On the Use of Reliable-Negatives Selection Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia

Edgardo Ferretti and Marcelo Errecalde
Universidad Nacional de San Luis
{ferretti,merreca}@unsl.edu.ar

Maik Anderka
University of Paderborn
maik.anderka@uni-paderborn.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

11th International Workshop on Text-based Information Retrieval, TIR'14
Munich, Germany

September 4th, 2014

Information Quality in Wikipedia

Situation

- extremely varying content quality
  - everyone can edit Wikipedia, even anonymously
  - heterogeneous community of Wikipedia authors
  - edits are not reviewed before publication
- comprehensive manual quality assurance is unfeasible
  - large data volumes, constantly evolving contents

Information Quality in Wikipedia

Situation

- extremely varying content quality
  - everyone can edit Wikipedia, even anonymously
  - heterogeneous community of Wikipedia authors
  - edits are not reviewed before publication
- comprehensive manual quality assurance is unfeasible
  - large data volumes, constantly evolving contents

Previous work

- research question: “Is an article featured or not?”

  [Hu et al., CIKM’07] [Blumenstock, WWW’08] [Dalip et al., JCDL’09] [Lipka and Stein, WWW’10]

➜ no practical support for Wikipedia’s quality assurance process

➜ less than 0.1% of the English Wikipedia articles are featured

Quality Flaw Prediction in Wikipedia

Question

- How to improve the 99.9% non-featured Wikipedia articles?

Central idea

- automatic exploitation of human-defined cleanup tags [Anderka et al., WWW'11]

Quality Flaw Prediction in Wikipedia

Question

- How to improve the 99.9% non-featured Wikipedia articles?

Central idea

- automatic exploitation of human-defined cleanup tags [Anderka et al., WWW'11]
  - each tag defines a specific quality flaw
  - tagged articles serve as human-labeled examples
  - machine learning is used to predict flaws in untagged articles

Existing flaw prediction approaches

- one-class classification [Anderka et al., WWW'11, SIGIR'12]
- binary classification [Ferschke et al., CLEF'12, ACL'13]
- **PU learning** [Ferretti et al., CLEF'12]

Outline

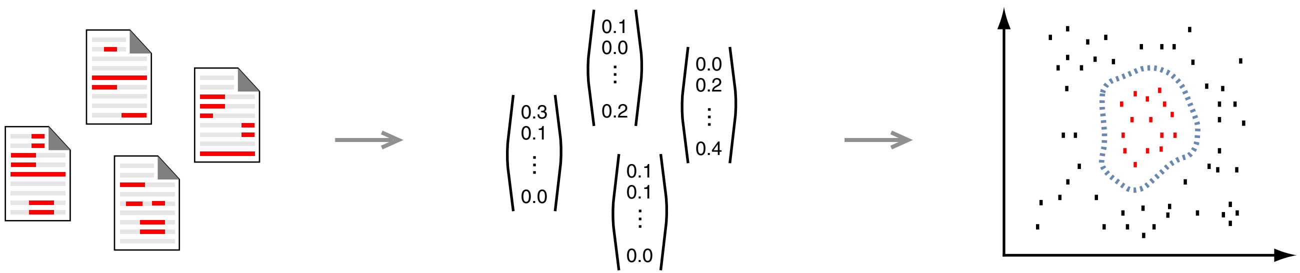
- Motivation
- Problem Statement
- Quality Flaw Prediction Using PU Learning
- Analysis and Empirical Evaluation
- Summary

Problem Statement

Quality flaw prediction in Wikipedia [Anderka et al., SIGIR'12]

- 3.8 M English Wikipedia articles ➜ D
- 445 quality flaws (cleanup tags) ➜ F

- Build a classifier $c : D \rightarrow \{1; 0\}$ for each flaw $f \in F$, given a sample of articles containing f.

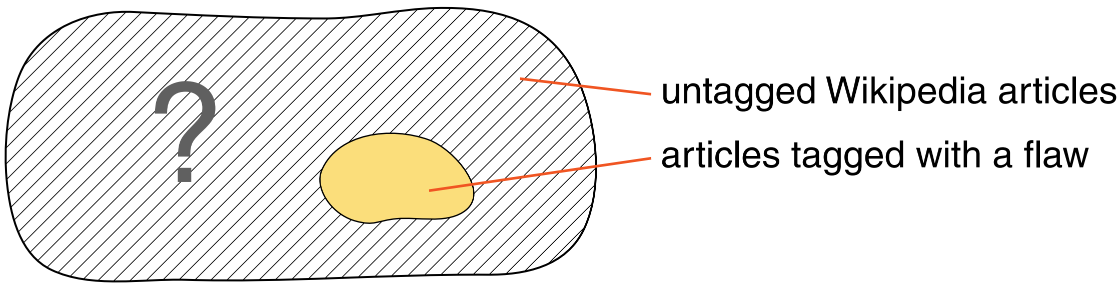


flawed articles — article representation (document model) — one-class classifier

Problem Statement

Quality flaw prediction using PU learning [Ferretti et al., CLEF'12]

- exploit untagged articles to improve the effectiveness of a classifier c



  – in Wikipedia, it is more than likely that many flaws are not yet identified

→ PU learning: learning from *Positive* and *Unlabeled* examples [Liu et al., ICML'02]

  – *positive* examples = articles tagged with a flaw
  – *unlabeled* examples = untagged articles (either flawed or flawless)

Problem Statement

Background: PU learning [Liu et al., ICML'02]

- set P of positive examples
- set U of unlabeled examples (containing both positive and negative examples)
- Build a classifier using P and U that can identify positive examples in U or in a separate test set.
- two-stage approach:
  1. identifying *reliable negatives*
     - train a binary classifier using P and U
     - apply this classifier to the examples in U
     - consider all examples not classified as "positive" as *reliable negatives*
  2. building the final classifier (non-iterative version)
     - train a binary classifier using P and the set of *reliable negatives*

Problem Statement

Crucial aspects in the Wikipedia setting

1. unknown (flaw-specific) class imbalances
   - 1^{st} stage: ratio between P and U
   - 2^{nd} stage: ratio between P and the set of *reliable negatives*
2. effects of sampling (essential in practice due to the large number of existing Wikipedia articles)
   - 1^{st} stage: U is very large for most flaws
   - 2^{nd} stage: the set of *reliable negatives* can become considerably large

- have not—or only partially—addressed by Liu et al. and Ferretti et al.

➜ we show where in the PU learning procedure sampling is useful

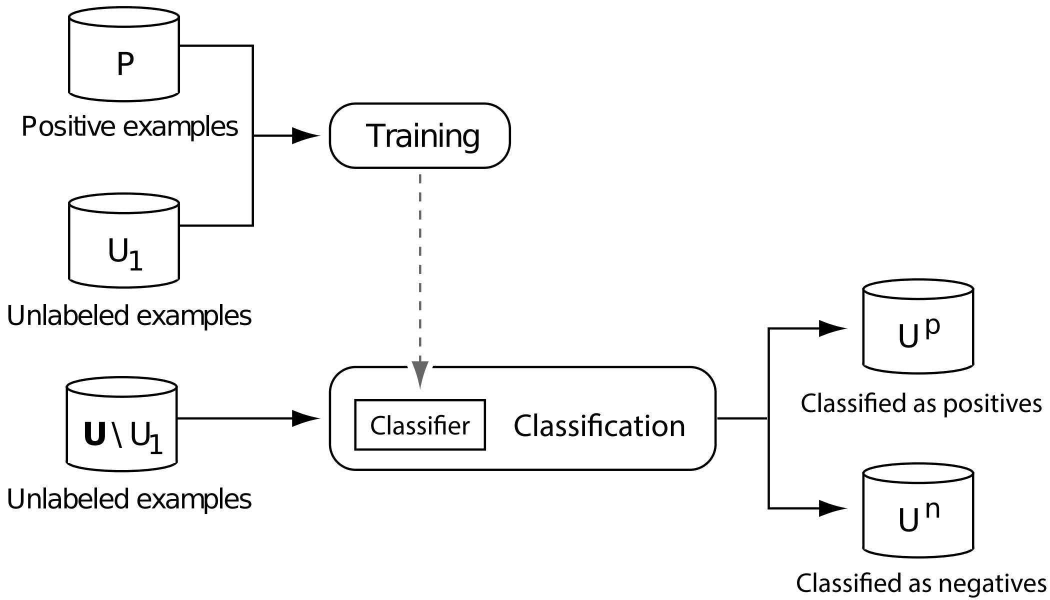
➜ we analyze how different sampling strategies affect the flaw prediction effectiveness

Outline

- Motivation
- Problem Statement
- Quality Flaw Prediction Using PU Learning
- Analysis and Empirical Evaluation
- Summary

Quality flaw prediction using PU learning

1^{st} stage: identifying *reliable negatives*



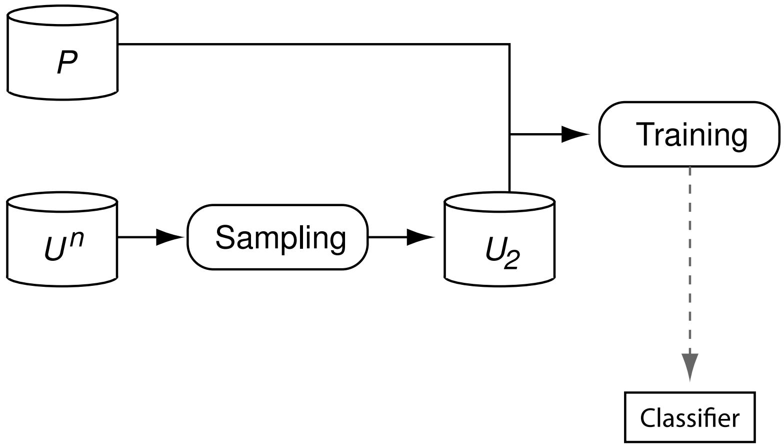
- U_1 is a sample from U
- training set is balanced, $|P| = |U_1|$

➜ sampling strategy does not affect the flaw prediction performance

➜ random sampling

Quality flaw prediction using PU learning

2^{st} stage: building the final classifier



- using $U_2 = U^n$ worsened the performance by up to 50% [Ferretti et al., CLEF'12]
- sampling strategies:

M_1 selecting $|P|$ articles by random from U^n

M_2 selecting the $|P|$ *best* articles from U^n
(those assigned the highest confidence values by the first-stage classifier)

M_3 selecting the $|P|$ *worst* articles from U^n
(those assigned the lowest confidence values by the first-stage classifier)

Outline

- Motivation
- Problem Statement
- Quality Flaw Prediction Using PU Learning
- Analysis and Empirical Evaluation
- Summary

Analysis and Empirical Evaluation

Experimental design

- evaluation corpus of the "1^{st} international competition on quality flaw prediction in Wikipedia"
  - 1,592,226 English Wikipedia articles
  - 208,228 tagged to contain one of ten important quality flaws

- 1^{st} stage classifier: Naïve Bayes
- 2^{nd} stage classifier: Support Vector Machine (SVM)

- balanced training sets: $|P| = |U_1|$ and $|P| = |U_2|$
- random sampling in the 1^{st} stage
- M_1, M_2, and M_3 in the 2^{nd} stage

Analysis and Empirical Evaluation

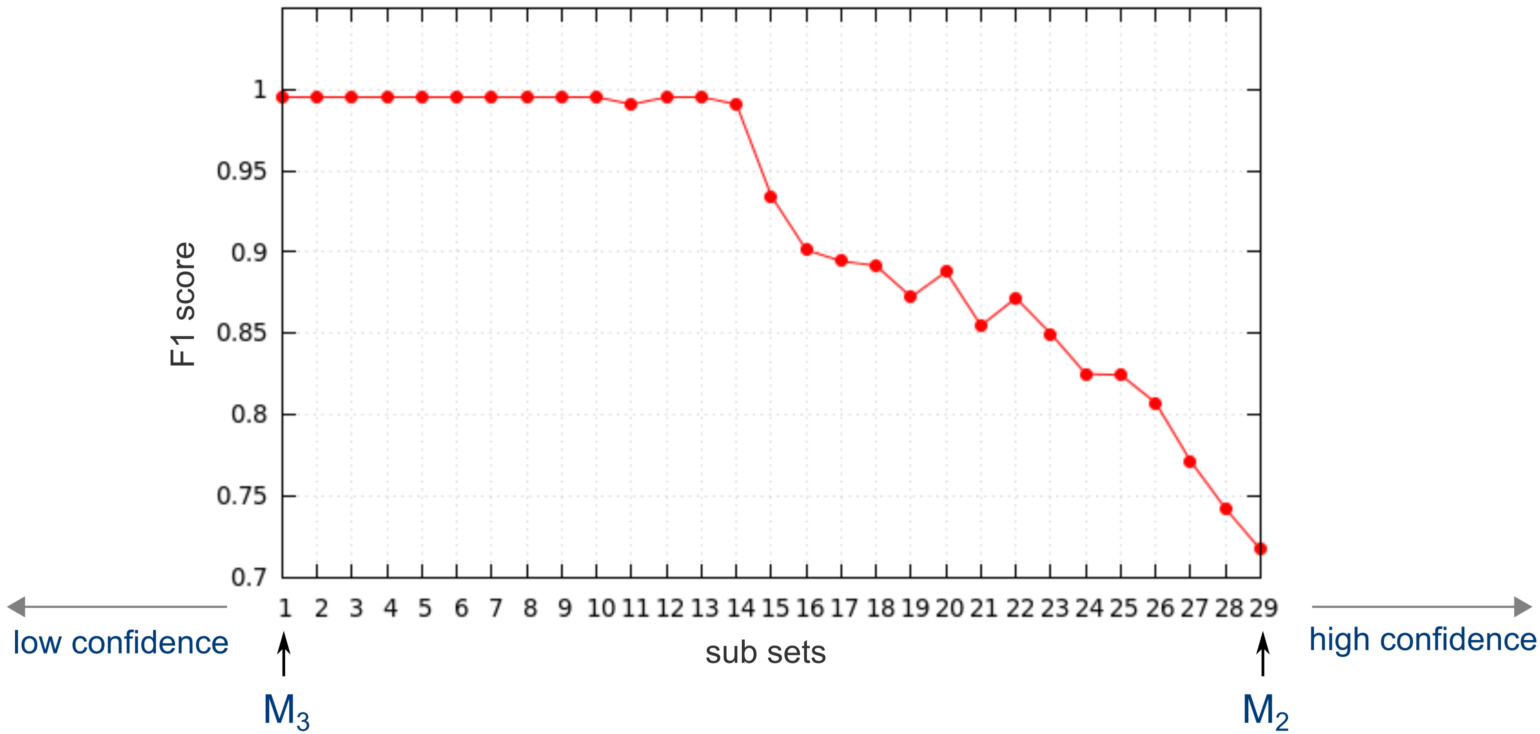
Selecting *reliable negatives* (2^{nd} stage sampling)

- flaw *Unreferenced*: $|U^n| =$ 29,635, $|P| = |U_2| =$ 1,000

Analysis and Empirical Evaluation

Selecting *reliable negatives* (2^{nd} stage sampling)

- flaw *Unreferenced*: $|U^n| = $ 29,635, $|P| = |U_2| = $ 1,000



➜ strategy M_3 outperforms M_2

➜ differences between M_3 and M_1 (random) are not statistically significant

Analysis and Empirical Evaluation

Flaw prediction effectiveness

effectiveness of PU learning in terms of F1 score for the ten quality flaws

flaw name	**baseline** [Ferretti et al., CLEF'12]	**proposed approach** using strategy M_3
Advert	0.8214	0.9440 (+14.93%)
Empty section	0.8216	0.9394 (+14.34%)
No footnotes	0.8264	0.9826 (+18.90%)
Notability	0.7944	0.9886 (+24.45%)
Orphan	0.8986	0.9960 (+10.84%)
Original research	0.7638	0.9338 (+22.26%)
Primary sources	0.8068	0.9891 (+22.60%)
Refimprove	0.8362	0.9382 (+12.20%)
Unreferenced	0.8365	0.9432 (+12.76%)
Wikify	0.7396	0.9818 (+32.75%)
averaged over all flaws	**0.8145**	**0.9637 (+18.31%)**

Outline

- Motivation
- Problem Statement
- Quality Flaw Prediction Using PU Learning
- Analysis and Empirical Evaluation
- Summary

Summary

What we have done

1. shed light on the effects of sampling in PU learning
   - ➜ sampling is necessary (in both stages)
   - ➜ in general, sampling strategy M_3 is favorable
2. improved PU learning approach for quality flaw prediction in Wikipedia
   - ➜ average improvement of 18.31% compared to the baseline

Summary

What we have done

1. shed light on the effects of sampling in PU learning
   - ➜ sampling is necessary (in both stages)
   - ➜ in general, sampling strategy M_3 is favorable
2. improved PU learning approach for quality flaw prediction in Wikipedia
   - ➜ average improvement of 18.31% compared to the baseline

Current work

- ❏ comparative study of the existing flaw prediction approaches

Thank you!

maik.anderka@uni-paderborn.de

Appendix

Article representation

- 65 state-of-the-art features, 30 new features

content	characters, words, syllables, sentences, readability, parts of speech, closed-class word sets, ...
structure	sections, tables, images, references, categories, templates, lists, specific sections, ...
network	internal-, external-, interwiki-, broken links, PageRank, citation measures, ...
edit history	age, currency, connectivity, revisions, reverts, editors, cooperation, ...