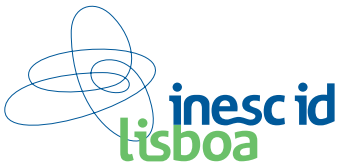
Combining LSI with other Classifiers to Improve Accuracy of Single-label Text Categorization

Ana Cardoso-Cachopo Arlindo Oliveira

Instituto Superior Técnico — Technical University of Lisbon / INESC-ID

EWLSATEL, March 2007



Outline

Outline

Outline

Outline

Outline

Outline

Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

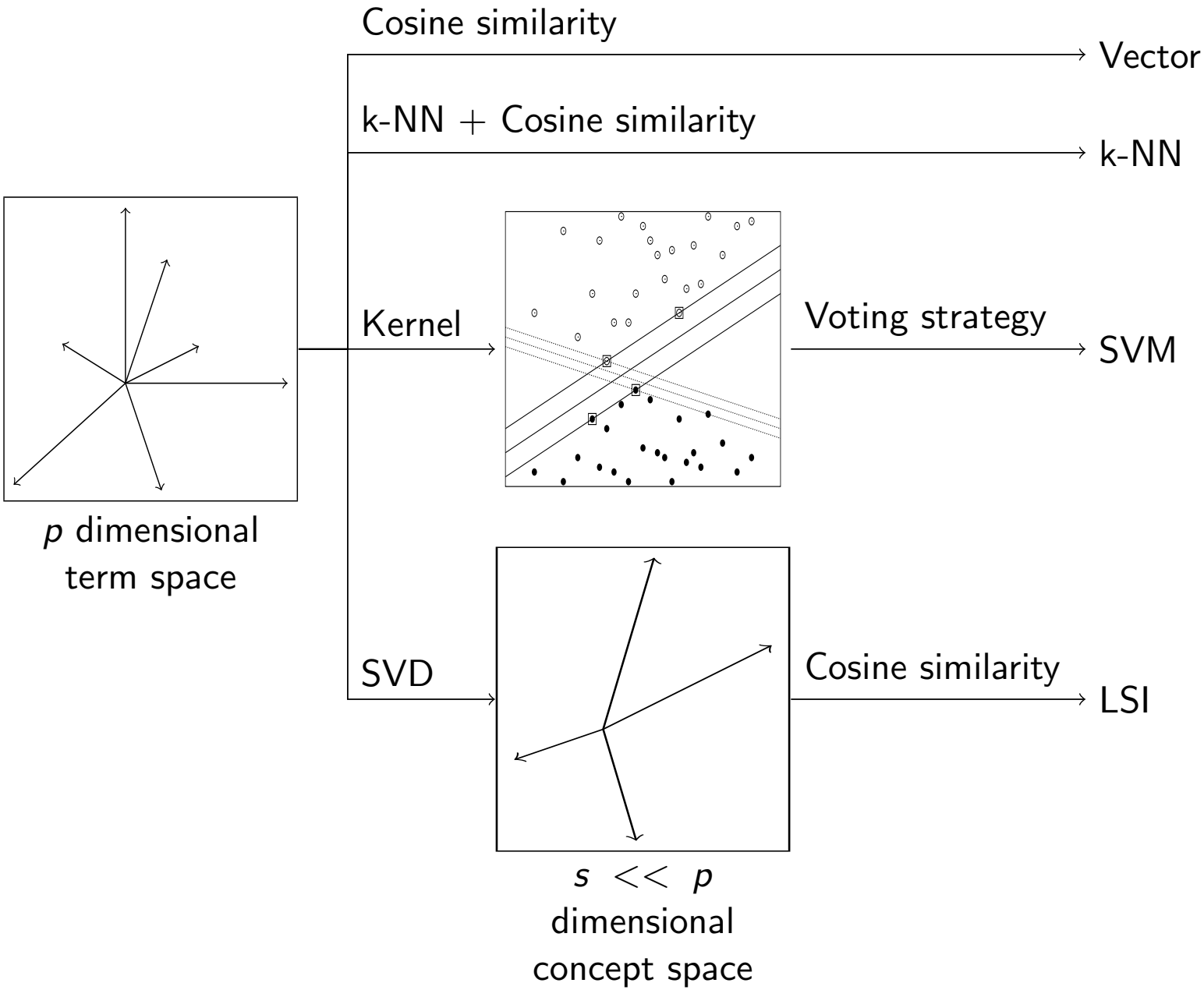
Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

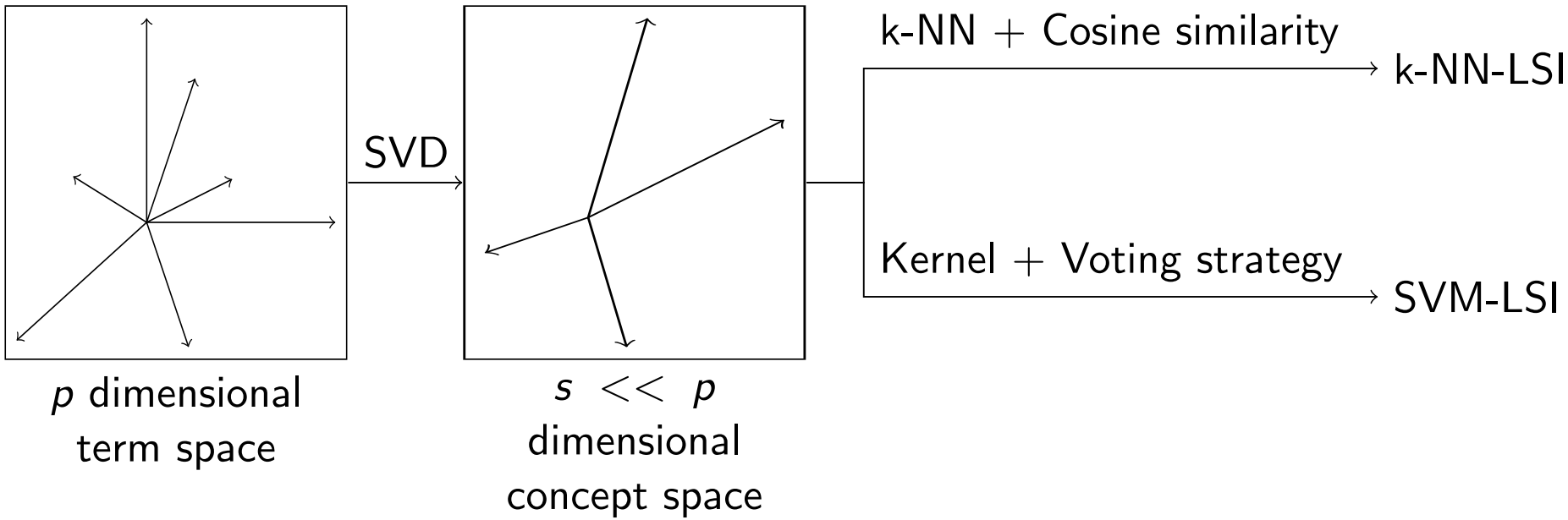
Introduction

- Text Classification
- Single-label
- Classification Methods
  - Vector
  - k-NN
  - SVM
  - LSI
- Goal: improve Accuracy

Classification Methods



Combinations Between Methods



Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank's Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - WebA
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank´s Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank´s Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank´s Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank´s Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank´s Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank´s Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Experimental Setup

- Methods (6 already mentioned + Dumb)
- Datasets
  - Bank´s Data - Bank37
  - Reuters 21578 - R8, R52
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12
- Evaluation Measure

$$Accuracy = \frac{\#\text{Correctly classified documents}}{\#\text{Total documents}}$$

Characteristics of the Datasets

	Train Docs	Test Docs	Total Docs	Smallest Class	Largest Class
Bank37	928	463	1391	5	346
20Ng	11293	7528	18821	628	999
R8	5485	2189	7674	51	3923
R52	6532	2568	9100	3	3923
Web4	2803	1396	4199	504	1641
Cade12	27322	13661	40983	625	8473

Numbers of documents for the datasets: number of training documents, number of test documents, total number of documents, number of documents in the smallest class, and number of documents in the largest class.

Characteristics of the Datasets

	Train Docs	Test Docs	Total Docs	Smallest Class	Largest Class
Bank37	928	463	1391	5	346
20Ng	11293	7528	18821	628	999
R8	5485	2189	7674	51	3923
R52	6532	2568	9100	3	3923
Web4	2803	1396	4199	504	1641
Cade12	27322	13661	40983	625	8473

Numbers of documents for the datasets: number of training documents, number of test documents, total number of documents, number of documents in the smallest class, and number of documents in the largest class.

Characteristics of the Datasets

	Train Docs	Test Docs	Total Docs	Smallest Class	Largest Class
Bank37	928	463	1391	5	346
20Ng	11293	7528	18821	628	999
R8	5485	2189	7674	51	3923
R52	6532	2568	9100	3	3923
Web4	2803	1396	4199	504	1641
Cade12	27322	13661	40983	625	8473

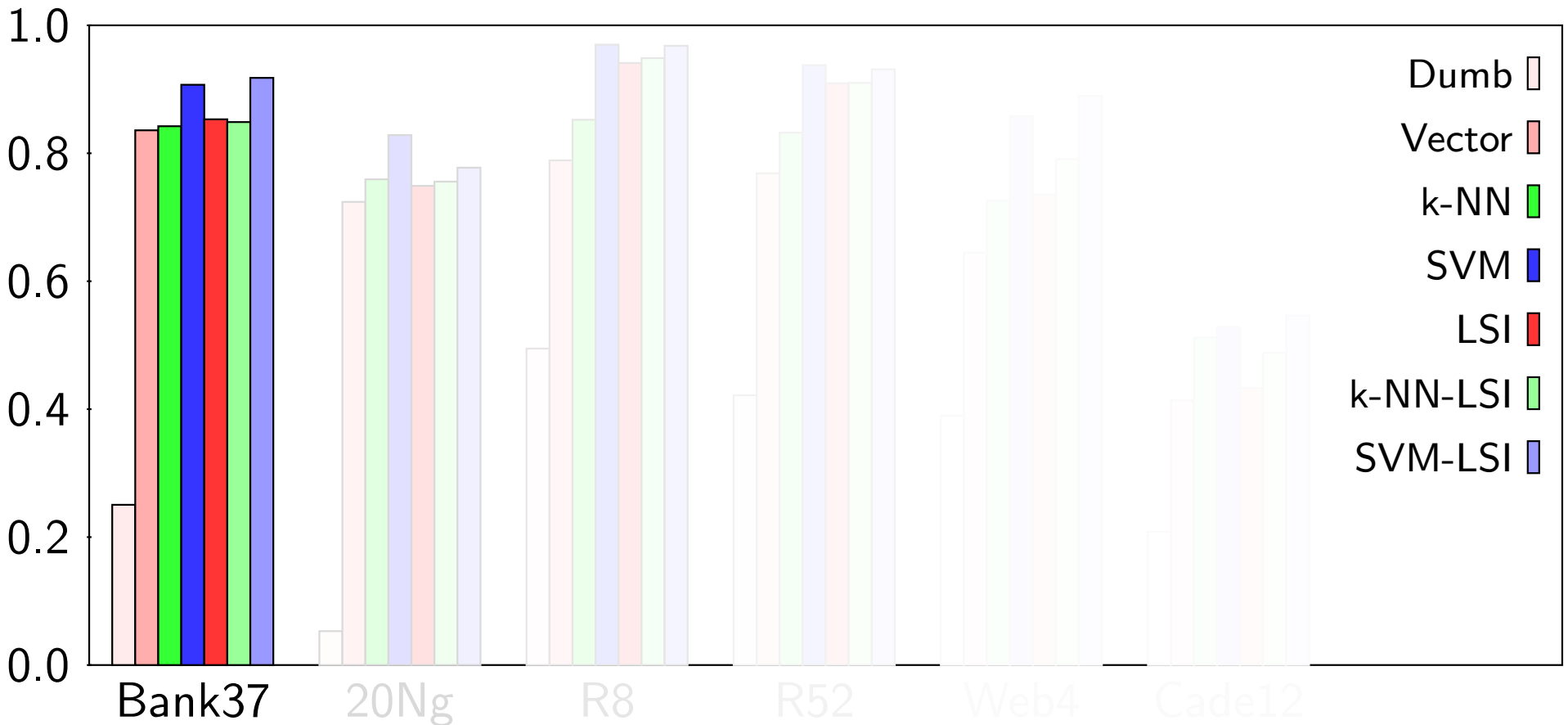
Numbers of documents for the datasets: number of training documents, number of test documents, total number of documents, number of documents in the smallest class, and number of documents in the largest class.

Characteristics of the Datasets

	Train Docs	Test Docs	Total Docs	Smallest Class	Largest Class
Bank37	928	463	1391	5	346
20Ng	11293	7528	18821	628	999
R8	5485	2189	7674	51	3923
R52	6532	2568	9100	3	3923
Web4	2803	1396	4199	504	1641
Cade12	27322	13661	40983	625	8473

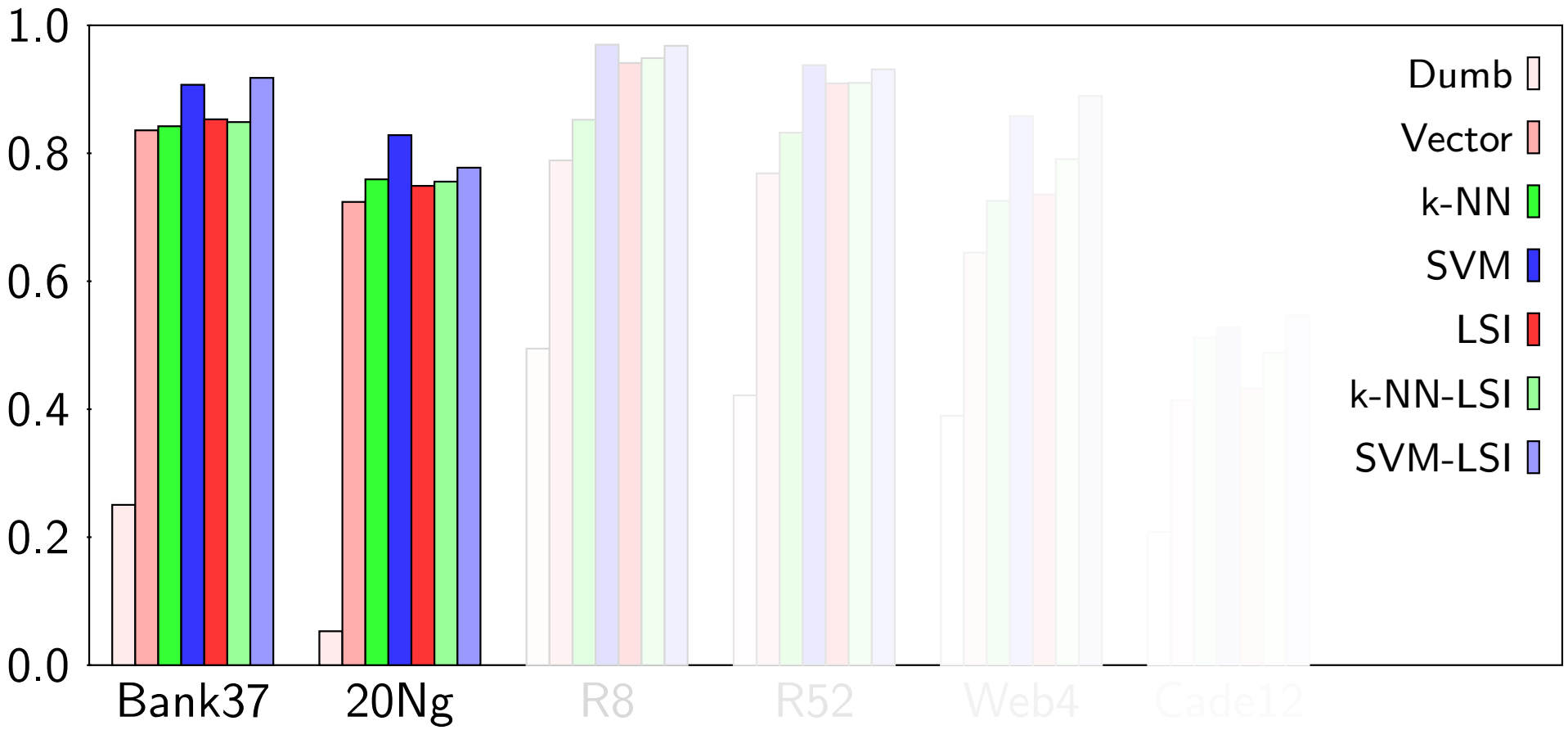
Numbers of documents for the datasets: number of training documents, number of test documents, total number of documents, number of documents in the smallest class, and number of documents in the largest class.

Experimental Results



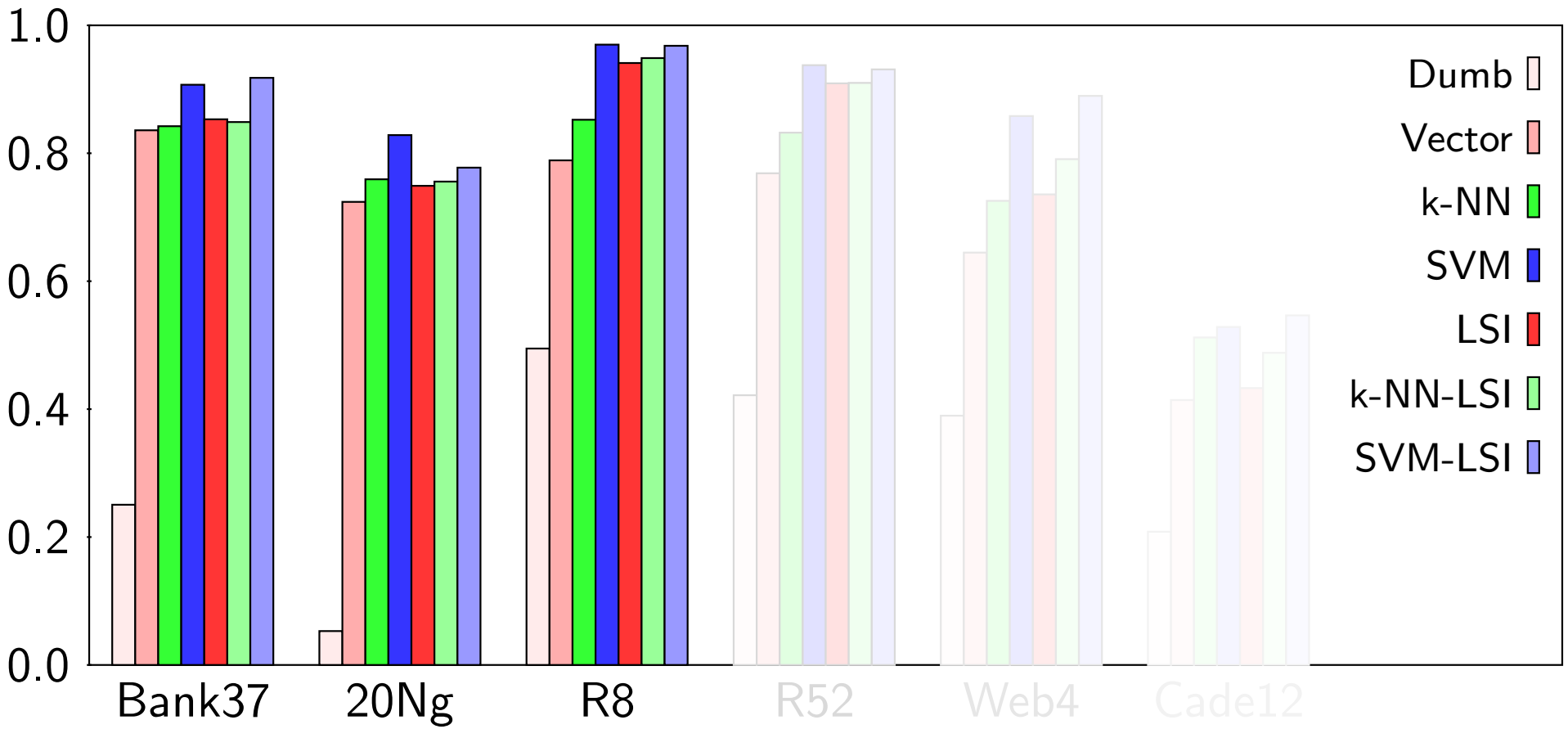
Accuracy values for the six datasets using each method.

Experimental Results



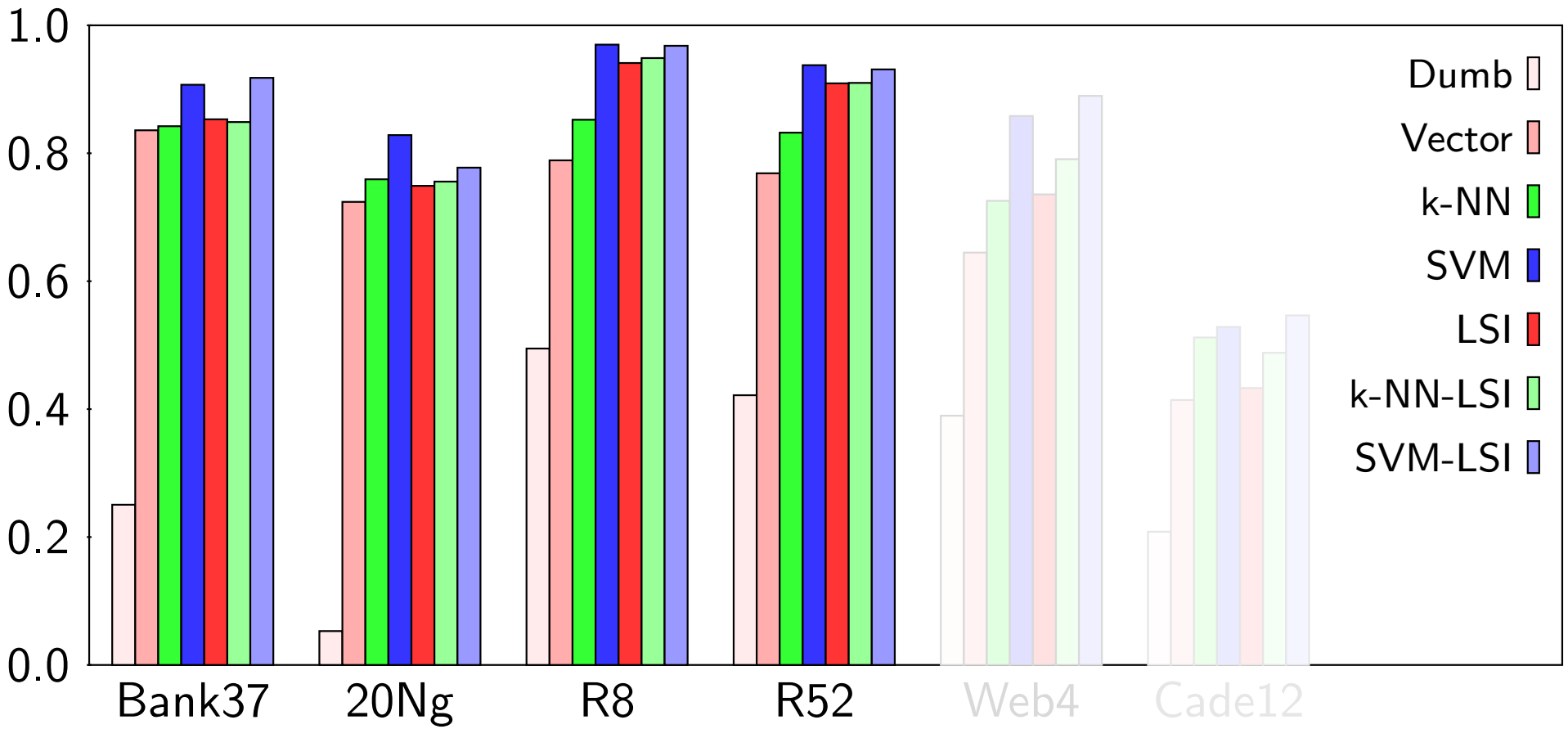
Accuracy values for the six datasets using each method.

Experimental Results



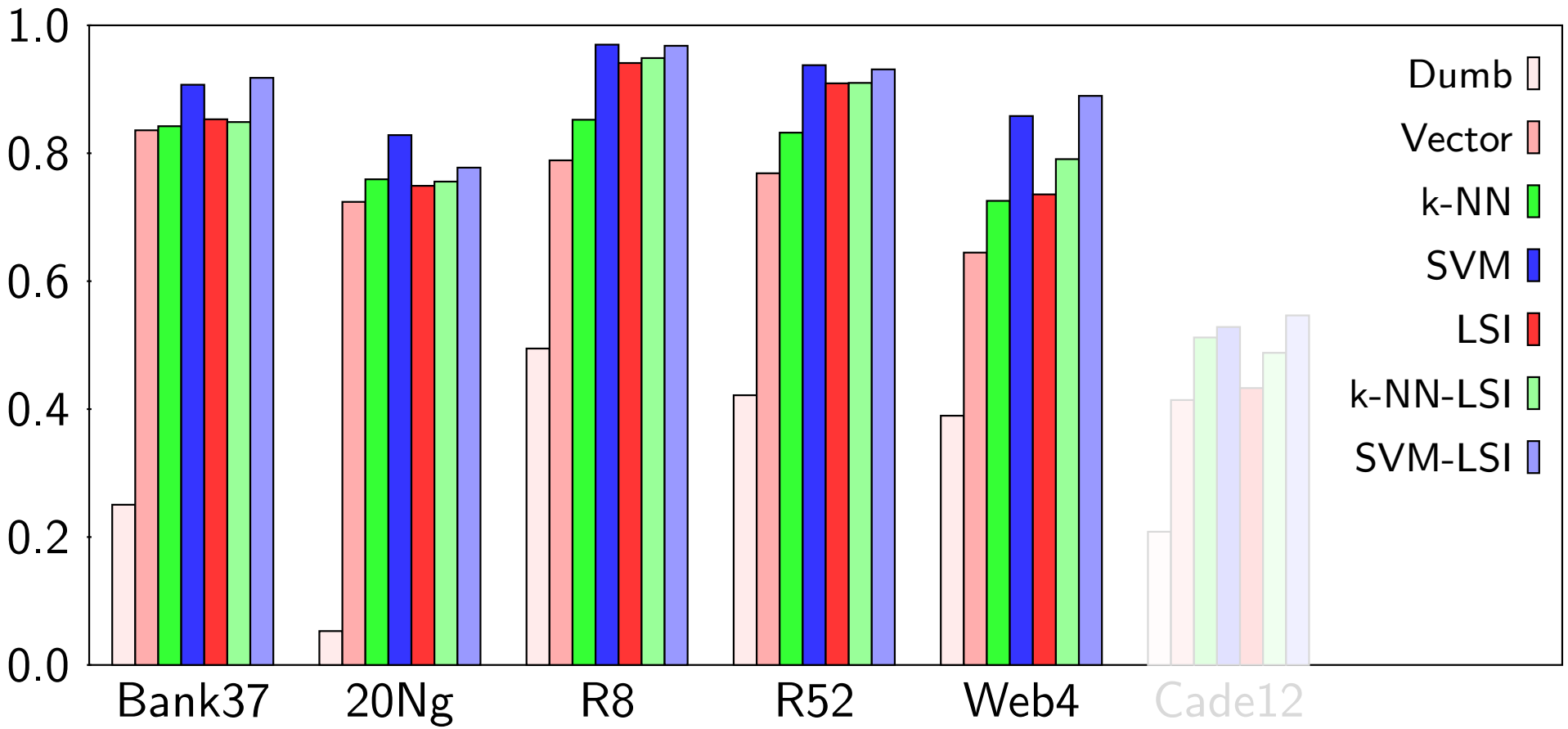
Accuracy values for the six datasets using each method.

Experimental Results



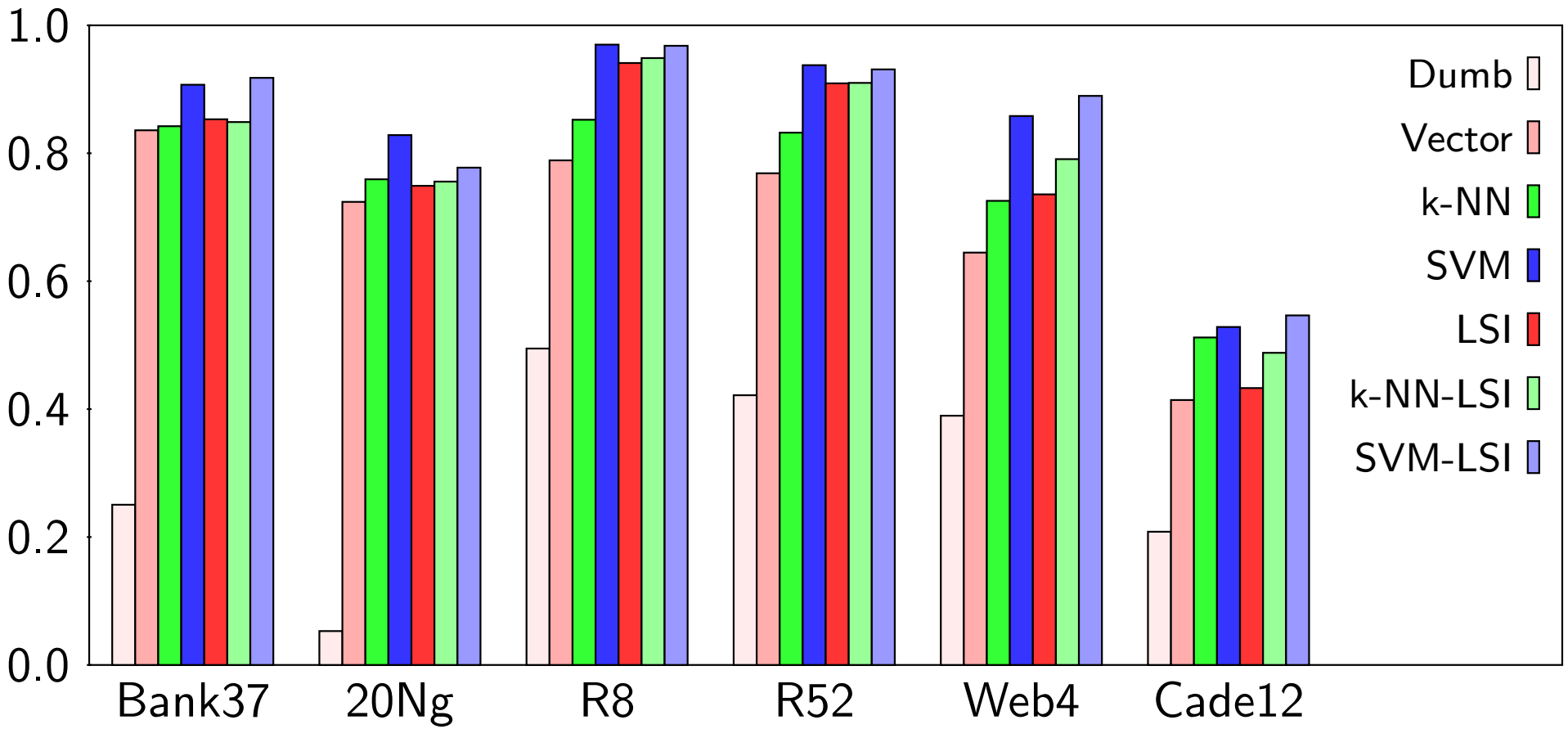
Accuracy values for the six datasets using each method.

Experimental Results



Accuracy values for the six datasets using each method.

Experimental Results



Accuracy values for the six datasets using each method.

Experimental Results

Dataset	Dumb	Vector	k-NN	SVM	LSI	k-NN LSI	SVM LSI
Bank37	0.2505	0.8359	0.8423	0.9071	0.8531	0.8488	0.9179
20Ng	0.0530	0.7240	0.7593	0.8284	0.7491	0.7557	0.7775
R8	0.4947	0.7889	0.8524	0.9698	0.9411	0.9488	0.9680
R52	0.4217	0.7687	0.8322	0.9377	0.9093	0.9100	0.9311
Web4	0.3897	0.6447	0.7256	0.8582	0.7357	0.7908	0.8897
Cade12	0.2083	0.4142	0.5120	0.5284	0.4329	0.4880	0.5465
Average	0.3030	0.6961	0.7540	0.8383	0.7702	0.7904	0.8385

Accuracy values for the six datasets using each method, and average Accuracy for each method over all the datasets.

Experimental Results

Dataset	Dumb	Vector	k-NN	SVM	LSI	k-NN LSI	SVM LSI
Bank37	0.2505	0.8359	0.8423	0.9071	0.8531	0.8488	0.9179
20Ng	0.0530	0.7240	0.7593	0.8284	0.7491	0.7557	0.7775
R8	0.4947	0.7889	0.8524	0.9698	0.9411	0.9488	0.9680
R52	0.4217	0.7687	0.8322	0.9377	0.9093	0.9100	0.9311
Web4	0.3897	0.6447	0.7256	0.8582	0.7357	0.7908	0.8897
Cade12	0.2083	0.4142	0.5120	0.5284	0.4329	0.4880	0.5465
Average	0.3030	0.6961	0.7540	0.8383	0.7702	0.7904	0.8385

Accuracy values for the six datasets using each method, and average Accuracy for each method over all the datasets.

Experimental Results

Dataset	Dumb	Vector	k-NN	SVM	LSI	k-NN LSI	SVM LSI
Bank37	0.2505	0.8359	0.8423	0.9071	0.8531	0.8488	0.9179
20Ng	0.0530	0.7240	0.7593	0.8284	0.7491	0.7557	0.7775
R8	0.4947	0.7889	0.8524	0.9698	0.9411	0.9488	0.9680
R52	0.4217	0.7687	0.8322	0.9377	0.9093	0.9100	0.9311
Web4	0.3897	0.6447	0.7256	0.8582	0.7357	0.7908	0.8897
Cade12	0.2083	0.4142	0.5120	0.5284	0.4329	0.4880	0.5465
Average	0.3030	0.6961	0.7540	0.8383	0.7702	0.7904	0.8385

Accuracy values for the six datasets using each method, and average Accuracy for each method over all the datasets.

Experimental Results

Dataset	Dumb	Vector	k-NN	SVM	LSI	k-NN LSI	SVM LSI
Bank37	0.2505	0.8359	0.8423	0.9071	0.8531	0.8488	0.9179
20Ng	0.0530	0.7240	0.7593	0.8284	0.7491	0.7557	0.7775
R8	0.4947	0.7889	0.8524	0.9698	0.9411	0.9488	0.9680
R52	0.4217	0.7687	0.8322	0.9377	0.9093	0.9100	0.9311
Web4	0.3897	0.6447	0.7256	0.8582	0.7357	0.7908	0.8897
Cade12	0.2083	0.4142	0.5120	0.5284	0.4329	0.4880	0.5465
Average	0.3030	0.6961	0.7540	0.8383	0.7702	0.7904	0.8385

Accuracy values for the six datasets using each method, and average Accuracy for each method over all the datasets.

Experimental Results

Dataset	Dumb	Vector	k-NN	SVM	LSI	k-NN LSI	SVM LSI
Bank37	0.2505	0.8359	0.8423	0.9071	0.8531	0.8488	0.9179
20Ng	0.0530	0.7240	0.7593	0.8284	0.7491	0.7557	0.7775
R8	0.4947	0.7889	0.8524	0.9698	0.9411	0.9488	0.9680
R52	0.4217	0.7687	0.8322	0.9377	0.9093	0.9100	0.9311
Web4	0.3897	0.6447	0.7256	0.8582	0.7357	0.7908	0.8897
Cade12	0.2083	0.4142	0.5120	0.5284	0.4329	0.4880	0.5465
Average	0.3030	0.6961	0.7540	0.8383	0.7702	0.7904	0.8385

Accuracy values for the six datasets using each method, and average Accuracy for each method over all the datasets.

Conclusions and Future Work

- Very good Accuracy for some datasets.
- It is worth pursuing this line of research by testing more combinations between the method´s parameters.

Conclusions and Future Work

- Very good Accuracy for some datasets.
- It is worth pursuing this line of research by testing more combinations between the method´s parameters.

Thank You.

Any Questions?