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
- Manuel Montes
- Rafael Guzmán

Methodological Design

PU Learning

Research questions

Conclusions
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

---


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]

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

---

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

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

What classifier in each stage?

1st stage
- U
- R
- P
  - Training
  - Test

Classifier 1
- Spy, 1-DNF, Rocchio, NB, KNN

2nd stage
- EM, SVM, SVM-I, SVM-1S

Classifier 2
- ?

Who are we? Methodological Design PU Learning Research questions #1 Conclusions
What classifier in each stage?

Spy, 1-DNF, Rocchio, NB, KNN

EM, SVM, SVM-I, SVM-IS

What classifier in each stage?

1st stage
- **U** (Unlabeled)
- **P** (Positive)
- **RNs**
  - **Test**
  - **Training**

**Classifier 1**
- Spy, 1-DNF, Rocchio, NB, KNN

2nd stage
  - **Training**
  - **Classifier 2**
    - EM, SVM, SVM-I, SVM-IS

---

What classifier in each stage?

1st stage
- U
- RNs
- P

2nd stage
- NB, KNN, SVM

Who are we?
Methodological Design
PU Learning
Research questions
Conclusions
What classifier in each stage?

Our choice: NB + SVM
Untagged sampling strategy

50000 untagged documents

1\textsuperscript{st} stage

\begin{itemize}
\item \textbf{RNs}
\item \textbf{U}
\end{itemize}

\textbf{Test}

\textbf{Training}

2\textsuperscript{nd} stage

\begin{itemize}
\item \textbf{NB Classifier}
\item \textbf{SVM Classifier}
\end{itemize}

Who are we? Methodological Design PU Learning Research questions Conclusions
Untagged sampling strategy

50000 untagged documents

10-fold cross-validation

Training

Test

PU Learning
Untagged sampling strategy

\[ |U_i| = 5000, \text{ for } i=1\ldots10 \]
**Untagged sampling strategy**

1-sample

<table>
<thead>
<tr>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{1.0} = U_1$</td>
<td>$U_{1.1} = U_1 + U_2$</td>
<td>$U_{1.2} = U_{1.1} + U_3$</td>
<td>$U_{1.3} = U_{1.2} + U_4$</td>
</tr>
</tbody>
</table>

2-sample

<table>
<thead>
<tr>
<th>2.0</th>
<th>2.1</th>
<th>2.2</th>
<th>2.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{2.0} = U_2$</td>
<td>$U_{2.1} = U_2 + U_3$</td>
<td>$U_{2.2} = U_{2.1} + U_4$</td>
<td>$U_{2.3} = U_{2.2} + U_5$</td>
</tr>
</tbody>
</table>

10-sample

<table>
<thead>
<tr>
<th>10.0</th>
<th>10.1</th>
<th>10.2</th>
<th>10.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{10.0} = U_{10}$</td>
<td>$U_{10.1} = U_{10} + U_1$</td>
<td>$U_{10.2} = U_{10.1} + U_2$</td>
<td>$U_{10.3} = U_{10.2} + U_3$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P size</strong></td>
<td><strong>Proportions</strong></td>
</tr>
<tr>
<td>1000</td>
<td>1:5, 1:10, 1:15, 1:20</td>
</tr>
</tbody>
</table>
Untagged sampling strategy

1-sample
\[ U_{1.0} = U_1 \]
\[ U_{1.1} = U_1 + U_2 \]
\[ U_{1.2} = U_{1.1} + U_3 \]
\[ U_{1.3} = U_{1.2} + U_4 \]

2-sample
\[ U_{2.0} = U_2 \]
\[ U_{2.1} = U_2 + U_3 \]
\[ U_{2.2} = U_{2.1} + U_4 \]
\[ U_{2.3} = U_{2.2} + U_5 \]

10-sample
\[ U_{10.0} = U_{10} \]
\[ U_{10.1} = U_{10} + U_1 \]
\[ U_{10.2} = U_{10.1} + U_2 \]
\[ U_{10.3} = U_{10.2} + U_3 \]

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

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>P size</td>
<td>Proportions</td>
</tr>
<tr>
<td>1000</td>
<td>1:5, 1:10, 1:15, 1:20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recall</th>
<th>Advert</th>
<th>Empty</th>
<th>No-foot</th>
<th>Notab</th>
<th>OR</th>
<th>Orphan</th>
<th>PS</th>
<th>Ref</th>
<th>Unref</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.58</td>
<td>0.98</td>
<td>0.57</td>
<td>0.99</td>
<td>0.3</td>
<td>1</td>
<td>0.74</td>
<td>0.61</td>
<td>0.99</td>
<td>0.97</td>
<td></td>
</tr>
</tbody>
</table>
Strategies to select negative set from RNs

1\textsuperscript{st} stage

U

no

Test

Training

Classifier 1

RN$\textsuperscript{s}$

P

2\textsuperscript{nd} stage

N

Training

Classifier 2

Who are we? Methodological Design PU Learning Research questions #1 #2 #3 Conclusions
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).

## Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

<table>
<thead>
<tr>
<th>Strategy</th>
<th>fn prediction rates</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Median</td>
</tr>
<tr>
<td>1</td>
<td>22.17</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4.48</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4.00</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4.17</td>
<td>1</td>
</tr>
</tbody>
</table>
Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

<table>
<thead>
<tr>
<th>Strategy</th>
<th>fn prediction rates</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Median</td>
</tr>
<tr>
<td>1</td>
<td>22.17</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4.48</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4.00</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4.17</td>
<td>1</td>
</tr>
</tbody>
</table>
Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

<table>
<thead>
<tr>
<th>Strategy</th>
<th>fn prediction rates</th>
<th>Recall</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Median</td>
<td>Minimum</td>
</tr>
<tr>
<td>1</td>
<td>22.17</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4.48</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4.00</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4.17</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3. Average recall values per flaw

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Flaws</th>
<th>Advert</th>
<th>Empty</th>
<th>No-foot</th>
<th>Notab</th>
<th>OR</th>
<th>Orphan</th>
<th>PS</th>
<th>Ref</th>
<th>Unref</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.58</td>
<td>0.98</td>
<td>0.57</td>
<td>0.99</td>
<td>0.30</td>
<td>1.00</td>
<td>0.74</td>
<td>0.61</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.90</td>
<td>0.99</td>
<td>0.86</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.90</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.95</td>
<td>0.98</td>
<td>0.94</td>
<td>0.99</td>
<td>0.97</td>
<td>0.99</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.90</td>
<td>0.99</td>
<td>0.89</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.89</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Strategies to select negative set from RNs

Table 2. Recall and fn values for RNs selection strategies

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Average fn prediction rates</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Median</td>
</tr>
<tr>
<td>1</td>
<td>22.17</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4.48</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4.00</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4.17</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Average recall values per flaw

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Flaws</th>
<th>Advert</th>
<th>Empty</th>
<th>No-foot</th>
<th>Notab</th>
<th>OR</th>
<th>Orphan</th>
<th>PS</th>
<th>Ref</th>
<th>Unref</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.58</td>
<td>0.98</td>
<td>0.57</td>
<td>0.99</td>
<td>0.30</td>
<td>1.00</td>
<td>0.74</td>
<td>0.61</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.90</td>
<td>0.99</td>
<td>0.86</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.90</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.95</td>
<td>0.98</td>
<td>0.94</td>
<td>0.99</td>
<td>0.97</td>
<td>0.99</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.90</td>
<td>0.99</td>
<td>0.89</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.89</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
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 $\rightarrow$ 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.

- Strategies for selecting RNs as true negatives
  - $2 \approx 4 > 3 > 1$, “$>$” 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 $\gamma$ 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\cite{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.} to assess Advert, Notability and Original Research quality flaws.
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

<table>
<thead>
<tr>
<th>Strategy</th>
<th>fn prediction rates</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Median</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>21.5</td>
</tr>
<tr>
<td>3</td>
<td>6.20</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>21</td>
</tr>
</tbody>
</table>

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

  Table 5. Best $\gamma$ values

<table>
<thead>
<tr>
<th>Advert</th>
<th>Empty</th>
<th>No-foot</th>
<th>Notab</th>
<th>OR</th>
<th>Orphan</th>
<th>PS</th>
<th>Ref</th>
<th>Unref</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^{-7}$</td>
<td>$2^{-7}$</td>
<td>$2^{-5}$</td>
<td>$2^{-11}$</td>
<td>$2^{-9}$</td>
<td>$2^{-9}$</td>
<td>$2^{-5}$</td>
<td>$2^{-9}$</td>
<td>$2^{-9}$</td>
<td>$2^{-9}$</td>
</tr>
</tbody>
</table>

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