Evaluation of Classification Models for Opinion Detection

J. Savoy, O. Zubaryeva
University of Neuchâtel

Outline

- **Context, Examples and Applications**
- Text Categorization
- Approaches used
  - Naïve Bayes
  - SVM
  - Z-Score
- Corpus Description
- Evaluation and Discussion

Context

- Terminology
  - Sentiment or opinion
- A thought, view, or attitude, especially one based mainly on emotion instead of reason
  - Opinion detection
- Sentiment analysis aka opinion mining
- Use of natural language processing (NLP) and machine learning methods to automate the extraction and classification of **opinion** from typically unstructured text

Introduction

- Two main types of textual information
  - Facts and opinions
- We will not cover sentiment (fear, joy, etc.)
- Most current text information processing methods (e.g., web search, text mining) work with factual information.
- We do not use a prior and precise definition
  - the users specify what is an opinion, what is a fact

Example of a Customer Review

- «*As far as I'm concerned*, the iPhone3G is already *the best* possible smartphone you can buy. Although some phones such as the PALM PRE and the TMOBILE G1 have their own niches, the iPhone is the best all around touchscreen device and if you are a newcomer to AT&T and iPhone… this is the phone *you want to have* if you want the *most versatile* and the *best supported* device on the market.
Unfortunately, as I pointed out in the 3G review, AT&T is the main problem for the iPhone. *They are basically extorting* us with ridiculous data and text prices. With no SMS and unlimited data added to the cheapest minutes plan, I am still paying over $80 a month for this phone…»
- Real word: spelling errors, SMS language, propaganda, …

5

Examples from Newspapers

- Fact
"Five years ago, there were no Internet-related information businesses."
- Negative opinion
Since the United States is Korea's most important trade partner, the Korean economy was also affected immediately."
- Positive opinion
"I believe that we have found the appropriate balance," he said.
- Mixed opinion
"However, it is important that we not place excessively high expectations on the summit."

6

Applications

- Automatic email filtering (spam, suspicious "terrorist" content)
- Consumer Information (product reviews, spam product reviews)
- Marketing (consumer attitudes, trends)
- Social (find individuals and communities by interest)
- Opinion Question/Answering
- Ads placements (placing ads in the user-generated content)
- Authorship identification (male/female, particular writer)
- Etc..

7

Example

Outline

- Context, Examples and Applications
- **Text Categorization**
- Approaches used
  - Naïve Bayes
  - SVM
  - Z-Score
- Corpus Description
- Evaluation and Discussion

9

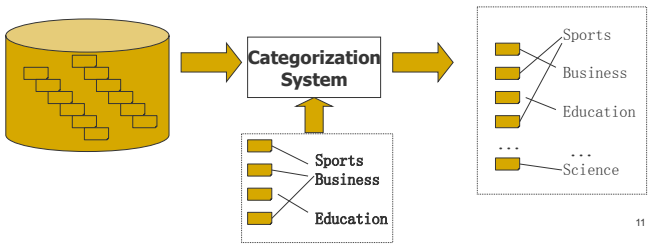
Using NLP methods?

- **Adjectives**
  - objective: **red, metallic**
  - positive: **honest, important, mature, large, patient**
  - negative: **harmful, hypocritical, inefficient**
  - subjective (but not positive or negative): **curious, peculiar, odd, likely, probable**
- **Verbs**
  - positive: **praise, love**
  - negative: **blame, criticize**
  - subjective: **predict**
- **Nouns**
  - positive: **pleasure, enjoyment**
  - negative: **pain, criticism**
  - subjective: **prediction, feeling**

10

Text Categorization

- Predefined categories C (categories may form hierarchy)
- Set of labeled document examples D (to learn)
- A standard classification (supervised learning) problem



11

Which features are useful?

- Which features to use?
  - Word (unigram)
  - Lemma
  - POS
  - Punctuation
  - *n*-gram
  - Phrase
- Sequence of features
  - Bag-of-words (IR)
  - Annotated lexicons (WordNet, SentiWordNet)
  - Syntactic patterns (JJR, NN JJ)
  - Paragraph structure, layout

12

Selecting features?

- Do we need to select them?
  - With a min. frequency (*tf*) (*hapax*)
  - Ignoring stopword? (the, in, of, has, year)
  - Stemming? (historical → historic)
  - Morphological analysis (said → say)
  - Using a weighting scheme?
    - mutual information
    - information gain
    - χ^2
    - *tf·idf*

13

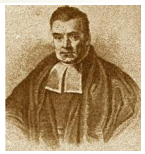
Outline

- Context, Examples and Applications
- Text Categorization
- **Approaches used**
  - Naïve Bayes
  - SVM
  - Z-Score
- Corpus Description
- Evaluation and Discussion

14

Bayes' Rule

Thomas Bayes (1702-1761)



- Probability of event *H* given evidence *E:*

$$Prob[H|E] = \frac{Prob[E|H] \cdot Prob[H]}{Prob[E]}$$

- *A priori* probability of *H* : Prob[H]
  - Probability of event *before* evidence is seen
- *A posteriori* probability of *H* : Prob[H|E]
  - Probability of event *after* evidence is seen
- Combining prior probabilities and the likelihood of the data (according to the hypothesis *H*)

$$Prob[H|E] = \frac{Prob[E|H] \cdot Prob[H]}{Prob[E]} \propto Prob[E|H] \cdot Prob[H]$$

Naïve Bayes

- Text Categorization, we have
  - Evidence *E* = new document, sentence, instance
  - Event h_j = class value for this new instance
- The evidence can be divided into parts (i.e. the various features / terms E = {e_1, e_2, …e_n})
- Classify according to

$$h_{MAP} = \arg\max_{h_j \in H} Prob[h_j|e_1, e_2, \ldots e_n]$$

$$h_{MAP} = \arg\max_{h_j \in H} \frac{Prob[e_1, e_2, \ldots e_n|h_j] \cdot Prob[h_j]}{Prob[e_1, e_2, \ldots e_n]}$$

$$= \arg\max_{h_j \in H} Prob[e_1, e_2, \ldots e_n|h_j] \cdot Prob[h_j]$$

16

Naïve Bayes Classifier

- The computation of $Prob[e_1, e_2, \ldots e_n | h_j]$ is in a general case too complex
- The naïve Bayes classifier (conditionally independence)

$$Prob[e_1, e_2, \ldots e_n | h_j] \rightarrow \prod_{i=1}^{n} Prob[e_i | h_j]$$

and thus

$$h_{NB} = \arg\ \max_{h_j \in H}\ Prob[h_j] \cdot \prod_{i=1}^{n} Prob[e_i | h_j]$$

Text Classification (Learning)

1. Collect all words, punctuation that occur in the C (Corpus)
   V ← the set of all distinct words or tokens (selection?, stemming?)
2. Compute the probability estimate $P[h_j]$ and $P[e_k|h_j]$ as
   doc_j ← the subset of documents from C having the target value is h_j
   $P[h_j] = |doc_j| / |C|$ (reasonable prior estimation)
   $Text_j$ = concatenation of all members of doc_j
   n ← total number of words in $Text_j$
   for each word w_k in Voc
      n_k ← number of times word e_k occurs in $Text_j$
      $P[e_k|h_j] = (n_k+1) / (n + |Voc|)$ (better than direct n_k / n)

18

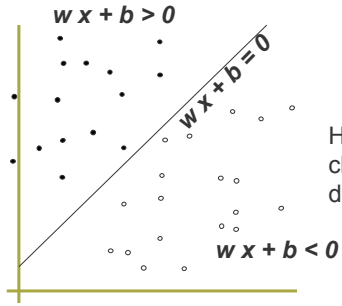
Naïve Bayes

- Pros:
  - Simple to implement
  - Fast and still pretty accurate in the performance
  - Easy to update with new data
  - Often used as a baseline to compare to other algorithms
- Cons:
  - Independence assumption which doesn`t always hold true!
  - Generative model: improving parameter estimates can hurt classification effectiveness
  - Be careful with the Prob[]

19

Support Vector Machine

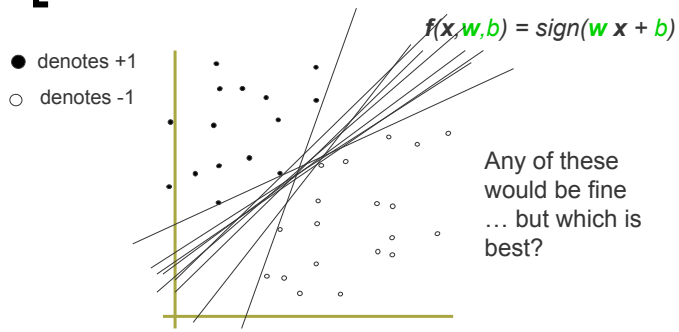
● denotes +1
○ denotes -1



How would you classify this data?

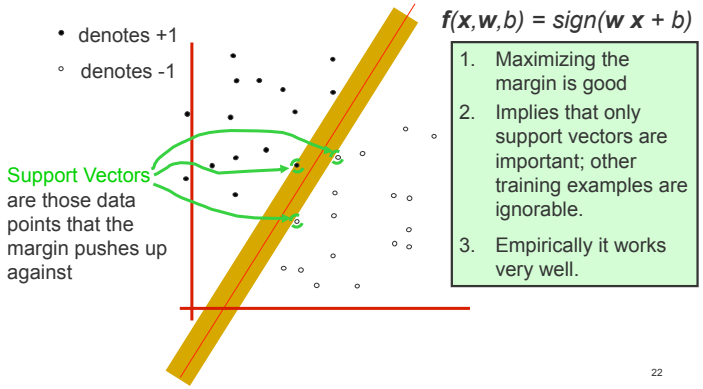
20

SVM: linear example



21

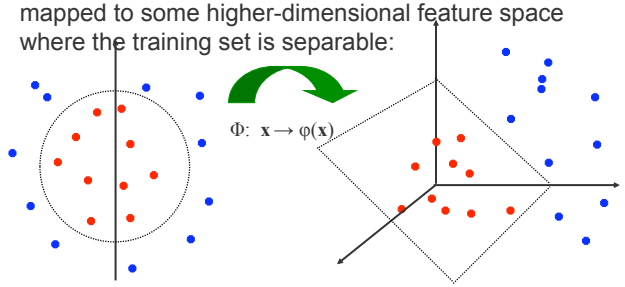
SVM: linear example



22

Non-linear SVM: Feature space

- What if the training sets are not separable?
- General idea: the original input space can always be mapped to some higher-dimensional feature space where the training set is separable:



23

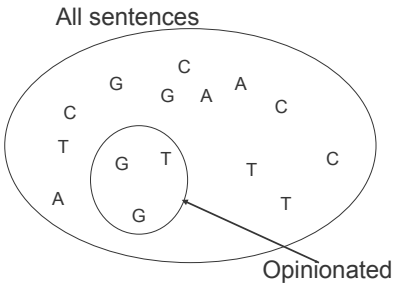
Support Vector Machine

- Pros:
  - Flexibility in choosing a similarity function
  - Sparseness of solution (only support vectors are used)
  - Ability to handle large feature spaces
  - Nice math property
- Cons:
  - It is sensitive to noise: a relatively small number of mislabeled examples can dramatically decrease the performance
  - It only considers two classes
  - Takes a long time in computation

24

Z score

Define the specific vocabulary of opinionated sentences



We have *n* the number of words in the opinionated sentences (*n* = 3)

Pr(ω): prob. that the word ω appears

e.g. Pr(G) = 4/16

We observe *a* = 2 time the word G in opinionated sent.

25

Z score

- In order to weigh the importance of a term in discriminating opinionated from factual sentence, we suggest computing the Z score

$$Z score(\omega) = \frac{a - n \cdot \Pr(\omega)}{\sqrt{n \cdot \Pr(\omega) \cdot (1 - \Pr(\omega))}}$$

- We assume that the word ω follows a binomial distribution with the parameters Pr(ω) and *n*.
- Pr(ω), where ω is a word in opinionated or factual corpora
- *n*: number of terms in opinionated documents
- *a*: the actual number (observed) of occurrences in the opinionated set

26

Z score

Terms having the largest Z score

	Opinionated		factural	
	Zscore	**term**	***Zscore***	**term**
1	7.17	should	5.96	year
2	4.95	we	4.03	last
3	4.71	must	3.94	billion
4	3.79	because	3.85	police
5	3.75	right	3.76	first
6	3.71	expressed	3.43	el
7	3.49	insisted	3.22	nino
8	3.48	need	3.21	Mainichi
9	3.47	I	3.19	Shimbun
10	3.43	believe	3.17	per

27

Z score

- Having a large positive Z score value (e.g., > 2) indicates that a term is over-used in opinionated sentences.
- The opposite is true for the large negative Z score value (e.g. < -2) (terms are under-used)
- The threshold is determined empirically but corrsponds to approx. 5% of the observations in a Normal distribution.

28

Z score

Procedure

- Represent a sentence as a set of words.
- Each word has a Z score for each category (opinionated, not opinionated)
- Sum of the Z Scores of terms over-used in opinionated class, and the sum of term under-used. Classify the sentence according to the max sum.

29

Outline

- Context, Examples and Applications
- Text Categorization
- Approaches used
  - Naïve Bayes
  - SVM
  - Z-Score
- **Corpus Description**
- Evaluation and Discussion

30

Corpus Description

- We use NTCIR6 and NTCIR7 corpus collections in English, traditional Chinese and Japanese languages
- The collections use newspaper articles from 1998 to 2001 with similar content in all languages
- For both collections we have:
  - 10 145 sentences
  - 24.6% opinionated
  - 75.4% not opinionated

31

Corpus Pre-processing

- Stemming: *cats* → *cat*
  (Harman S-stemmer)
- Extracting lemmas: given a word, obtain its` vocabulary definition

  said → *say*

Toutanova K., Klein D., Manning C. & Singer Y. « Feature-rich part-of-speech tagging with a cyclid dependency network », *Proceedings of HLT-NAACL 2003*, Edmonton, 27 May – 2 June 2003, p. 252-259.

- Stopword elimination
  - 41 words (the, is, of, and,…)
  - No meaning added

32

Pre-Processing

15 most frequent words in the collection

tf – term frequency in the collection

df – number of sentences with at least one occurrence of the term

The most frequent terms are not the solution!

	Opinionated Sentences			Not opinionated Sentences		
	tf	term	*df*	*tf*	term	*df*
1	536	said	529	772	said	754
2	422	not	398	646	not	609
3	290	he	254	552	he	487
4	201	we	169	423	japan	386
5	175	I	152	394	two	383
6	166	US	143	386	US	359
7	166	government	161	371	government	353
8	158	should	151	368	korean	314
9	153	more	139	354	korea	318
10	141	japan	126	329	other	315
11	139	world	132	329	after	323
12	138	chinese	119	325	more	311
13	133	korea	120	315	south	297
14	127	economic	123	311	economic	292
15	116	other	111	306	countr	292

33

Outline

- Context, Examples and Applications
- Text Categorization
- Approaches used
  - Naïve Bayes
  - SVM
  - Z-Score
- Corpus Description
- **Evaluation and Discussion**

34

Evaluation

- Effectiveness measure on *unseen* examples (train & test)
- Contingency table for each category c_i

TP: True positive
TN: True negative
FP: False positive
FN: False negative

Category c_i		Expert	
		Yes	No
Classifier decision	Yes	TP_i	FP_i
	No	FN_i	TN_i

- We can also create a global contingency table with all decisions (all documents)

35

Evaluation

- Precision (only the true) and Recall (all the truth)
- F_β measure (combining precision & recall)

with $F_1 = F = (2 \cdot P \cdot R) / (P+R)$

$$Prec = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

36

Evaluation

Model, parameter	Precision	Recall	$F_{(1)}$
Naïve Bayes, lemma	20.15 %	63.03 %	*30.49 %*
SVM, term, *tf idf*	33.63 %	65.76 %	44.37 %
SVM, lemma, *tf idf*	32.42 %	66.80 %	43.33 %
Z Score, term, λ, min: 4	44.23 %	82.72 %	56.30 %
Z Score, term, λ, min: 0	43.93 %	**84.49 %**	56.50 %
Z Score, term, min: 4	**45.54 %**	81.00 %	57.01 %
Z Score, term, min: 0	45.40 %	82.97 %	**57.34 %**
Z Score, lemma, λ, min: 0	39.19 %	83.80 %	53.23 %
Z Score , lemma, min: 0	41.46 %	79.91 %	54.36 %

- λ - smoothing parameter
- min:0 or min:4 – terms with frequencies less than 4 are eliminated

37

Failure Analysis

Opinionated sentence (mixed)
"Half of the job is psychiatry. "
With "psychiatry" (*tf* = 1, *hapax*)

NB: (0.179 / 0.821) → without opinion

half (3.23 / 5.91) job (2.61 / 2.13) psychiatry (-)

Z score: (0.0 / -1.83) → without opinion

half (-1.83) job (0.16) psychiatry (-)

38

Failure Analysis

Opinionated sentence (negative)
"You were often abused and humiliated "
with "humiliated " (*tf*=1, *hapax*)

NB: (0.397 / 0.603) → without opinion

you (12.65 / 7.7) often (4.17 / 3.39) abused (-)

Z score: (1.76 / 0) → with opinion

you (1.76) often (0.26) abused (-0.15) humiliated (-)

39

Results and Discussion

- Z Score clearly outperforms the other two baselines
- Classification of terms rather than lemmas gives better results
- Smoothing techniques have a tendency to improve recall but at the same time lower the precision
- Eliminating words with the frequency lower than 4 allows to significantly reduce the feature space with only a slight impact on the performance
- Number of sentences in the training corpus limits the possibility of the algorithm to improve the classification

40