A Model for Term Selection in Text Categorization Problems

Laura Maria Cannas, Nicoletta Dessì, Stefania Dessì
University of Cagliari
Italy

Text Categorization

- TC is the study of assigning natural language documents to one or more category labels

- TC is receiving a crescent interest because of the need to *automatically* organize the increasing number of digital documents available.

TC and high-dimensionality

- TC approaches have to face the problem of high-dimensionality

- *Term selection*: feature selection process that reduces the dimensionality of the feature space by only retaining the most informative or discriminative terms.

Feature Selection

- FS algorithms can be divided in two categories:
  - FILTER: fast, not so smart, which threshold?
  - WRAPPER: smarter, not so fast

Wrappers have been shown to generally perform better than filters, but their time-consuming behavior has made the use of filters prominent

Feature Selection

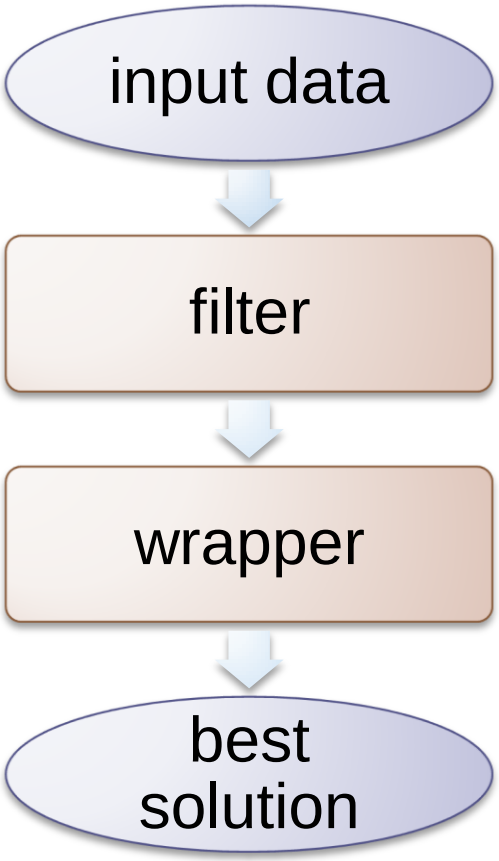
- HYBRID approaches
  - combine the use of filter and wrapper approaches to take advantage of the strengths of both while avoiding their drawbacks

A hybrid model for term selection to perform text categorization

The proposed model

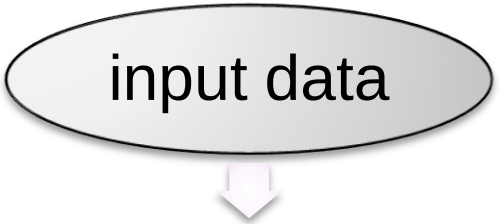
- Hybrid model for term selection
- Filter + wrapper
- Different search spaces
- Genetic Algorithm as search strategy within the wrapper

Genetic Wrapper Model (GWM)



The proposed model

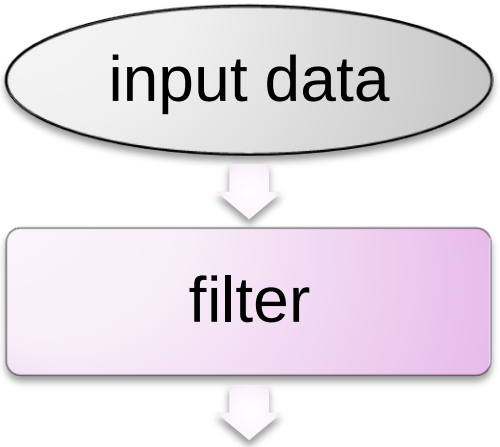
- GWM resolves *binary TC problems*
- It first selects the most representative terms for a given category c_i
- Then performs a binary classification process on this selection



GWM - input

- A matrix where each row represents a document d_j and columns are the related terms $\{w_1, w_2, \ldots, w_M\}$
- Each document is assigned to either the category c_i or its complement $\overline{c_i}$
- This is our training set

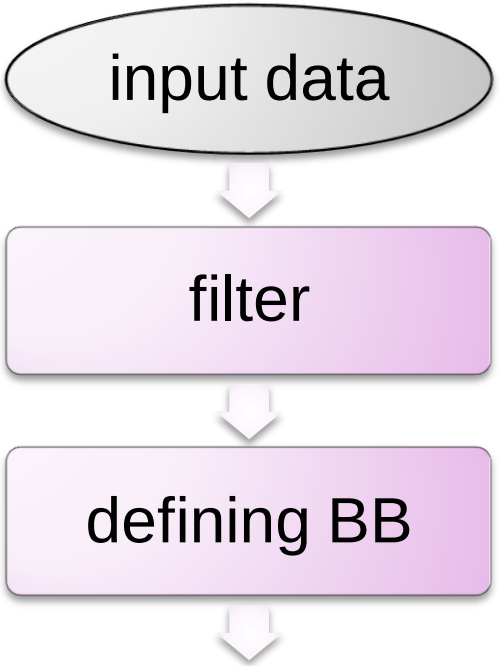
GWM - filter



- A filter assesses the scores of terms
- It outputs an ordered list where terms appear in descending order of relevance

The aim is to guide the term research at the initial stage and ensure that useful terms are unlikely to be discarded

GWM – search spaces



- The ordered list is cut using different threshold values
- Different term subsets of increasing size are constructed

Building Blocks (BBs)

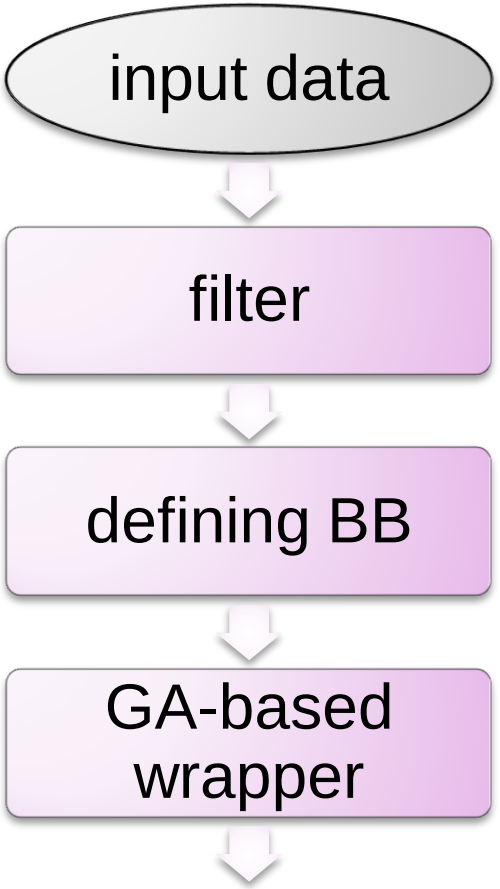
- It results in a sequence of Q nested BBs: $BB_1 \subset BB_2 \subset \ldots \subset BB_Q$

GWM – wrapper

- Each BB is refined by a wrapper that uses a GA as search strategy

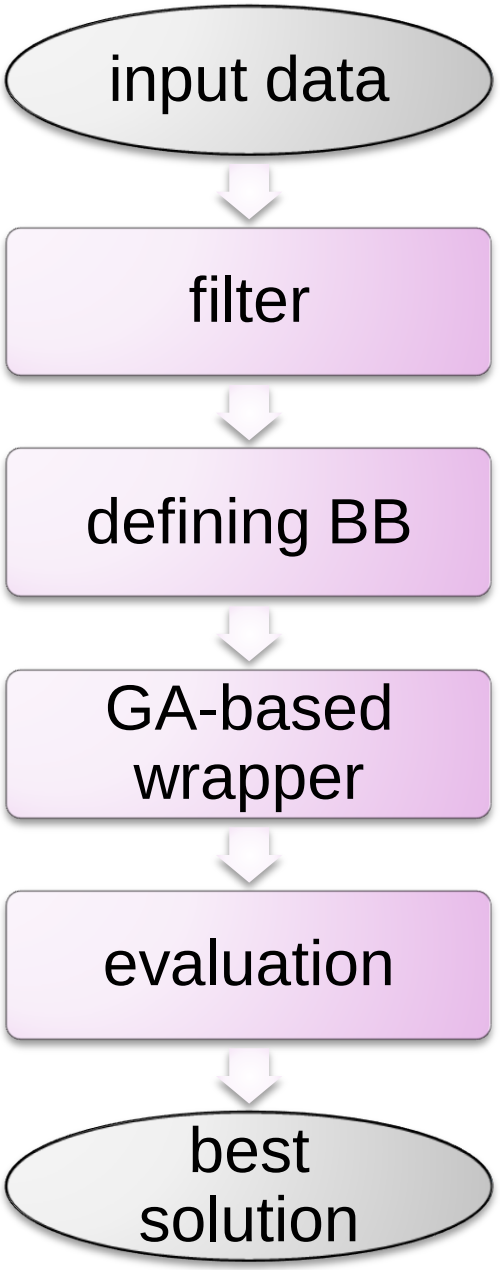
The aim is to is to remove redundant terms and obtain more accurate and small-sized subsets of terms for categorization

- As output we obtain the subset of terms that best categorizes the BB



GWM – evaluation

- Using a test set, solutions from each BB are evaluated and compared
- The best one is selected and returned by the GWM
- This solution is the subset of terms that best categorizes the given category c_i



GWM - settings

filter	Information Gain (IG) and χ^2 (CHI) as filter metrics
defining BB	BB10, BB20, …, BB200
GA-based wrapper	Fitness function: accuracy Naïve Bayes Multinomial classifier 3 runs
evaluation	F-measure, Break Even Point (BEP), and μ-BEP

Dataset

- Reuters-21578 test collection
  - 12,902 documents clustered in 135 categories
- Mod-Apté split
  - Training set: 9,603 docs
  - Test set: 3,299 docs

Category	No. of terms
acq	7,495
corn	8,302
crude	14,466
earn	9,500
grain	12,473
interest	10,458
money-fx	7,757
ship	9,930
trade	7,600
wheat	8,626

- We considered the 10 categories with the highest number of positive training examples (R10)

Experimental Results

- Results for a single category (grain)
- Results for all the categories (R10)
- Comparison with other approaches presented in literature

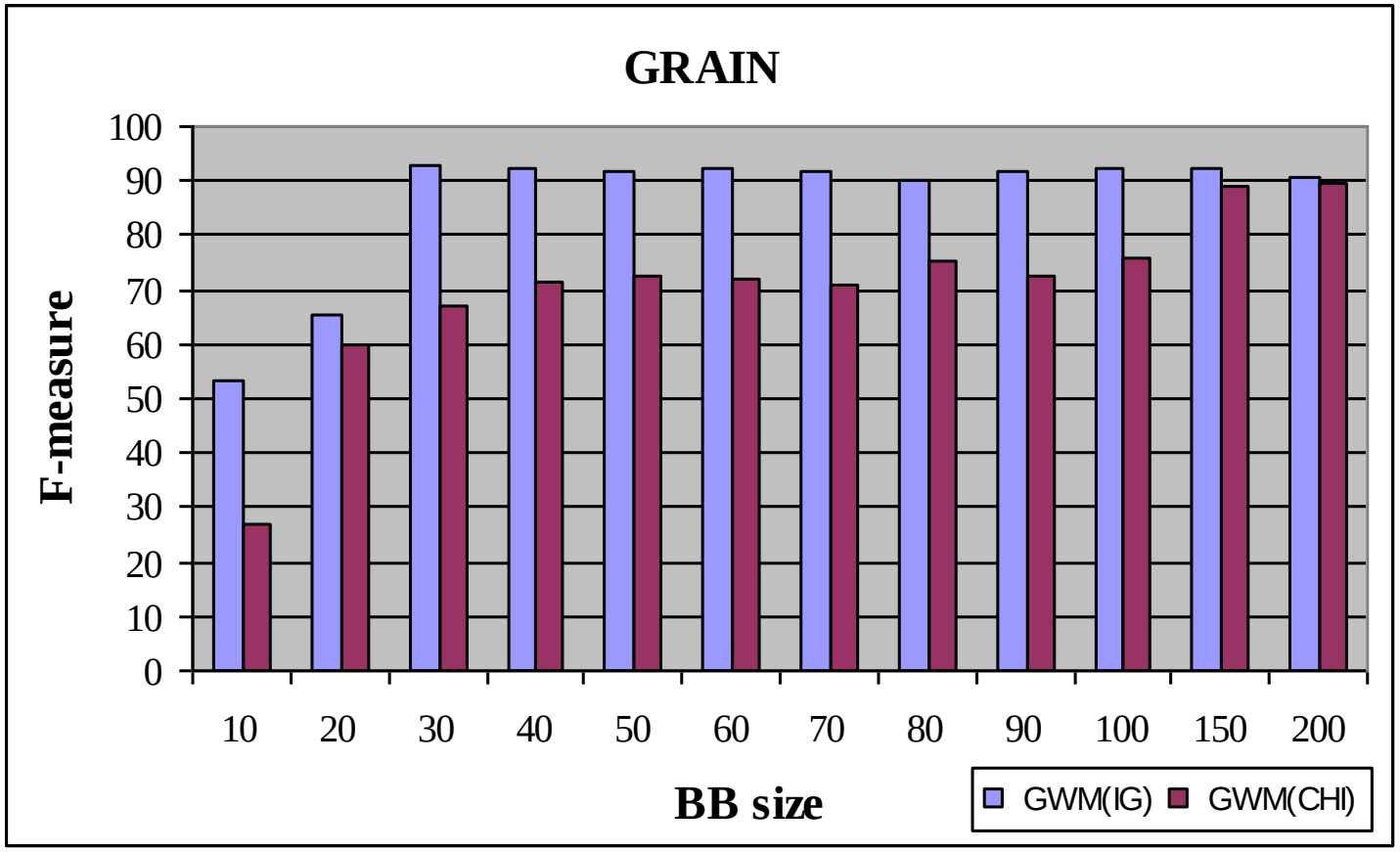
Experimental Results

- Avg and best values using GWM(IG) - (cat. grain)

	Average Values		Best Values	
BB size	F-measure	Selected Terms	F-measure	Selected Terms
10	53.16	9	53.16	9
20	65.48	18	65.48	18
30	92.28	12	92.78	13
40	91.59	15	92.45	14
50	91.16	16	91.56	17
60	90.77	19	92.26	17
70	90.30	24	91.66	21
80	89.03	24	90.36	24
90	89.61	30	91.61	29
100	90.09	27	92.26	19
150	89.70	46	92.26	36
200	89.16	63	90.37	58

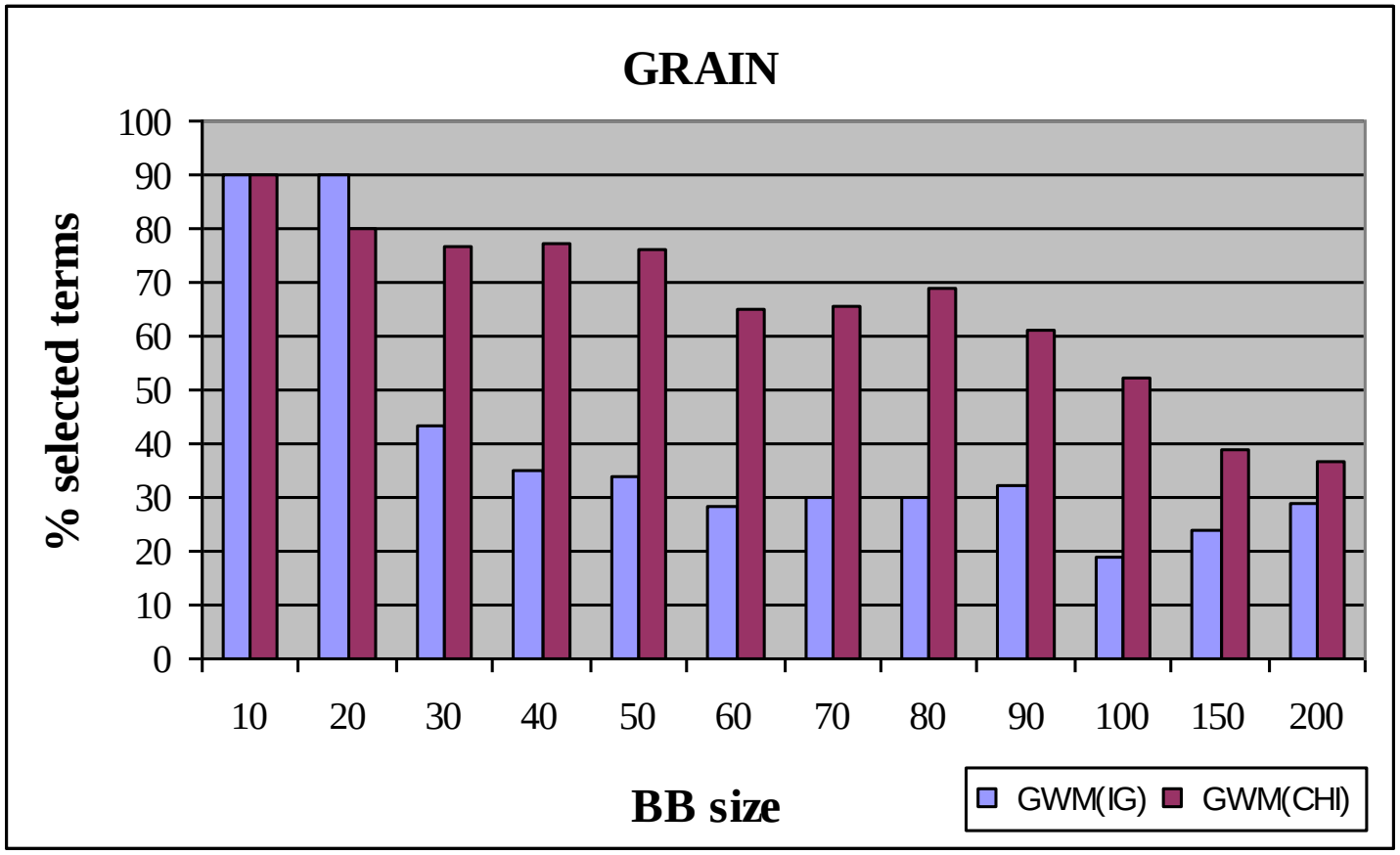
Experimental Results

- Best F-measure obtained within each BB (cat. grain)



Experimental Results

- Percentage of selected terms from each BB (cat. grain)



Experimental Results

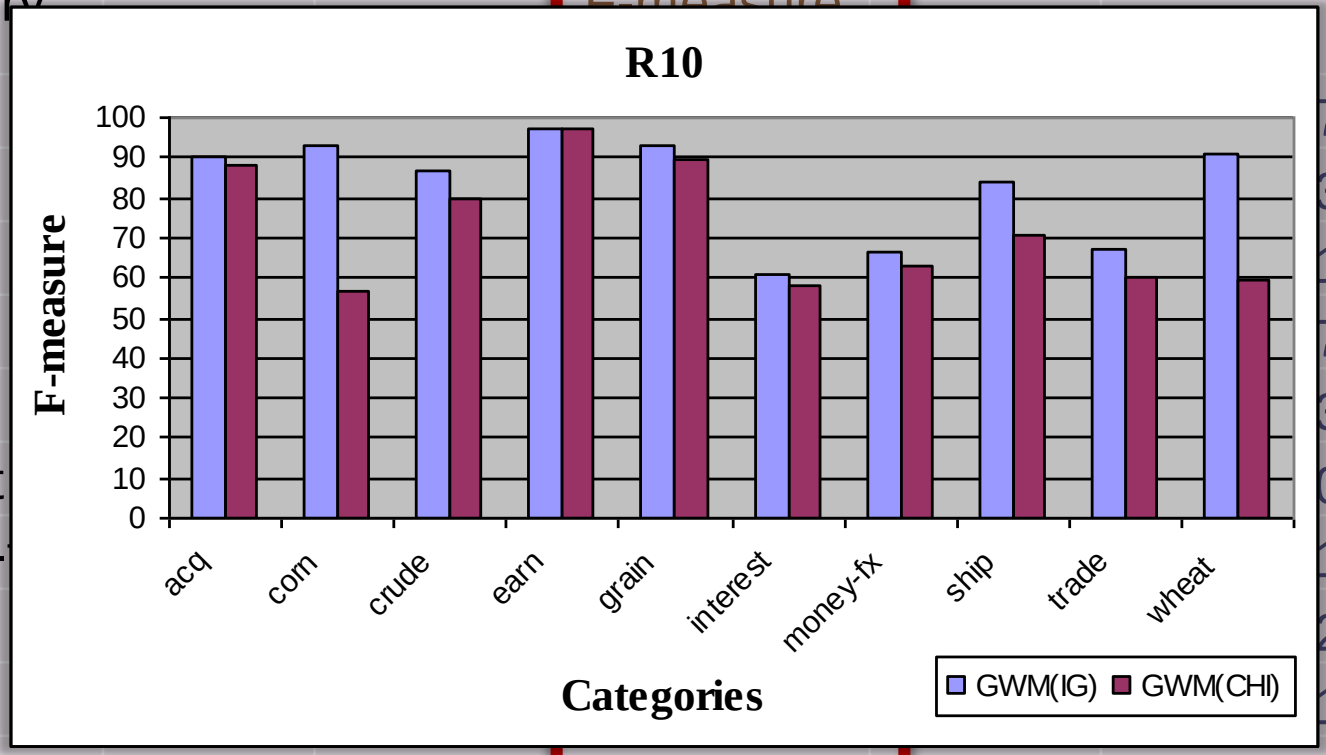
- Best F-measure value (R10)

	GWM(IG)			GWM(CHI)		
Category	BB size	Selected Terms	F-measure	BB size	Selected Terms	F-measure
acq	200	105	**90.36**	200	107	88.46
corn	150	30	**93.09**	200	123	56.52
crude	50	33	**86.52**	200	111	79.91
earn	150	73	96.90	200	97	**97.05**
grain	30	13	**92.79**	200	73	89.82
interest	90	34	**60.68**	200	110	58.29
money-fx	150	69	**66.51**	200	111	63.21
ship	90	47	**84.09**	200	122	70.74
trade	60	30	**67.29**	200	101	60.48
wheat	40	5	**90.81**	150	98	59.29

Experimental Results

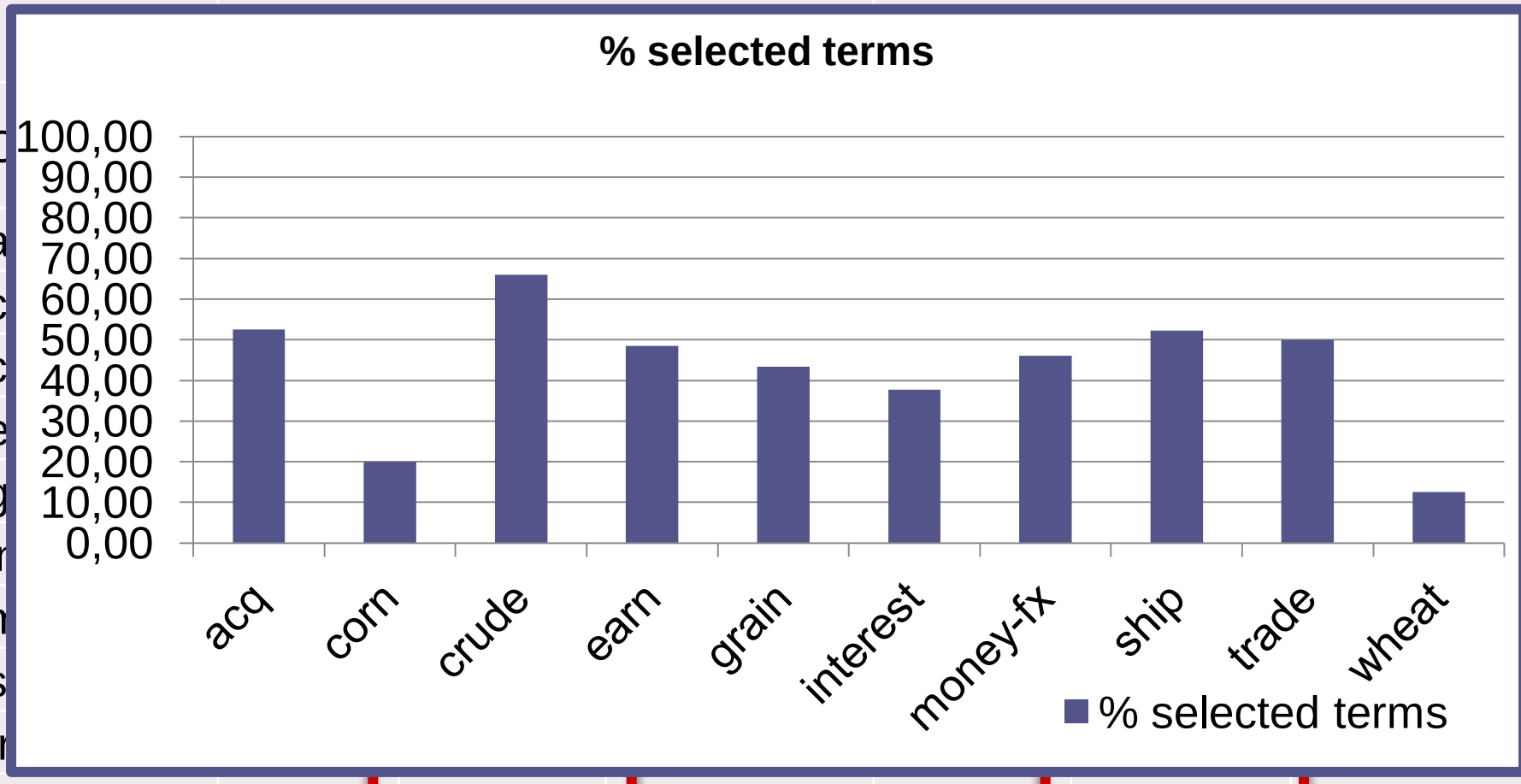
- Best F-measure value (R10)

	GWM(IG)			GWM(CHI)		
Category	BB	Selected	F-measure	BB	Selected	F-measure
acq					7	88.46
corn					3	56.52
crude					1	79.91
earn					7	**97.05**
grain					3	89.82
interest					0	58.29
money-fx					1	63.21
ship					2	70.74
trade					1	60.48
wheat	40	5	**90.81**	150	98	59.29



Experimental Results

- Best F-measure value (R10)



wheat	40	5	**90.81**	150	98	59.29

Experimental Results

- Best F-measure value(R10)

	GWM(IG)			GWM(CHI)		
Category	BB size	Selected Terms	F-measure	BB size	Selected Terms	F-measure
acq	200	105	**90.36**	200	107	88.46
corn	150	30	**93.09**	200	123	56.52
crude	50	33	**86.52**	200	111	79.91
earn	150	73	96.90	200	97	**97.05**
grain	30	13	**92.79**	200	73	89.82
interest	90	34	**60.68**	200	110	58.29
money-fx	150	69	**66.51**	200	111	63.21
ship	90	47	**84.09**	200	122	70.74
trade	60	30	**67.29**	200	101	60.48
wheat	40	5	**90.81**	150	98	59.29

Comparison (using BEP and μ–BEP values)

Category	Naïve Bayes	C4.5	Ripper	SVM		Olex		GWM
				Poly	rbf	Greedy	GA	
acq	90.29	85.59	86.63	90.37	90.83	84.32	87.49	90.40
corn	59.41	86.73	91.79	87.16	84.74	89.38	91.07	93.20
crude	78.84	82.43	81.07	87.82	86.17	80.84	77.18	86.85
earn	96.61	95.77	95.31	97.32	96.57	93.13	95.34	97.05
grain	77.82	89.69	89.93	92.47	88.94	91.28	91.75	92.85
interest	61.71	52.93	63.15	68.16	58.71	55.96	64.59	60.70
money-fx	56.67	63.08	62.94	72.89	68.22	68.01	66.66	66.95
ship	68.68	71.72	75.91	82.66	80.40	78.49	74.81	84.10
trade	57.90	70.04	75.82	77.77	74.14	64.28	61.81	67.70
wheat	71.77	91.46	90.66	86.13	89.25	91.46	89.86	91.20
μ-BEP	**82.52**	**85.82**	**86.71**	**89.91**	**88.80**	**84.80**	**86.40**	**89.06**

Conclusions

- We presented a hybrid model for term selection supporting TC problems
- An extensive validation has been carried out on the standard data collection Reuters
- Experimental results confirm the effectiveness of our model

Conclusions

- We presented a hybrid model for term selection supporting TC problems
  - GA-based learning approaches have remained isolated attempts
  - the hybrid approach combines effectiveness and efficiency

Future work

- Our proposal seems to offer several research perspectives:
  - the choice of the specific filter
  - the choice of the values for building the nested Building Blocks
- Validate the proposed model on other benchmarks for text analysis

Thanks for your attention

Questions?

Suggestions?