Maschinelles Lernen und Experemintelles Design (SS 2011) Dr. Caroline Sporleder

Thema: Spam-Filterung mit Naive Bayes

Ghamdan Atef

Spam

- Unerwünschte E-Mails
- Spamfilter filtern relativ gut
- Klassifizieren einer Nachricht
- Ziel: Keine guten E-mails in Spam haben

3 Spamfilter-Methoden

1. Basiert auf Regeln, die vom Programmierer kodiert werden, z.B.: "buy now”

2. Machine Learning-Methoden (Spam Assassin)

3. Mit Hilfe von Naive Bayes

Naiver Bayes-Klassifikator

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_i P(a_i|c)$$

C: {Spam, Nicht-Spam }
P: Wahrscheinlichkeit

- Annahme: Alle Attribute sind stochastisch unabhängig von Wörtern.

Spamfilter mit Naive Bayes

- Training des Naive-Bayes-Klassifikators
- Generieren eines Modells mithilfe von annotierten Daten (Spam und Nicht-Spam)
- Auswahl von Attributen (Eigenschaften), die eine E-mail als Spam oder Nicht-Spam beschreiben
- Erzielung von besseren Klassifikationsergebnissen durch Ausschliessung statistisch irrelevanter Daten

Spamfilter mit Naive Bayes

- Mutual Information:

➢ Zeigt, wie stark das Merkmal X zwischen den Klassen unterscheidet.

➢ Dies ermoeglicht, dass man, wenn man wenig Merkmale aussuchen moechte, die jenigen benutzt, die das beste Klassifikationsergebnis ermoeglichen.

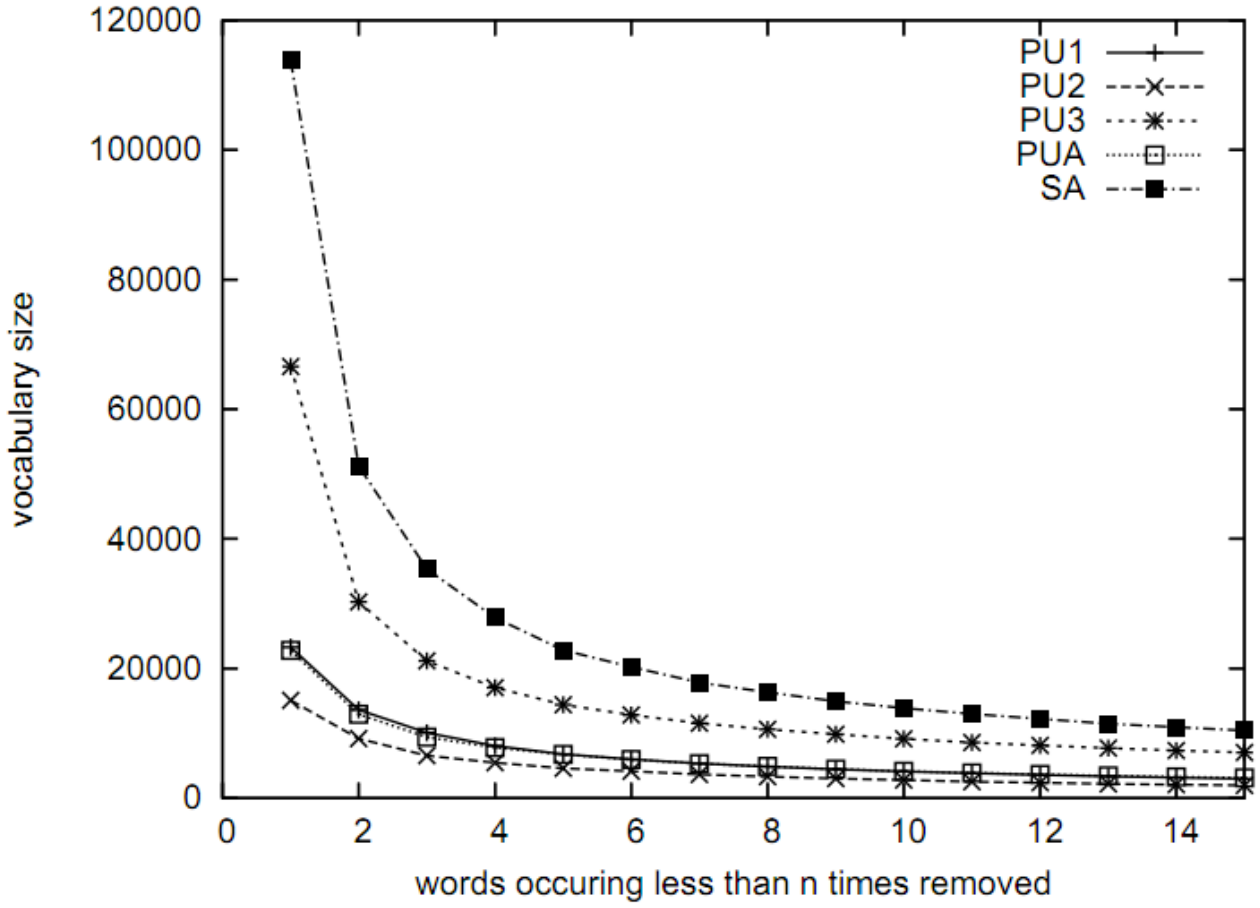
$$\sum_{x \in \{0,1\}} \sum_{c \in C} P(x,c) log \frac{P(x,c)}{P(x)P(c)}$$

Neue Methoden

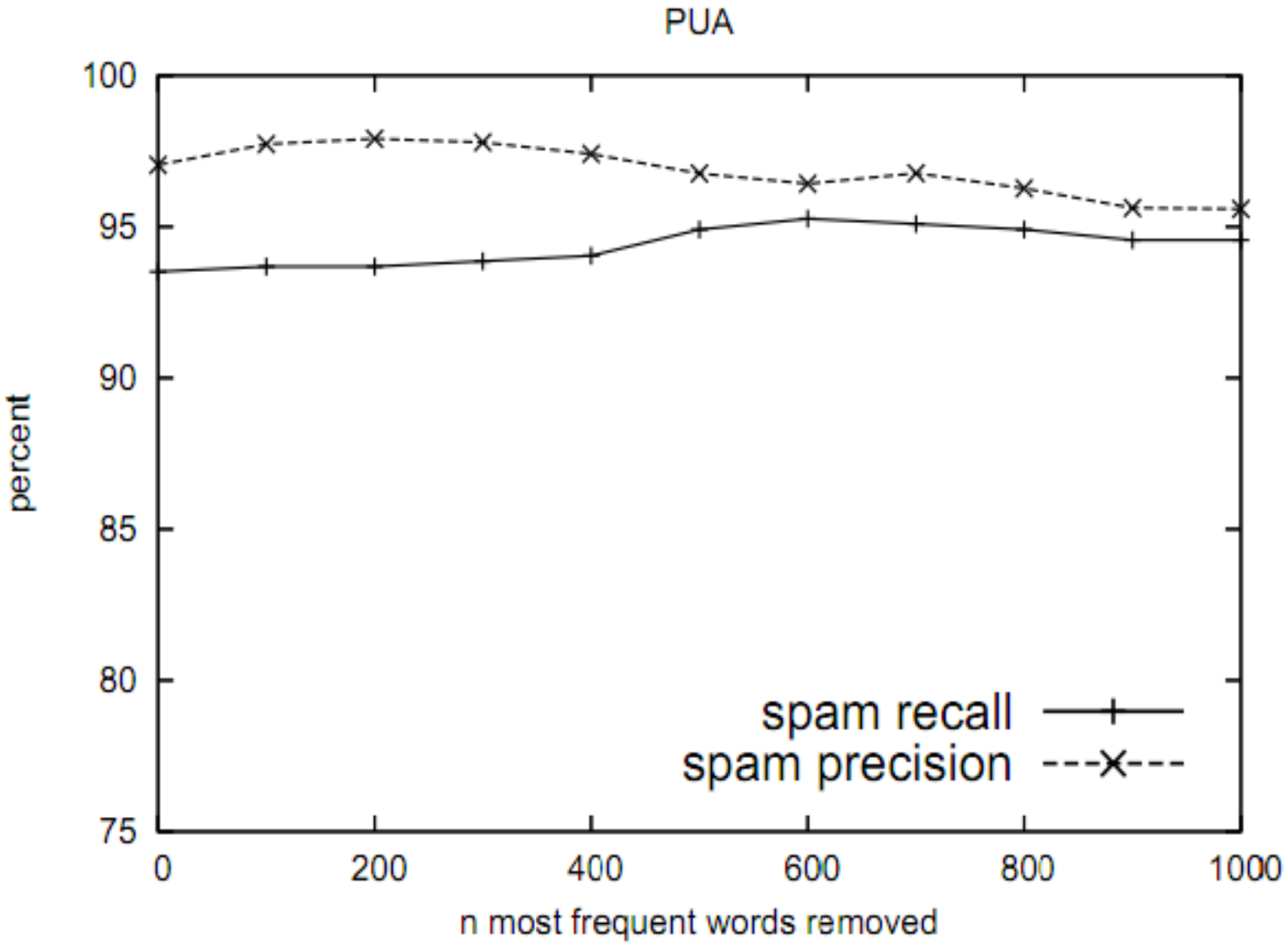
- Es werden häufige Wörter entfernt wie z.B.: der, die, das, ich, und, mit, in, Jahr usw.
- Diese Woerter sind für alle Klassifikationsaufgaben uninteressant
- Vokabular wird kleiner

➢ Ergebnis: schnellere Verarbeitung

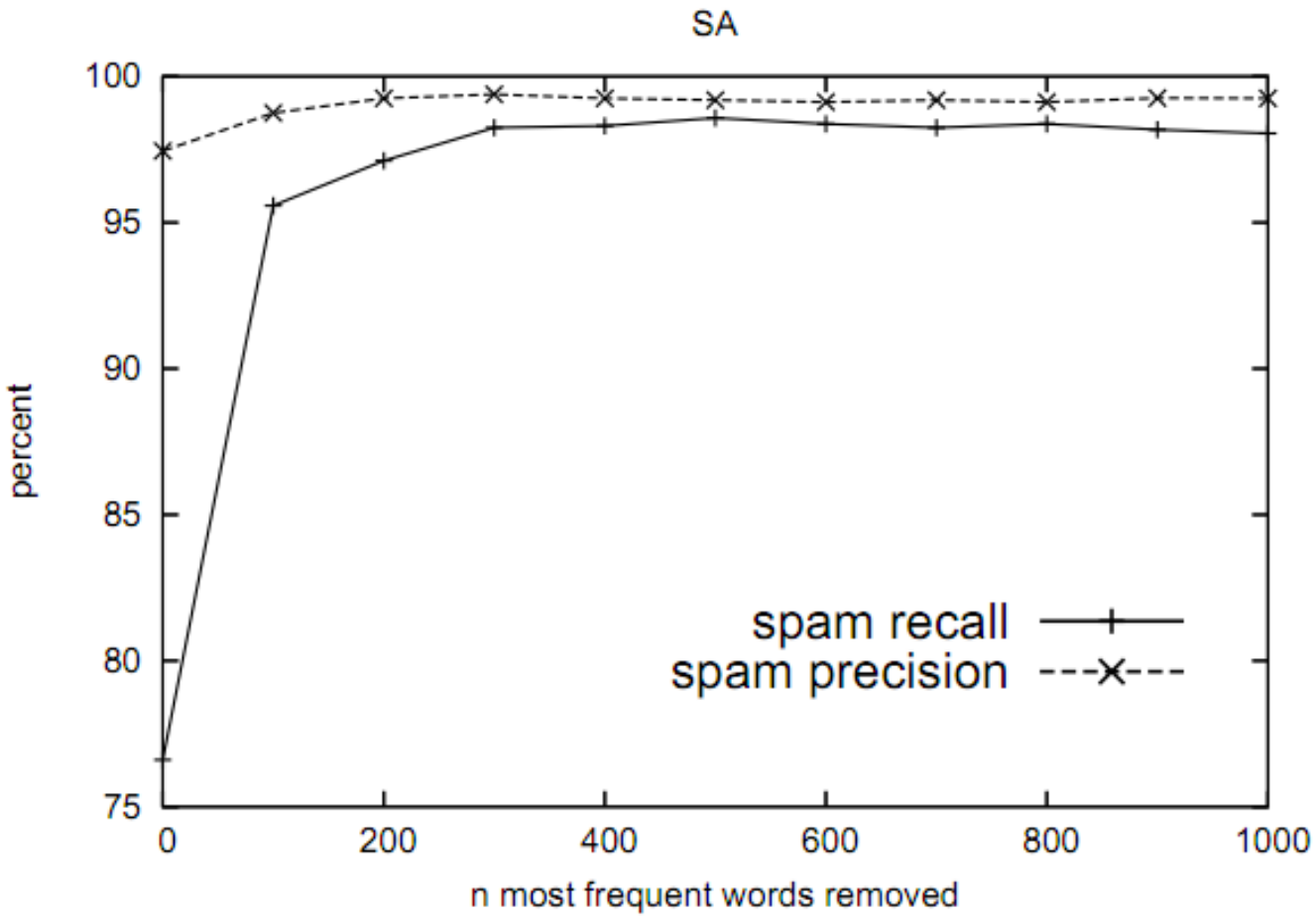
Vokabulargröße wird durch Entfernung von haufigen Woertern beeinflusst



Worthäufigkeit bei PUA



Worthäufigkeit bei SA



Neue Methoden

- Sehr seltene Wörter wie z.B *Notebookstaschenreißverschlussdefekt oder Aölkjasödfkajsdfkadsfjöadsfkl* werden entfernt (sind keine geeignete Basis, um ein Modell zu bilden)

Neue Methoden

- Es wird nicht laenger die Haeufigkeit der einzelnen Wörter beachtet, sondern die Wortposition mit Bi- und Trigrammen

➢ Vokabular waechst wieder (statt einzelne Wörter Bi- und Trigramme)

➢ Vokabular hat in etwa wieder die urspruengliche Größe

Evaluation

- Precision:

$$P = \frac{|S \rightarrow S|}{|S \rightarrow S| + |L \rightarrow S|}.$$

- Recall:

$$R = \frac{|S \rightarrow S|}{|S \rightarrow S| + |S \rightarrow L|}$$

Evaluation

- Testmaterial: 5 Corpora

corpus	messages	spam ratio
PU1	1099	44%
PU2	721	20%
PU3	4139	44%
PUA	1142	50%
SA	6047	31%

Vergleich von verschiedenen Varianten:

- Unigramm-Modell
- Bi-Trigrammmodell

n-grams	R	P	Acc
PU1			
$n = 1$	98.12	95.35	97.06
$n = 1, 2, 3$	99.17	96.19	97.89
PU2			
$n = 1$	97.14	87.00	96.20
$n = 1, 2, 3$	95.00	93.12	96.90
PU3			
$n = 1$	96.92	96.02	96.83
$n = 1, 2, 3$	96.59	97.83	97.53
PUA			
$n = 1$	93.68	97.91	95.79
$n = 1, 2, 3$	94.74	97.75	96.23
SA			
$n = 1$	97.12	99.25	98.95
$n = 1, 2, 3$	92.26	98.70	97.42
$n = 1, 2, 3$	98.46	99.66	99.46

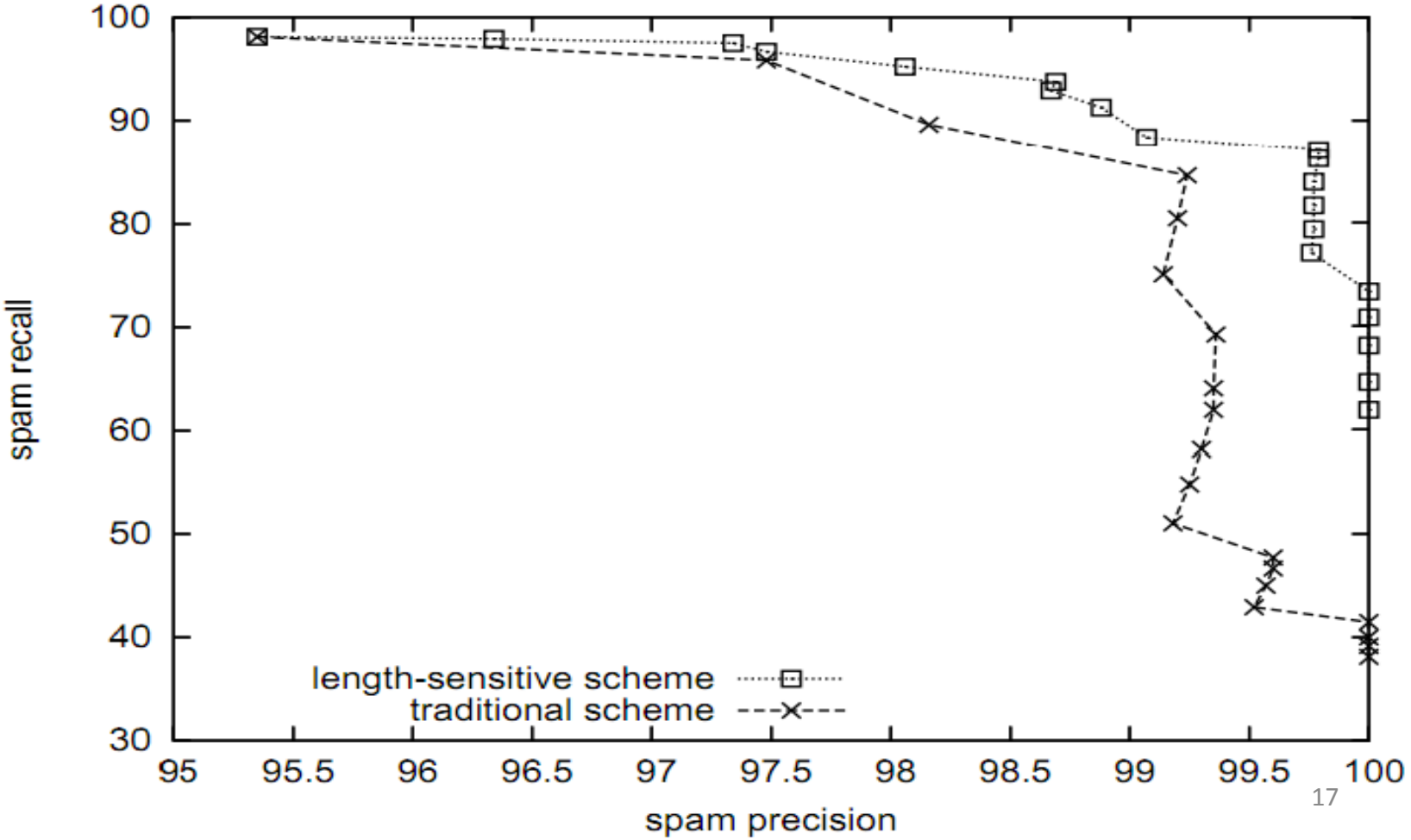
Evaluation

Ergebnisse im Vergleich zu anderen Naive-Bayes-Methoden

learner	R	P	Acc
PU1			
Androutsopoulos	99.38	89.58	94.59
Hovold	98.12	95.35	97.06
Flexible Bayes	97.08	96.92	97.34
PU2			
Androutsopoulos	90.00	80.77	93.66
Hovold	97.14	87.00	96.20
Flexible Bayes	79.29	90.57	94.22
PU3			
Androutsopoulos	94.84	93.59	94.79
Hovold	96.92	96.02	96.83
SVM	94.67	96.48	96.08
PUA			
Androutsopoulos	94.04	95.11	94.47
Hovold	93.68	97.91	95.79
Flexible Bayes	91.58	96.75	94.21

Evaluation

- Unterschiedliche Weighting-Schemas im Vergleich:



Danke für die Aufmerksamkeit!!!

Literatur

- Johan Hovold

Naive Bayes Spam Filtering Using Word-Position-Based Attributes

Diskussion