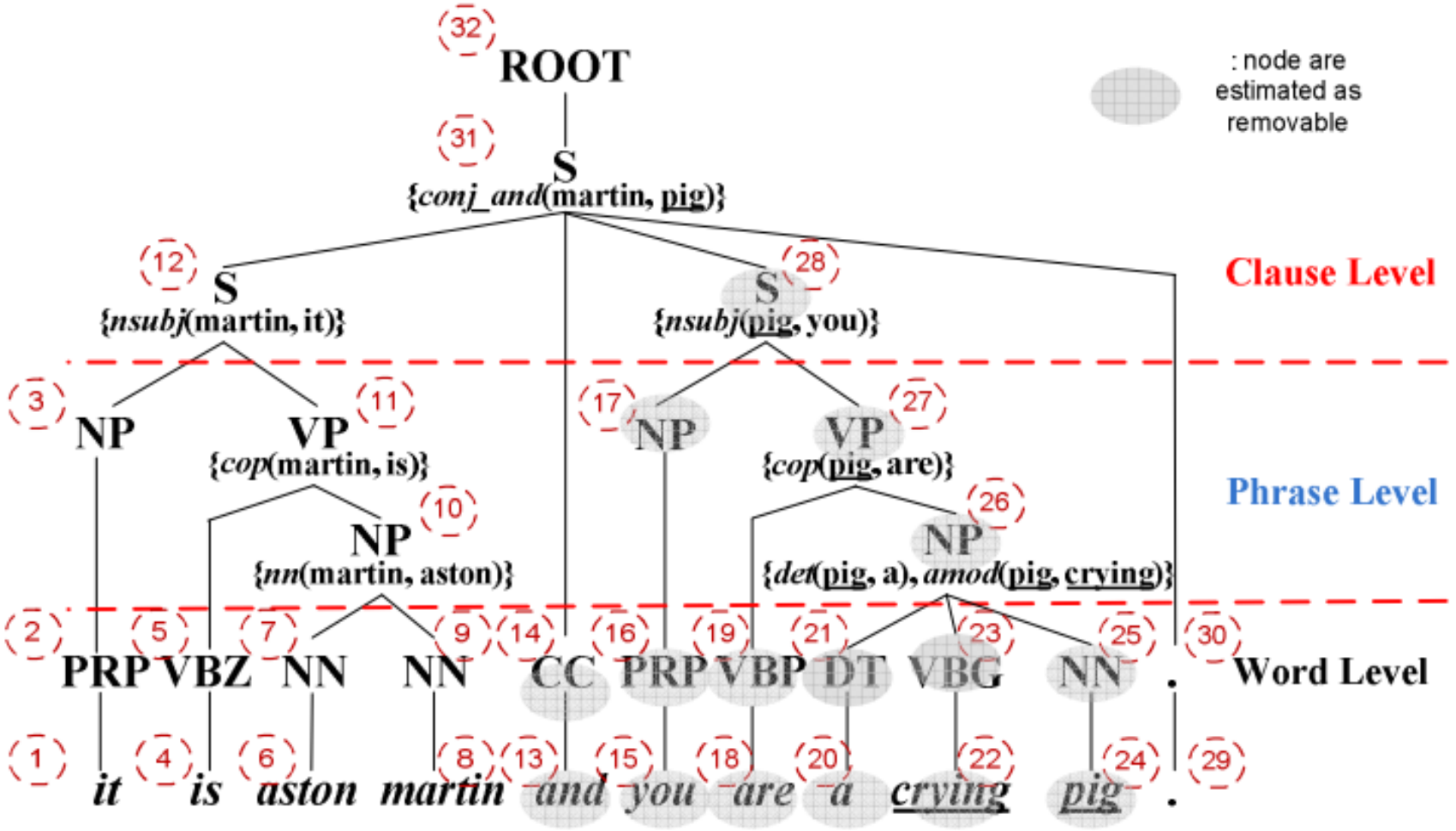
Все листья помечаются проверкой по badword list.

Затем производится обход дерева RelTree в обратном порядке (postorder). Для каждого не листового узла n:

> Просматриваем и размечаем все зависимости n.rel. Если помечается зависимость T(G,D), то прямые потомки n, содержащие G и D, также помечаются.
>
> Если все зависимости n.rel оказались отмечены, отмечаем узел n.
>
> Если n не отмечен, применяем дополнительную эвристику.

```
input  : a RelTree RelTree,
         a blacklist of offensive words Blacklist,
output: a labeled RelTree LebelRelTree
1  LebelRelTree ← RelTree;
2  Label all leaf nodes with offensive words by
   "removable" in LabelRelTree ;
3  Traverse LabelRelTree in postorder  foreach node
   n visited do
4      if n is a leaf node then
5          ignore; /* already labeled */
6      end
7      if n is not a leaf node then
8          if n only has one child node then
9              n.label ← n.child.label;
10         end
11         if n has more than one child node then
12             Estimate the label for n by its associated
               labels, using proposed estimation
               function and heuristic rules;
13         end
14     end
15 end
16 Return LabelRelTree;
```

Статистический парсинг. Фильтр для социальных сетей



Статистический парсинг. Фильтр для социальных сетей

Результат

Тестовая выборка из 2063 предложений, содержащий оскорбления (комментарии с YouTube).

Недостаточная фильтрация: 2.81%

Избыточная фильтрация: 6.25%

Качество: 90.94%

Средняя скорость обработки комментария: 48мс

Спасибо за внимание!
Вопросы?