On the Use of PU Learning for Quality Flaw Prediction in Wikipedia

Edgardo Ferretti, Donato Hernández, Rafael Guzmán, Manuel Montes, Marcelo Errecalde & Paolo Rosso

September 19th, PAN@CLEF'12, Rome

Who are we?

Edgardo Ferretti

Marcelo Errecalde

Paolo Rosso

Donato Hernández

Donato Hernández

Rafael Guzmán

Manuel Montes

Methodological Design

- Using a state-of-the-art document model
- Finding a good algorithm for classification tasks
  - Exploiting the characteristics of this algorithm

Methodological Design

- Using a state-of-the-art document model
  - 73 features from the document model used in [1]. They were selected following the guidelines in [2].

Text Features

LENGTH: character / sentence / word count, etc.
STRUCTURE: mandatory sections count, tables count, etc.
STYLE: prepositions / stop words / questions rate, etc.
READABILITY: Gunning-Fog / Kincaid indexes, etc,

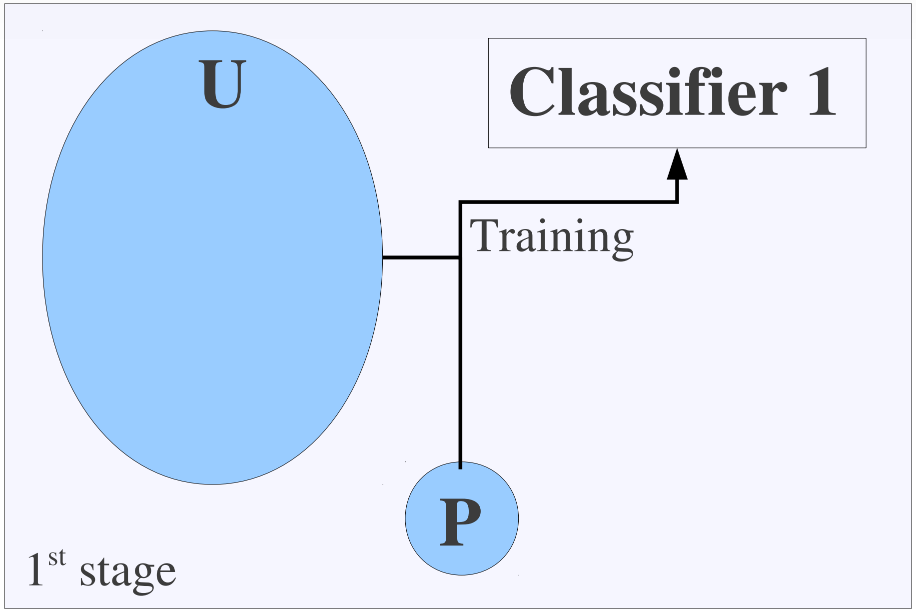
Network Features

In-link count,
Internal link count,
Inter-language link count

[1] Anderka, M., Stein, B., Lipka, N.: Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In: 35rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2012)
[2] Dalip, D., Goncalves, M., Cristo, M., Calado, P.: Automatic quality assessment of content created collaboratively by Web communities: a case study of Wikipedia. In: 9th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM (2009).

PU Learning

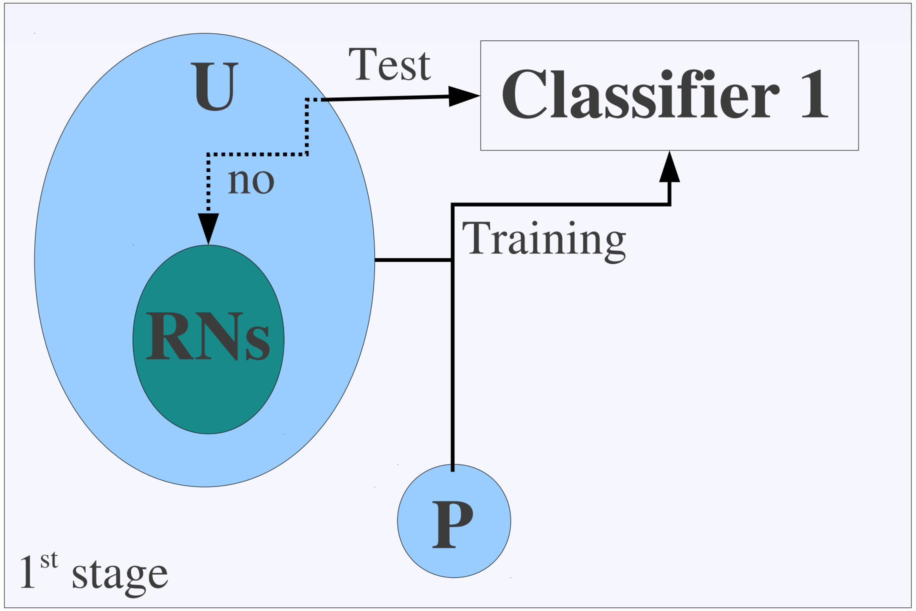
- This method uses as input a small labelled set of the positive class to be predicted and a large unlabelled set to help learning.[3]



[3] Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.

PU Learning

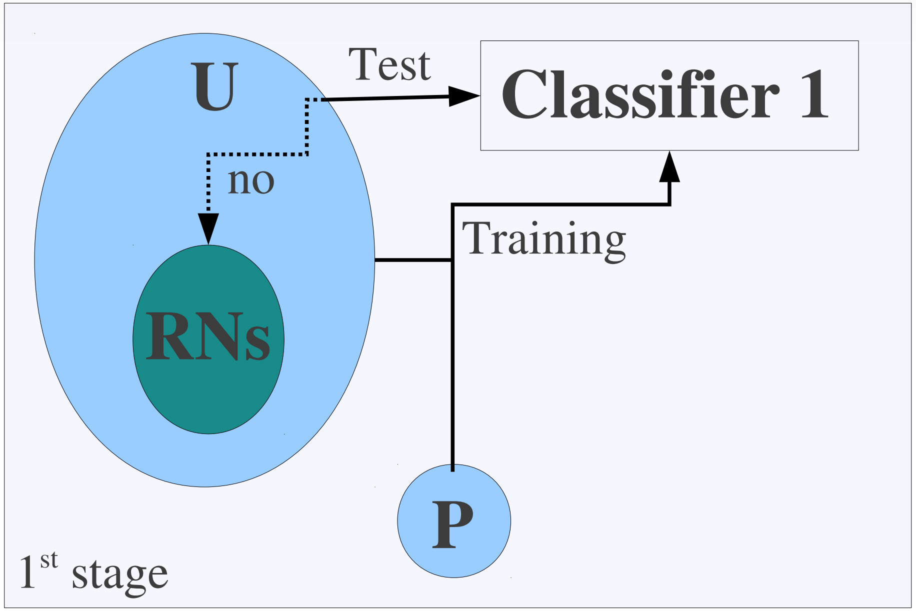
- This method uses as input a small labelled set of the positive class to be predicted and a large unlabelled set to help learning.[3]

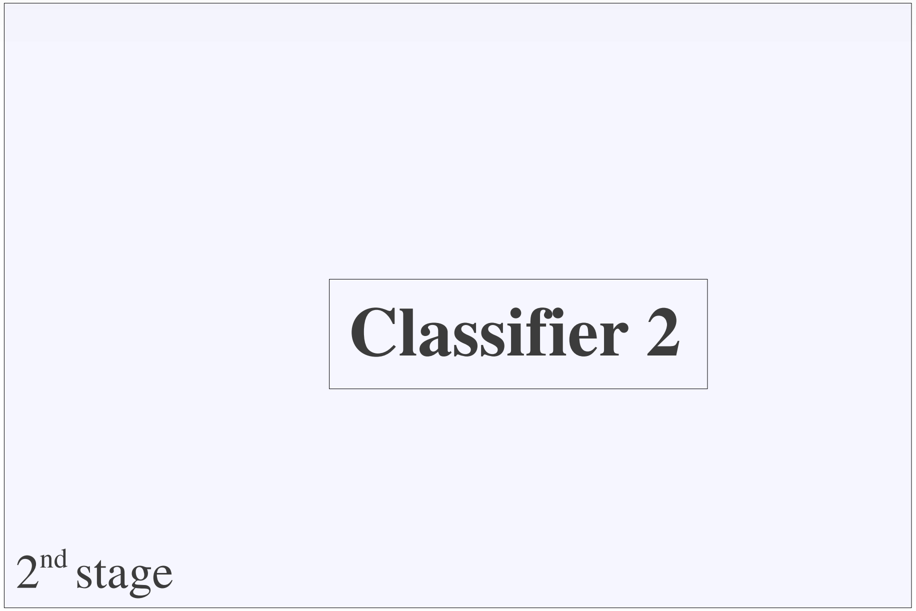


[3] Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.

PU Learning

- This method uses as input a small labelled set of the positive class to be predicted and a large unlabelled set to help learning.[3]

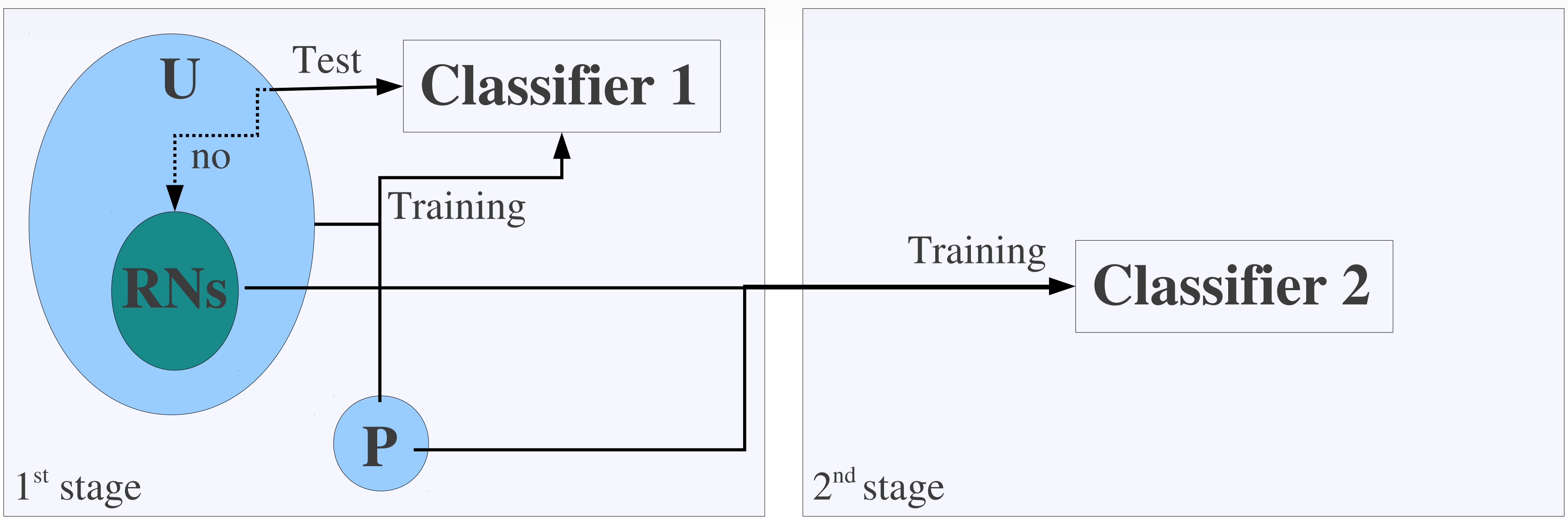




[3] Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.

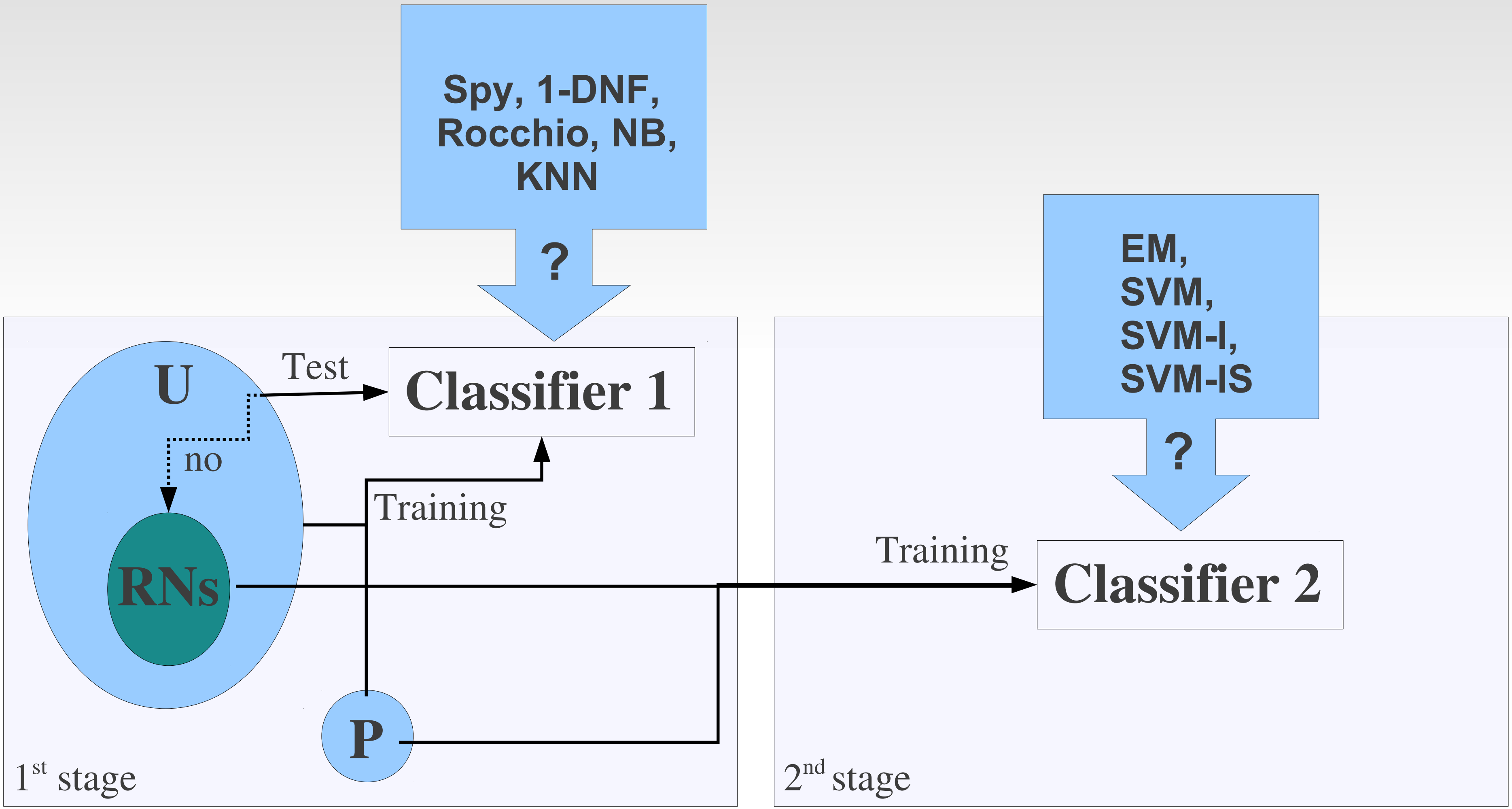
PU Learning

- This method uses as input a small labelled set of the positive class to be predicted and a large unlabelled set to help learning.[3]

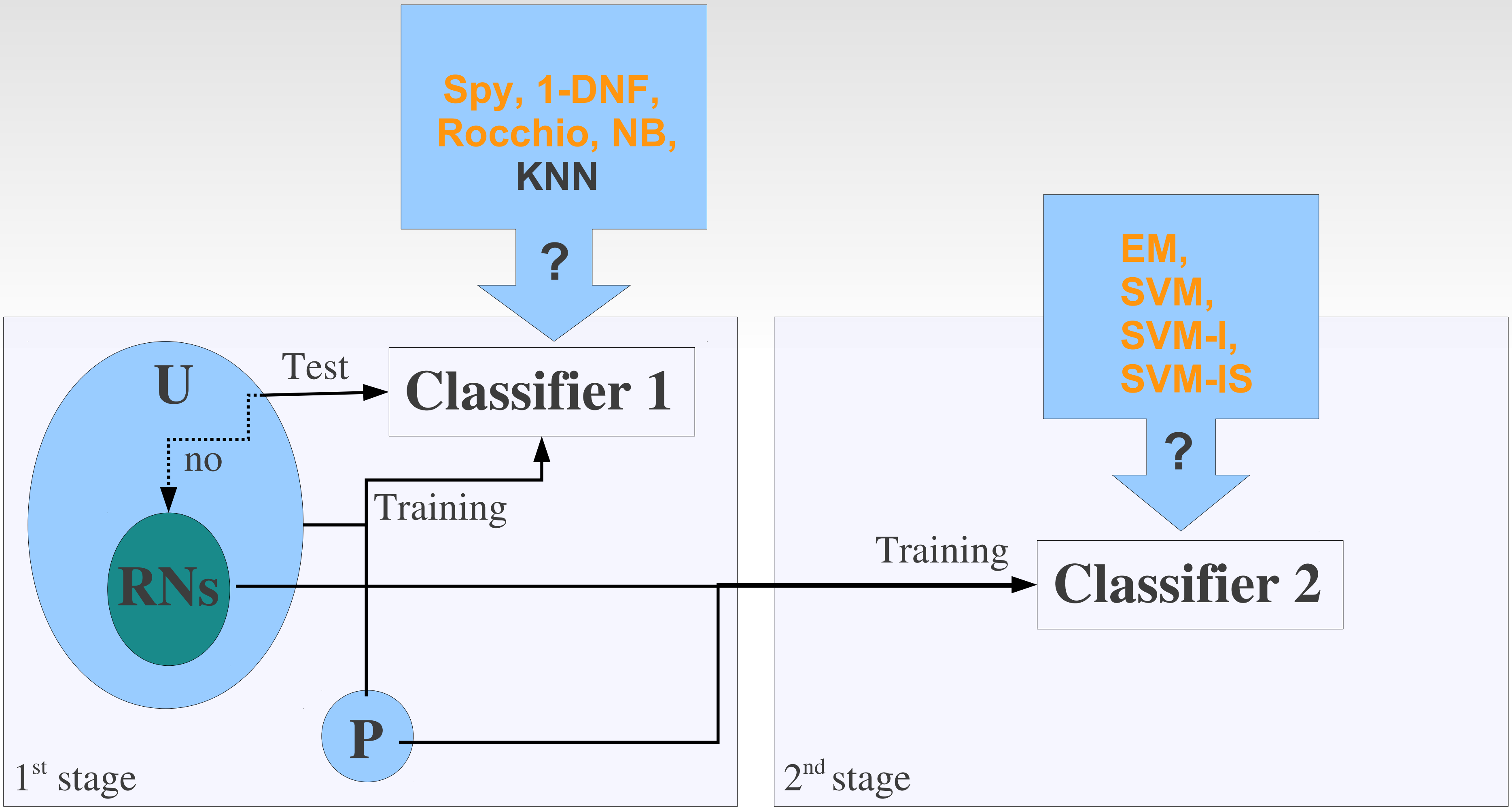


[3] Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.

What classifier in each stage?

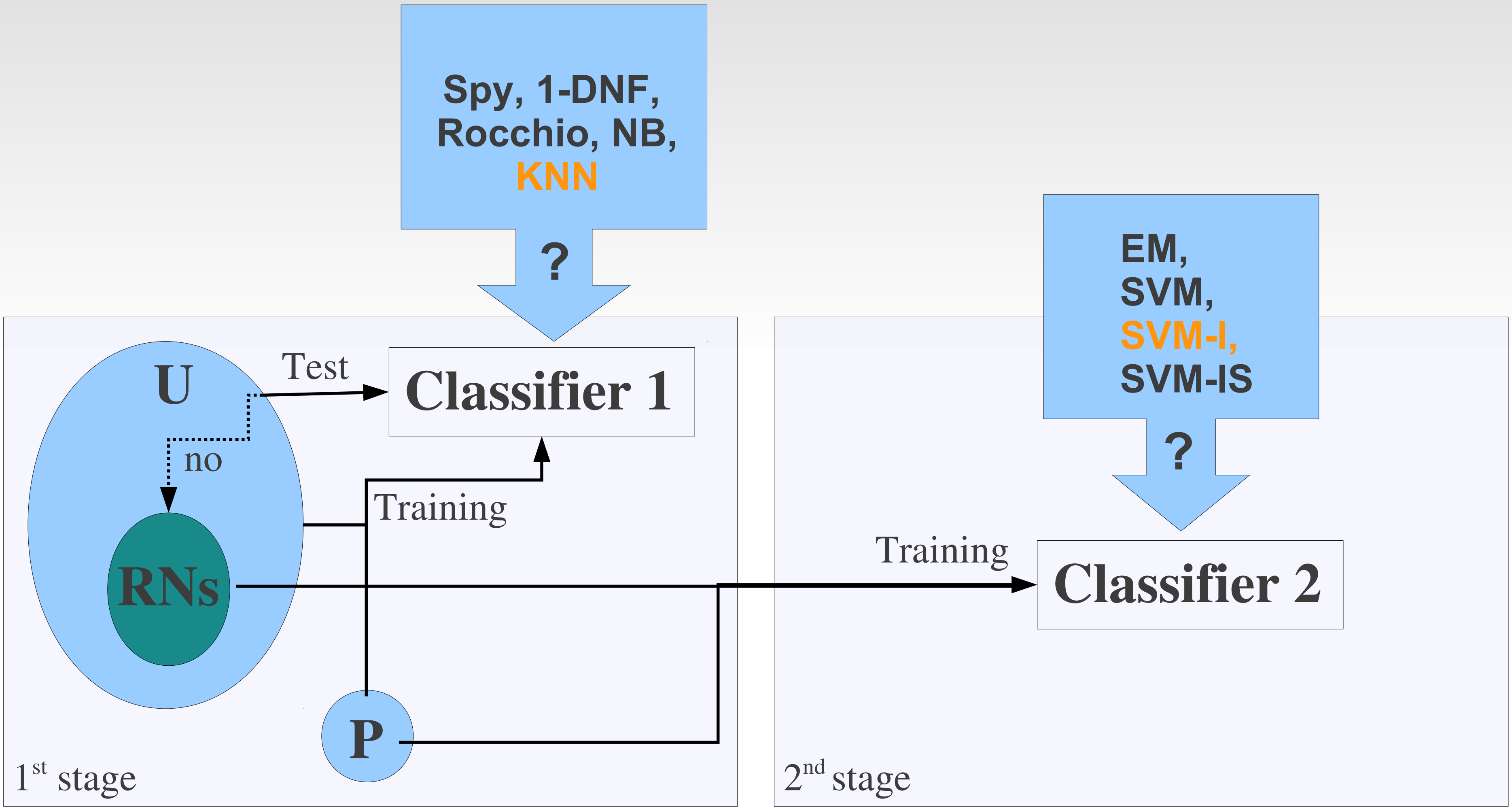


What classifier in each stage?



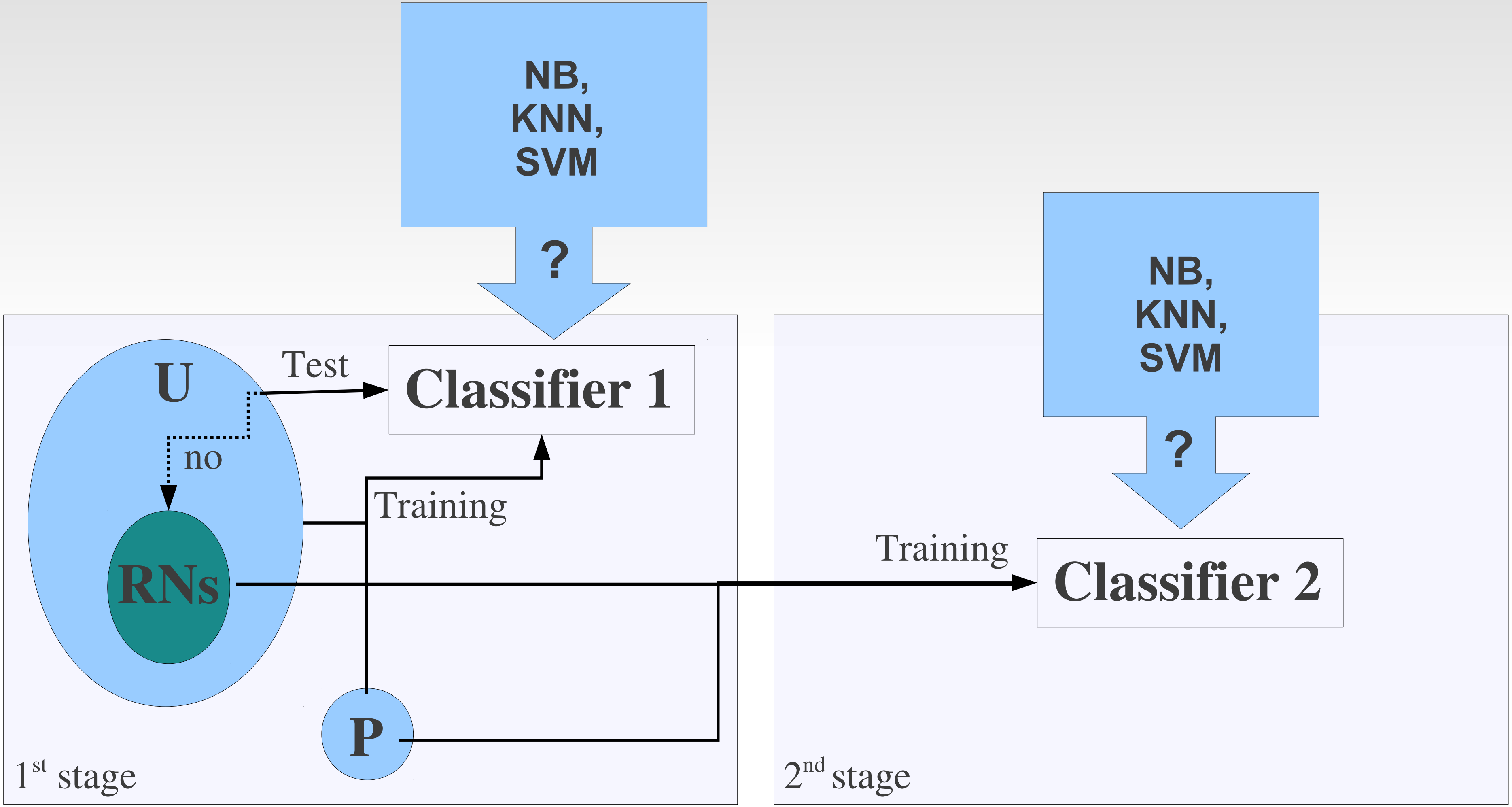
Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.

What classifier in each stage?



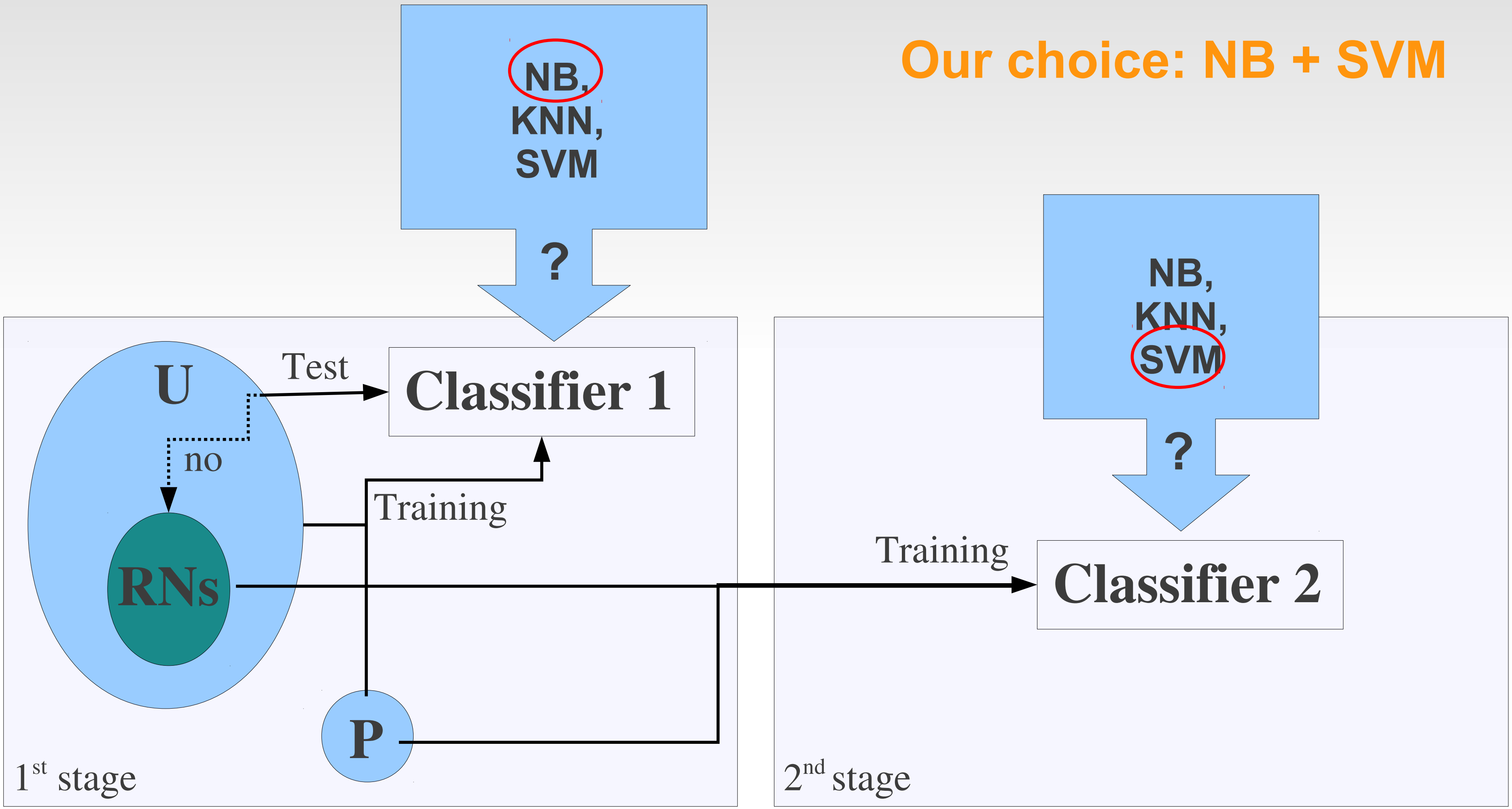
B. Zhang and W. Zuo. Reliable Negative Extracting Based on kNN for Learning from Positive and Unlabeled Examples . Journal of Computers, 4(1):94–101, 2009.

What classifier in each stage?



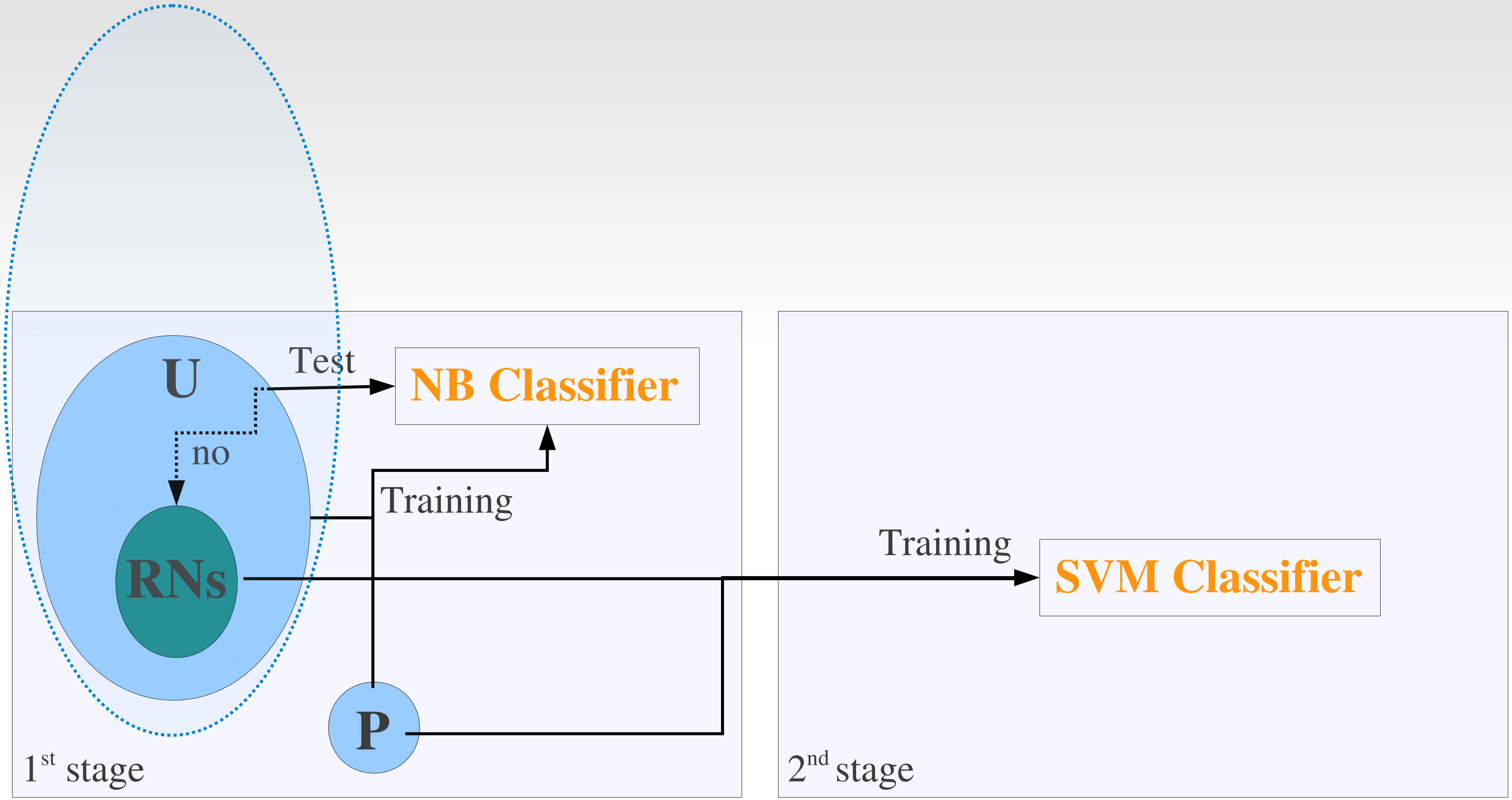
What classifier in each stage?

Our choice: NB + SVM

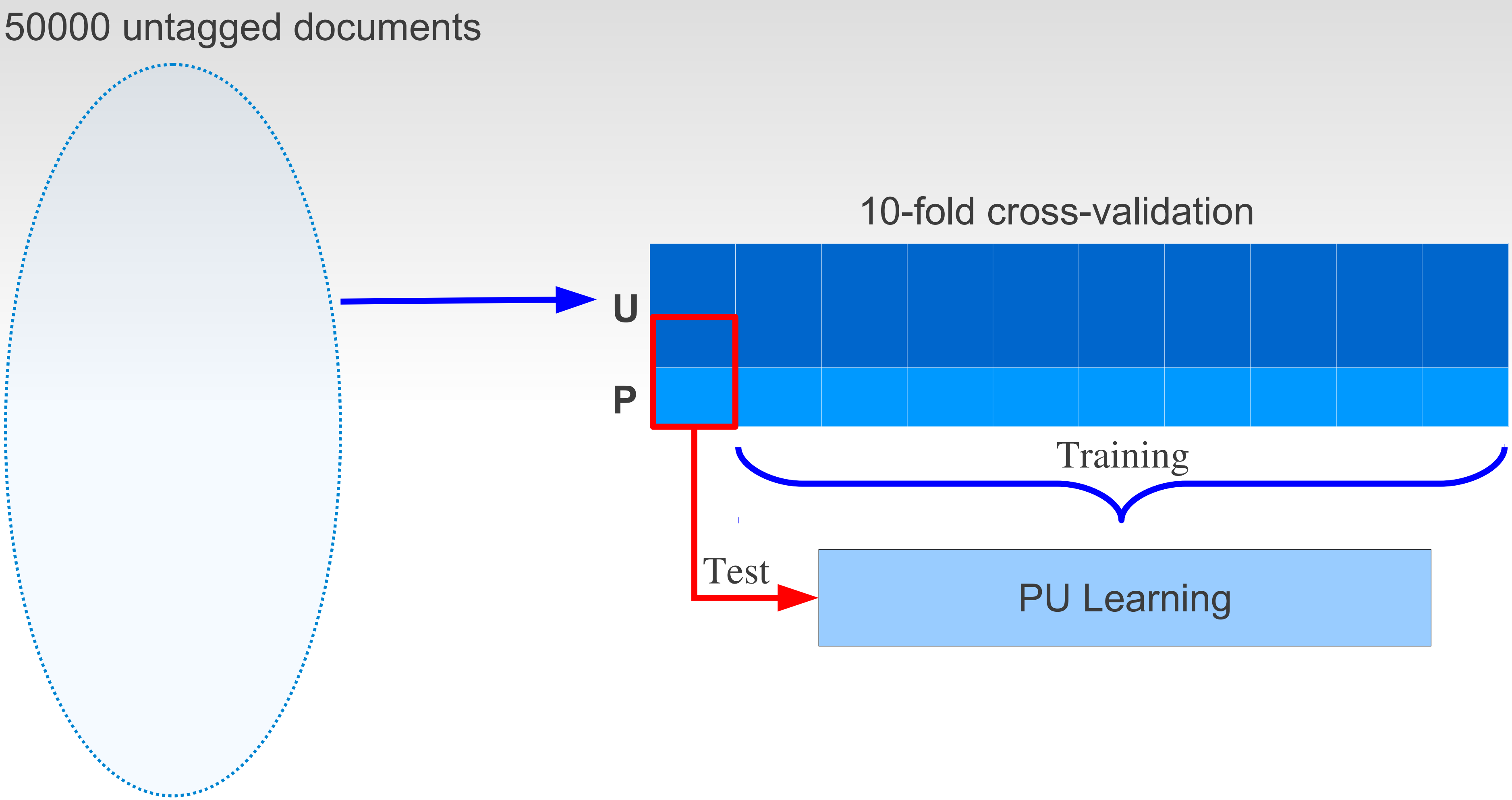


Untagged sampling strategy

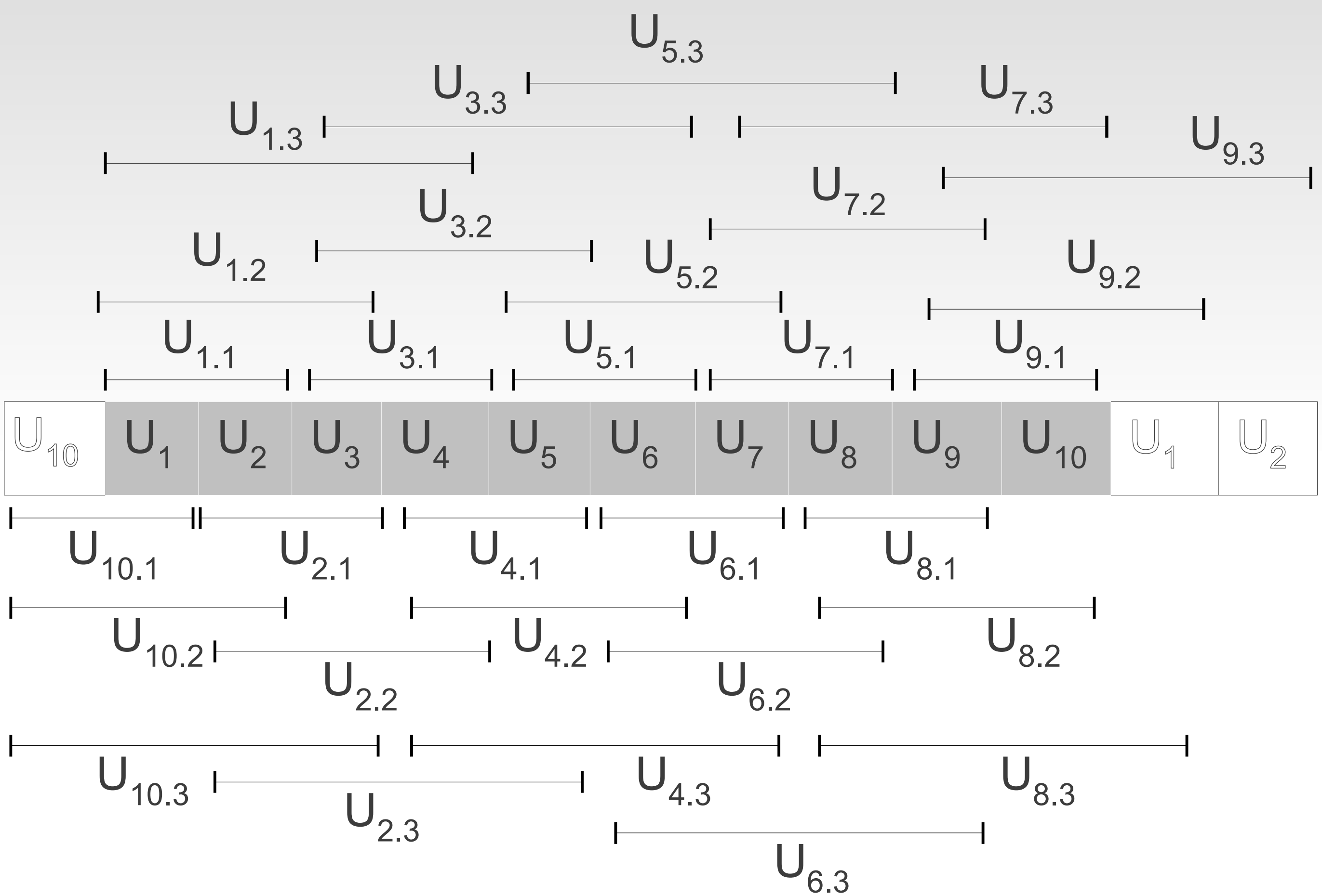
50000 untagged documents



Untagged sampling strategy



Untagged sampling strategy



$|U_i| = 5000$, for i=1..10

Untagged sampling strategy

1-sample	2-sample		10-sample
$U_{1.0}=U_1$	$U_{2.0}=U_2$		$U_{10.0}=U_{10}$
$U_{1.1}=U_1+U_2$	$U_{2.1}=U_2+U_3$		$U_{10.1}=U_{10}+U_1$
$U_{1.2}=U_{1.1}+U_3$	$U_{2.2}=U_{2.1}+U_4$	. . . .	$U_{10.2}=U_{10.1}+U_2$
$U_{1.3}=U_{1.2}+U_4$	$U_{2.3}=U_{2.2}+U_5$		$U_{10.3}=U_{10.2}+U_3$

$(P + U_{i.j})$, i=1..10, j=0..3 $\Rightarrow$ 40 different training sets

Training		**Test**
P size	**Proportions**	**P size**
1000	1:5, 1:10, 1:15, 1:20	110

Untagged sampling strategy

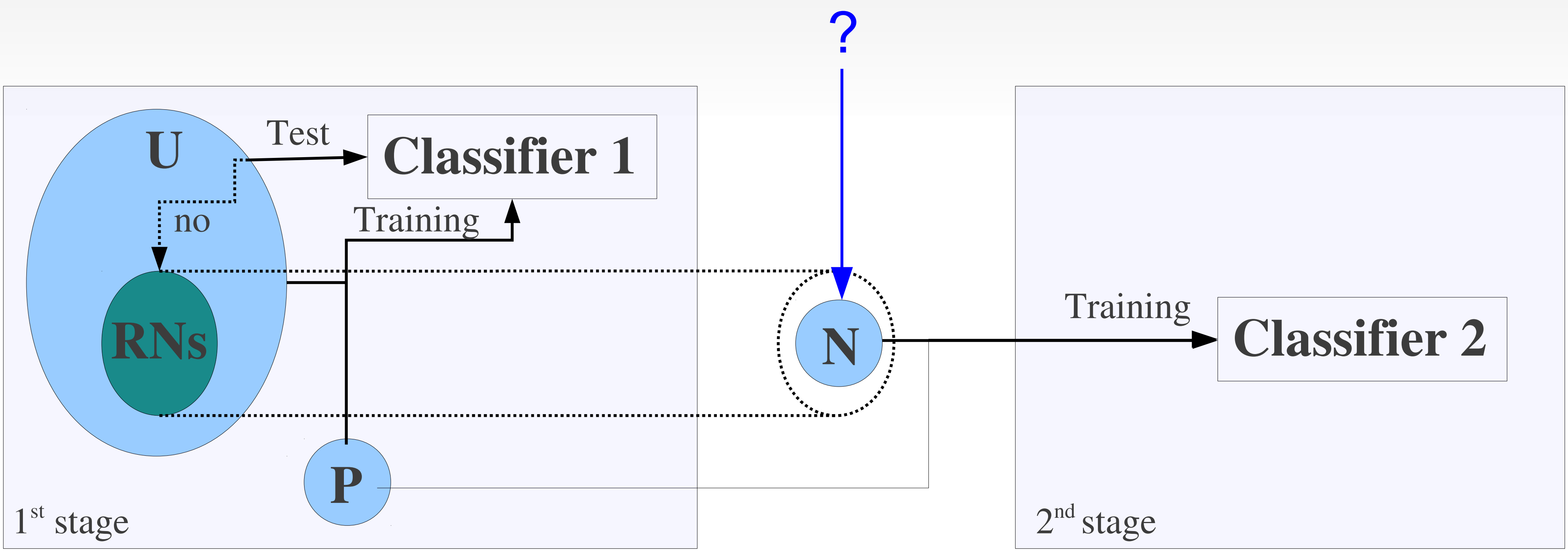
1-sample	2-sample		10-sample
$U_{1.0}=U_1$	$U_{2.0}=U_2$	. . . .	$U_{10.0}=U_{10}$
$U_{1.1}=U_1+U_2$	$U_{2.1}=U_2+U_3$		$U_{10.1}=U_{10}+U_1$
$U_{1.2}=U_{1.1}+U_3$	$U_{2.2}=U_{2.1}+U_4$		$U_{10.2}=U_{10.1}+U_2$
$U_{1.3}=U_{1.2}+U_4$	$U_{2.3}=U_{2.2}+U_5$		$U_{10.3}=U_{10.2}+U_3$

$(P + U_{i.j})$, i=1..10, j=0..3 $\Rightarrow$ 40 different training sets

Training		**Test**
P size	**Proportions**	**P size**
1000	1:5, 1:10, 1:15, 1:20	110

	Advert	Empty	No-foot	Notab	OR	Orphan	PS	Ref	Unref	Wiki
Recall	0.58	0.98	0.57	0.99	0.3	1	0.74	0.61	0.99	0.97

Strategies to select negative set from RNs



Strategies to select negative set from RNs

1. Selecting all RNs as negative set.[3]
2. Selecting |P| documents by random from RNs set.
3. Selecting the |P| best RNs (those assigned the highest confidence prediction values by classifier 1).
4. Selecting the |P| worst RNs (those assigned the lowest confidence prediction values by classifier 1).

[3] Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.

Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

Strategy	*fn* prediction rates				Recall	
	Average	*Median*	*Minimum*	*Maximum*	*Average*	*Median*
1	22.17	3	0	110	0.80	0.97
2	4.48	1	0	26	0.96	0.99
3	4.00	4	0	10	0.96	0.96
4	4.17	1	0	30	0.96	0.99

Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

Strategy	*fn* prediction rates				Recall	
	Average	*Median*	*Minimum*	*Maximum*	*Average*	*Median*
1	22.17	3	0	110	0.80	0.97
2	4.48	1	0	26	0.96	0.99
3	4.00	4	0	10	0.96	0.96
4	4.17	1	0	30	0.96	0.99

Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

Strategy	*fn* prediction rates				Recall	
	Average	*Median*	*Minimum*	*Maximum*	*Average*	*Median*
1	22.17	3	0	110	0.80	0.97
2	4.48	1	0	26	0.96	0.99
3	4.00	4	0	10	0.96	0.96
4	4.17	1	0	30	0.96	0.99

Table 3. Average recall values per flaw

Strategy	Flaws									
	Advert	Empty	No-foot	Notab	OR	Orphan	PS	Ref	Unref	Wiki
1	0.58	0.98	0.57	0.99	0.30	1.00	0.74	0.61	0.99	0.97
2	0.90	0.99	0.86	0.99	1.00	1.00	0.90	0.99	0.99	0.98
3	0.95	0.98	0.94	0.99	0.97	0.99	0.95	0.96	0.97	0.95
4	0.90	0.99	0.89	0.99	1.00	1.00	0.89	0.99	0.99	0.99

Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

Strategy	*fn* prediction rates				Recall	
	Average	*Median*	*Minimum*	*Maximum*	*Average*	*Median*
1	22.17	3	0	110	0.80	0.97
2	4.48	1	0	26	0.96	0.99
3	4.00	4	0	10	0.96	0.96
4	4.17	1	0	30	0.96	0.99

Table 3. Average recall values per flaw

Strategy	Flaws									
	Advert	Empty	No-foot	Notab	OR	Orphan	PS	Ref	Unref	Wiki
1	0.58	0.98	0.57	0.99	0.30	1.00	0.74	0.61	0.99	0.97
2	0.90	0.99	0.86	0.99	1.00	1.00	0.90	0.99	0.99	0.98
3	0.95	0.98	0.94	0.99	0.97	0.99	0.95	0.96	0.97	0.95
4	0.90	0.99	0.89	0.99	1.00	1.00	0.89	0.99	0.99	0.99

SVM: Which kernel?

- Linear SVM (WEKA's default parameters)
- RBF SVM
  - $\gamma \in \{2^{-15}, 2^{-13}, 2^{-11},\ldots, 2^{1}, 2^{3}\}$
  - $C \in \{2^{-5}, 2^{-3}, 2^{-1},\ldots, 2^{13}, 2^{15}\}$

Conclusions

- What classifier in each stage?

  NB + SVM

- Untagged sampling strategy

  Some unlabelled sets are more promising

  - RBF kernel: U_6 sub-sample → 60% of the flaws.
  - Linear kernel: U_4 sub-sample → 60% of the flaws
  - In general, $U_{i,j}$, i=1..10, j=2 or j=3 → best results.

- Strategies for selecting RNs as true negatives

  - 2 ≈ 4 > 3 > 1, ">" means "better than".

Conclusions

- Which SVM kernel and parameters?
  - RBF was better than Linear kernel.
  - High penalty value for the error term ($C = 2^{15}$) and very low γ values ($\gamma \in \{2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}\}$).
- Semi-supervised methods seem very promising.
- As current work, we are developing new features based on factual content measures[4] to assess Advert, Notability and Original Research quality flaws.

[4] E. Lex, M. Völske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer. Measuring the quality of web content using factual information. In Proceedings of the 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12), pages 7–10. ACM, April 2012.

Questions?

Thanks very much for your attention!

SVM: Which kernel?

- Linear SVM (WEKA's default parameters)

Table 4. Recall and fn values for RNs selection strategies

Strategy	*fn* prediction rates				Recall	
	Average	*Median*	*Minimum*	*Maximum*	*Average*	*Median*
2	21	21.5	4	49	0.81	0.80
3	6.20	6	0	20	0.94	0.94
4	20	21	1	44	0.82	0.81

- RBF SVM
  - $\gamma \in \{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{1}, 2^{3}\}$

Table 5. Best γ values

Advert	Empty	No-foot	Notab	OR	Orphan	PS	Ref	Unref	Wiki
2^{-7}	2^{-7}	2^{-5}	2^{-11}	2^{-9}	2^{-9}	2^{-5}	2^{-9}	2^{-9}	2^{-9}

  - $C \in \{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$ ⇨ $C = 2^{15}$