Design, Prototypical Implementation, and Evaluation of an Active Machine Learning Service in the Context of Legal Text Classification

Johannes Muhr, Feb 13th 2017, Munich

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de
### Key Facts

- **Title (German)**: Design, Prototypische Implementierung, und Evaluation eines Active Machine Learning Services im Kontext von Rechtstexten

- **Advisor**: Bernhard Waltl

- **Supervisor**: Prof. Dr. Florian Matthes

- **Project**: LexAlyze – Analysis of Legal Texts

- **Chair**: Software Engineering for Business Information Systems (SEBIS)

- **Student**: Johannes Muhr

- **Start**: January, 15th 2017

- **Submission**: July, 15th 2017
Outline

1. Motivation
2. Active Learning
3. Research Questions
4. Solution Approach
5. Roadmap
Motivation

- Huge amount of legal documents are produced every day
- Many different kinds of legal documents

Motivation

• Manual document classification is very expensive and time consuming
  • 13,5 Million $ were spent for classifying 1,6 Million items needing 4 month (= 8,50$ per document) [1]
  • A lot of time is wasted with (document) discovery [2]

<table>
<thead>
<tr>
<th></th>
<th>Senior Partner</th>
<th>Associate</th>
<th>Senior Associate</th>
<th>Junior Partner</th>
<th>Junior Partner</th>
<th>Senior Partner</th>
<th>All Others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Documents</td>
<td>60</td>
<td>311</td>
<td>162</td>
<td>95</td>
<td>57</td>
<td>39</td>
<td>197</td>
<td>921</td>
</tr>
<tr>
<td>Depositions</td>
<td>105</td>
<td>125</td>
<td>125</td>
<td>59</td>
<td>83</td>
<td>2</td>
<td>26</td>
<td>525</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>436</td>
<td>287</td>
<td>154</td>
<td>141</td>
<td>41</td>
<td>223</td>
<td>1,446</td>
</tr>
<tr>
<td>Communications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal</td>
<td>155</td>
<td>160</td>
<td>101</td>
<td>53</td>
<td>54</td>
<td>35</td>
<td>80</td>
<td>628</td>
</tr>
<tr>
<td>Opposition</td>
<td>68</td>
<td>149</td>
<td>86</td>
<td>24</td>
<td>44</td>
<td>0</td>
<td>41</td>
<td>411</td>
</tr>
<tr>
<td>Client</td>
<td>82</td>
<td>45</td>
<td>46</td>
<td>33</td>
<td>30</td>
<td>13</td>
<td>7</td>
<td>256</td>
</tr>
<tr>
<td>Total</td>
<td>303</td>
<td>344</td>
<td>232</td>
<td>110</td>
<td>128</td>
<td>48</td>
<td>128</td>
<td>1,295</td>
</tr>
<tr>
<td>Pleadings &amp; Research</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pleadings</td>
<td>35</td>
<td>282</td>
<td>162</td>
<td>44</td>
<td>37</td>
<td>18</td>
<td>65</td>
<td>643</td>
</tr>
<tr>
<td>Legal Research</td>
<td>8</td>
<td>126</td>
<td>114</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>126</td>
<td>377</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>408</td>
<td>276</td>
<td>44</td>
<td>38</td>
<td>19</td>
<td>191</td>
<td>1,019</td>
</tr>
<tr>
<td>Expert Witness Support</td>
<td>35</td>
<td>21</td>
<td>26</td>
<td>263</td>
<td>22</td>
<td>70</td>
<td>70</td>
<td>457</td>
</tr>
<tr>
<td>Administration</td>
<td>-</td>
<td>6</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>973</td>
<td>983</td>
<td>130,683</td>
</tr>
<tr>
<td>All Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Settlement</td>
<td>32</td>
<td>69</td>
<td>19</td>
<td>21</td>
<td>5</td>
<td>14</td>
<td>-</td>
<td>159</td>
</tr>
<tr>
<td>Other</td>
<td>12</td>
<td>24</td>
<td>29</td>
<td>12</td>
<td>4</td>
<td>7</td>
<td>20</td>
<td>109</td>
</tr>
<tr>
<td>Hearings</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>96</td>
<td>53</td>
<td>33</td>
<td>11</td>
<td>21</td>
<td>21</td>
<td>286</td>
</tr>
<tr>
<td>Total Hours</td>
<td>597</td>
<td>1,310</td>
<td>878</td>
<td>595</td>
<td>341</td>
<td>199</td>
<td>1,586</td>
<td>5,486</td>
</tr>
<tr>
<td>Total Dollars</td>
<td>305,117</td>
<td>310,202</td>
<td>258,724</td>
<td>250,005</td>
<td>250,005</td>
<td>146,439</td>
<td>138,305</td>
<td>297,531</td>
</tr>
</tbody>
</table>

Hours: $\frac{1,446}{5,486} = 26,4\%$

Dollars: $\frac{483,986}{1,697,322} = 28,5\%$

Result

- **Document** and **Sentence classification** is a hot topic
- Manual classification is very expensive and time-consuming
- Machine learning approach is supposed to help here

Solution Approaches

1. Use of (Ruta) Rules
2. Active Machine Learning (AL)
3. Combination of Ruta Rules and AL
Outline

1. Motivation

2. Active Learning

3. Research Questions

4. Solution Approach

5. Roadmap
Active Learning – Motivation

• **Why using Active Machine Learning for Document- & Sentence Classification?**
  • Detection of rules is limited
    • Minor linguistic variations are enough that sentences are not classified accordingly
      • “Im Sinne des Gesetzes” !≠ “Im Sinne der Gesetze”
  • Active learning has already been successfully applied in
    ✓ text classification [3]
    ✓ and also within the legal environment [4]

[3] Novak, Mladenič, & Grobelnik, 2006; S. Tong & Koller, 2002; Segal, Markowitz, & Arnold, 2006
Active Learning – Overview

• Subfield of machine learning with people in the loop (iterative & interactive form)
• **Goal**: Reduce size of needed trainings data by labelling those instances that are especially helpful
• Many influencing factors need to be considered (e.g. classifier, query strategy)
Active Learning – Data Set

• **Document classification**
  • >100 000 documents
  • Manually labelled set of documents received from Datev

• **Sentence Classification**
  • Available from laws (Lexia)
  • Manual classification with the help of Elena Scepankova
1. Motivation

2. Active Learning

3. Research Questions

4. Solution Approach

5. Roadmap
What are common concepts, strategies and technologies used in the context of text classification?

How can (active) machine learning support the classification of legal documents and their content (sentences)?

What does the concept and design of an active machine learning service look like?

How well does the active machine learning service in the classification of legal documents and their content (sentences) perform?
Outline

1. Motivation

2. Active Learning

3. Research Questions

4. Solution Approach

5. Roadmap
Literature Study and Framework Assessment

- Machine Learning
- (Legal) Text Classification
- Active Learning

Analysis of Machine Learning Frameworks

February 13th 2017, Kick-off presentation Johannes Muhr
Preliminary Architecture

Scope of thesis

Lexia

Rest API

Machine Learning Microservice

Machine Learning Framework

Model Store
Outline

1. Motivation
2. Active Learning
3. Research Questions
4. Solution Approach
5. Roadmap
Johannes Muhr
Advisor: Bernhard Waltl

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel  +49.89.289.17132
Fax  +49.89.289.17136

matthes@in.tum.de
www.matthes.in.tum.de
Bibliography


Backup – Literature study

• **Use of online Platforms like**
  • Google Scholar,
  • Web of Science,
  • Institute of Electrical and Electronics Engineers (IEEE),
  • or Online Public Access Catalogue (OPAC) and Google Books

• **Backwards Search**