A PU Learning Approach to Quality Flaw Prediction in Wikipedia

Edgardo Ferretti[†]

Web-Technology and Information Systems Group
Bauhaus-Universität Weimar

June 26th - 2012

[†]Universidad Nacional de San Luis, Departamento de Informatica.

Outline

- Motivation
- State of the Art
- PU Learning
- Research questions
- Results
- Conclusions

Motivation

- Web Information Quality $\rightarrow$ critical task.
  - Increasing popularity of user-generated Web content.
  - Unavoidable divergence of the content's quality.
- Wikipedia is a paradigmatic undertaking.
  - Content contributed by millions of users.
    - Main strength
    - Main challenge

State of the Art

- Featured articles identification:
  - Number of edits and editors.[1]
  - Character trigrams distributions.[2]
  - Number of words.[3]
  - Factual information.[4]

[1] D. Wilkinson and B. Huberman. Cooperation and quality in Wikipedia. In Proceedings of the 3th international symposium on wikis and open collaboration (WikiSym'07), ACM, 2007.
[2] N. Lipka and B. Stein. Identifying featured articles in Wikipedia: writing style matters. In Proceedings of the 19th international conference on World Wide Web, ACM, 2010.
[3] J. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In Proceedings of the 17th international conference on World Wide Web (WWW'08), pages 1095–1096. ACM, 2008.
[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.

State of the Art

- Development of quality measurement metrics.[5-7]
- Quality flaws detection.[8-10]

[5] D. Dalip, M. Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by Web communities: a case study of Wikipedia. In Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries (JCDL'09), pages 295–304. ACM, 2009.

[6] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in Wikipedia: models and evaluation. In Proceedings of (CIKM'07, ACM, 2007.

[7] B. Stvilia, M. Twidale, L. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In Proceedings of ICIQ'05, MIT, 2005.

[8] M. Anderka, B. Stein, and N. Lipka. Detection of text quality flaws as a one-class classification problem. In Proceedings of CIKM'11, ACM, 2011.

[9] M. Anderka, B. Stein, and N. Lipka. Towards Automatic Quality Assurance in Wikipedia. In Proceedings of the 20th international conference on World Wide Web (WWW'11), pages 5–6. ACM, 2011.

[10] M. Anderka, B. Stein, and N. Lipka. Using Cleanup Tags to Predict Quality Flaws in User-generated Content. In Proceedings of SIGIR'12, ACM, 2012.

State of the Art

- Quality flaws detection.[8-10]
  - One-class classification problem: impossibility to properly characterize the "class" of documents not containing a particular flaw.
  - Supervised learning approaches.

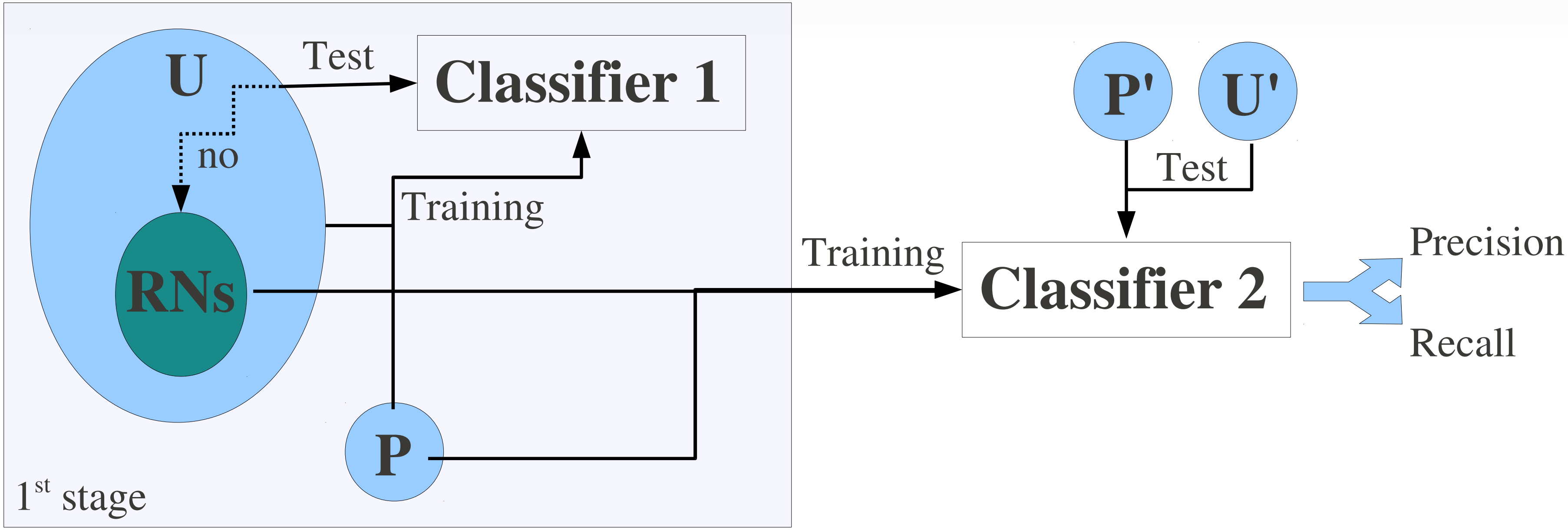
[8] M. Anderka, B. Stein, and N. Lipka. Detection of text quality flaws as a one-class classification problem. In Proceedings of CIKM'11, ACM, 2011.

[9] M. Anderka, B. Stein, and N. Lipka. Towards Automatic Quality Assurance in Wikipedia. In Proceedings of the 20th international conference on World Wide Web (WWW'11), pages 5–6. ACM, 2011.

[10] M. Anderka, B. Stein, and N. Lipka. Using Cleanup Tags to Predict Quality Flaws in User-generated Content. In Proceedings of SIGIR'12, ACM, 2012.

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.[11]



[11] 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?

- Liu's benchmark system on 20-Newsgroups and Reuters-21578:[11]
  - 1st stage: Spy, 1-DNF, Rocchio, NB.
  - 2nd stage: EM, SVM, SVM-I, SVM-IS.
- kNN as 1st stage classifier.[12]
- Our choice: NB + SVM.

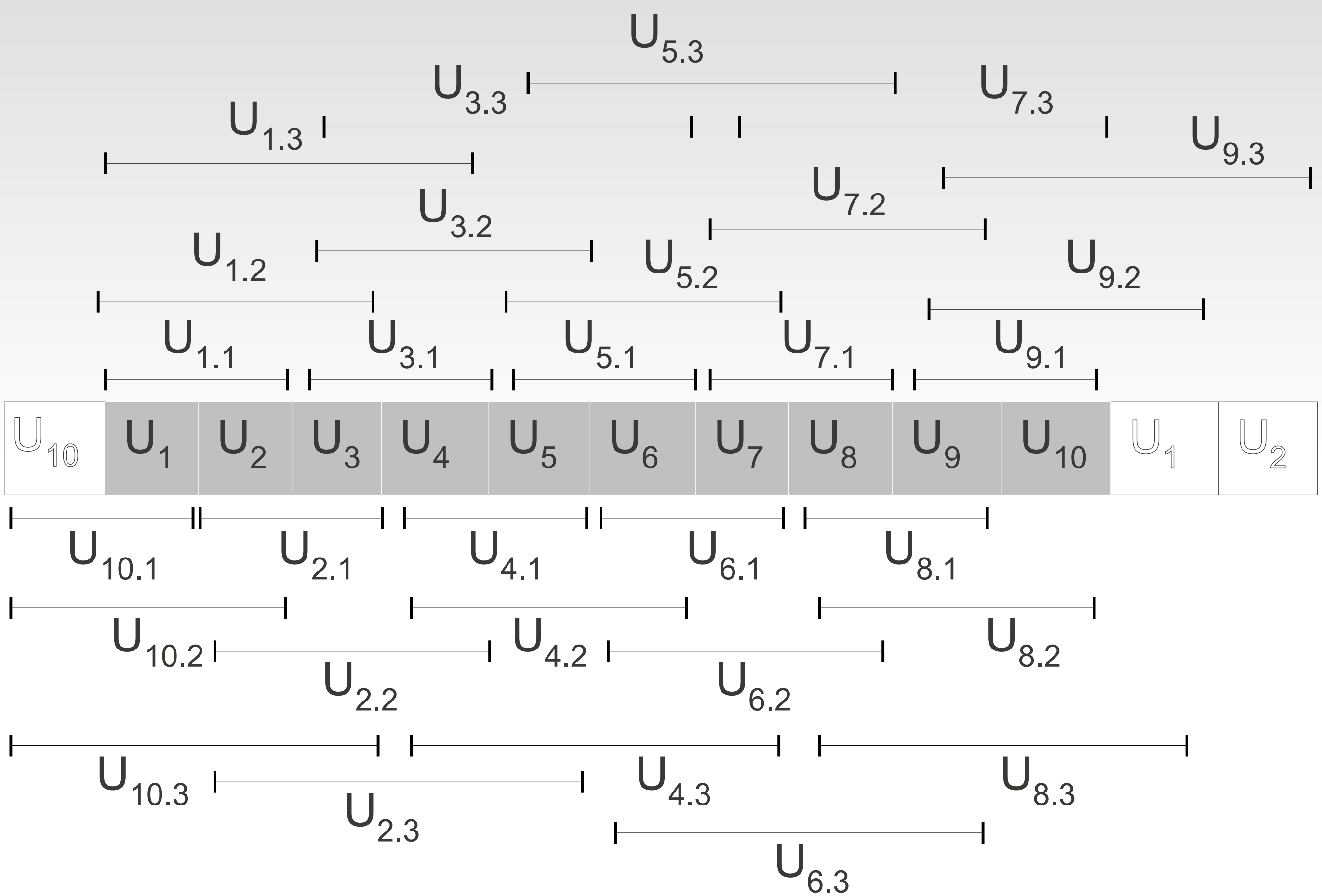
[11] Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.

[12] 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.

PAN@CLEF training release

Flaw name	# Articles	Description
Unreferenced	37572	The article does not cite any references or sources.
Orphan	21356	The article has fewer than three incoming links.
Refimprove	23144	The article needs additional citations for verification.
Empty section	5757	The article has at least one section that is empty.
Notability	3150	The article does not meet the notability guideline.
No footnotes	6068	The article lacks of inline citations.
Primary sources	3682	The article relies on references to primary sources.
Wikify	1771	The article needs to be wikified (links and layout).
Advert	1109	The article is written like an advertisement.
Original research	507	The article contains original research.
	50000	**Untagged articles**

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**	**Proportions**
1000	1:5	110	1:1
2500	1:10		
	1:15		
	1:20		

Strategies to select negative set from RNs

0. Selecting all RNs as negative set.[11]
1. Selecting |P| documents by random from RNs set.
2. Selecting the |P| best RNs (those assigned the highest confidence prediction values by classifier 1).
3. Selecting the |P| worst RNs (those assigned the lowest confidence prediction values by classifier 1).

[11] 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.

SVM: Which kernel?

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

Results for RBF SVM: m1 for RNs

Flaw	Training Set	SVM (Param.)		Precision	Recall	F1	Precision	Recall	F1
		C	γ				Benchmarks[8]		
Advert	U06.2	2^{15}	2^{-7}	0.802	0.955	**0.871**	0.650	0.580	**0.613**
Empty	U06.2	2^{15}	2^{-7}	1.000	0.991	**0.995**	0.740	0.700	**0.719**
No-foot	U04.2	2^{15}	2^{-5}	0.911	0.927	**0.919**	0.590	0.590	**0.590**
Notab	U06.2	2^{15}	2^{-11}	1.000	0.991	**0.995**	0.660	0.610	**0.634**
OR	U06.1	2^{15}	2^{-9}	0.681	0.991	**0.807**	0.560	0.800	**0.659**
Orph	U10.2	2^{15}	2^{-9}	0.991	1.000	**0.995**	0.720	0.590	**0.649**
PS	U04.2	2^{15}	2^{-5}	0.842	0.918	**0.878**	0.610	0.590	**0.600**
Ref	U06.3	2^{15}	2^{-9}	1.000	0.991	**0.995**	0.570	0.560	**0.565**
Unref	U09.3	2^{15}	2^{-9}	1.000	0.991	**0.995**	0.630	0.630	**0.630**
Wiki	U06.3	2^{15}	2^{-9}	0.991	0.991	**0.991**	0.640	0.580	**0.609**

0.944 ← **AVG.** → **0.629**

[8] M. Anderka, B. Stein, and N. Lipka. Detection of text quality flaws as a one-class classification problem. In: Proceedings of CIKM'11, ACM, 2011.

Results for linear SVM: m1 for RNs

Flaw	Training Set	Precision	Recall	F1	Precision	Recall	F1
					Benchmarks[8]		
Advert	U04.2	0.862	0.909	**0.885**	0.650	0.580	**0.613**
Empty	U04.1	0.936	0.927	**0.932**	0.740	0.700	**0.719**
No-foot	U04.3	0.860	0.782	**0.819**	0.590	0.590	**0.590**
Notab	U07.2	0.908	0.900	**0.904**	0.660	0.610	**0.634**
OR	U04.1	0.877	0.845	**0.861**	0.560	0.800	**0.659**
Orph	U07.2	0.913	0.955	**0.933**	0.720	0.590	**0.649**
PS	U04.2	0.898	0.800	**0.846**	0.610	0.590	**0.600**
Ref	U04.2	0.835	0.736	**0.783**	0.570	0.560	**0.565**
Unref	U01.1	0.955	0.964	**0.959**	0.630	0.630	**0.630**
Wiki	U08.3	0.967	0.791	**0.870**	0.640	0.580	**0.609**

0.879 ← **AVG.** → **0.629**

[8] M. Anderka, B. Stein, and N. Lipka. Detection of text quality flaws as a one-class classification problem. In: Proceedings of CIKM'11, ACM, 2011.

Conclusions

- RQ#1: NB + SVM
- RQ#2: Some unlabelled sets are more promising
  - RBF kernel: U_6 sub-sample $\rightarrow$ 60% of the flaws.
  - Linear kernel: U_4 sub-sample $\rightarrow$ 60% of the flaws
  - In general, $U_{i.j}$, i=1..10, j=2 or j=3 $\rightarrow$ best results.
- RQ#3: Method for selecting RNs as true negatives
  - 1 > 2 > 3 > 0, “>” means “better than”.

Conclusions

- RQ#4: SVM kernels and parameters
  - Linear and RBF kernels' best results are statistically significant than the benchmarks.[8]
  - RBF is statistically better than Linear kernel.
  - The optimistic setting in [8] → not statistically significant than Linear and RBF best results .
  - High penalty value for the error term (C) and very low γ values.
- Semi-supervised methods seem very promising.

[8] M. Anderka, B. Stein, and N. Lipka. Detection of text quality flaws as a one-class classification problem. In: Proceedings of CIKM'11, ACM, 2011.

Questions?

Thanks very much for your attention!