Automated extraction of balance sheet information

KNIME Summit 2019
1 Smart Data at LBBW

2 Automated extraction of balance sheet information
LBBW: A brief overview

**Medium-sized universal bank**
LBBW is a medium-sized universal bank with deep roots in Baden-Württemberg

**Strong partner for our customers**
We are a partner for SMEs, listed corporations, institutional customers, savings banks and private customers

**Data and facts**
- Head offices: Stuttgart, Mainz, Karlsruhe and Mannheim
- Group employees December 2017: Approximately 10,300
- Total assets December 2017: EUR 238bn

**Specialized Group subsidiaries**
- LBB Asset Management
- LBB Immobilien
- MKB Bank
- Südleasing
- SüdFactoring
Our four priorities

Medium-sized universal bank

- Solid customer relationships
- Owners with long-term orientation
- Strong identification of employees
- Deeply rooted in the region
- Strong capital based
- Innovations & process excellence

Business Focus

- Digitalization
- Sustainability
- Agility

Output-oriented  Consistent  Responsible
The Smart-Data Concept

Digital touch points

Archived states

Turning Data into Value

Mathematics and statistical methods

Better Decisions

Better Decisions
## Achieve impact at all touch points

<table>
<thead>
<tr>
<th>People</th>
<th>Markets</th>
<th>Companies</th>
<th>Cooperations</th>
</tr>
</thead>
<tbody>
<tr>
<td>§ Product affinity §</td>
<td>§ Risk prediction §</td>
<td>§ Balance sheet analysis §</td>
<td>§ „Data Analytics as a Service“ §</td>
</tr>
<tr>
<td>§ Networks and Relations §</td>
<td>§ Risk prediction §</td>
<td>§ Implementation Industry 4.0 §</td>
<td>for our corporate customer §</td>
</tr>
<tr>
<td>§ Forecast behaviour §</td>
<td>§ Sensitivity analysis §</td>
<td>§ Machine-Data §</td>
<td>§ Scientific cooperation with §</td>
</tr>
<tr>
<td>§ purchasing §</td>
<td>§ Market and competition §</td>
<td>§ Document management §</td>
<td>Universities and the KIT(^1) §</td>
</tr>
<tr>
<td>§ Fraud detection and §</td>
<td>§ Network analysis and §</td>
<td>§ Fraud detection and §</td>
<td>§ Saving Bank Finance §</td>
</tr>
<tr>
<td>§ money laundering §</td>
<td>§ behavior forecasting §</td>
<td>§ money laundering §</td>
<td>Group §</td>
</tr>
<tr>
<td>§ Process automation §</td>
<td>§ Trend analysis §</td>
<td>§ Transaction Management §</td>
<td>§ Technical Cooperation with §</td>
</tr>
<tr>
<td>§ Risk prediction §</td>
<td>§ … §</td>
<td>§ for Data §</td>
<td>the HLRS(^2) §</td>
</tr>
<tr>
<td>§ … §</td>
<td>§ … §</td>
<td>§ … §</td>
<td>§ … §</td>
</tr>
</tbody>
</table>

1\(^{1}\)Karlsruher Institut für Technologie
2\(^{2}\)Höchstleistungsrechenzentrum Stuttgart
One technical platform for all business units

Mission: implement Artificial Intelligence in the business processes
Some examples

**Product affinity**

Cluster analysis

**Network analysis and relationships**

Text mining
1 Smart Data at LBBW

2 Automated extraction of balance sheet information

- data
- goal
- pilot project in 4 use cases
  - method
  - results
- next steps
Goals of Balance Analysis

Motivation

- departments like Research and Advisory need all these information to provide their clients with tailor-made and targeted advice
- important for investments and comparison with other companies in the same peer group

Previous manual approach

- reading thousands of balances manually
- needing all information on each page
- extract special key figures, hints and facts about risks, opportunities and other KPIs of a balance and relatable companies in the same branch

Benefits of doing it automatically

- faster
- cost saving
- concentrated on what departments do
- no need to read everything
Pilot Project – In 4 Use Cases

- Why extract key figures?
- Examples: revenue, EBIT (earnings before interest) etc.
- How?

Models learn **word patterns** in order to identify and distinguish key figures, text section candidates and industries.
Extract key figures from balance tables

Benefits & Drawbacks

• Almost all key figures we want
• Nicely structured information …
• …but difficult to extract from PDF files.
• Almost impossible to identify after parsing PDF files into plain text.
• Common KNIME Parser Nodes not applicable.
Solution: New PDF – Parser Node
Features: Words in context

- Words close by that frequently cooccur
- Create word vectors
Model training & Evaluation

- Train & evaluate best performing model
- Deep Learning is not the solution for everything
Extract key figures from free text

Why free text search?

- Not necessarily all key figures present in tables
- Additional contextual information
- Why not?
Difficulties

Large data volume
- #pages ≈ 50 - 450
- #words ≈ 10k – 500k

Relevance
- only key figures representing forecasts development
- Classify key figures (4 different classes in training)

One Sentence ≠ one Class
XXYZ hat als Zielgröße einen mittelfristigen Umsatz von rund 1 Milliarde € definiert und erneuert den mittelfristig angestrebten EBITDA-Margen-Korridor von 18–20%.

Structure matters

Value & Type

Slide 17
Identify structure of balance

- Infer reading directions from word positions

Thanks for making it possible
Classify relevant sentences & type of key figure

Text sections

Sentence extraction

Word vector creation

Relevance Prediction

Predict Key Figure X

Predict Key Figure Y

Predict Key Figure Z

> Für die Umsatzerlöse wird mit einer Steigerung um mindestens 9 % auf mindestens 350 Mio. € gerechnet.

Relevant: YES
Figure Type: Umsatzerlöse

March 2019  ·  KNIME Summit 2019  ·  Slide 19
Model evaluation & Overall accuracy

### Precision (%) | Recall (%)
---|---
80.6 | 73.5

### Ø Precision (%) | Ø Recall (%)
---|---
76.2 | 78.5

| Precision (%) | Recall (%) | Accuracy (%) |
---|---|---|
58.2 | 65.4 | \(~60\)

**Ways to improve**

- More training data for underrepresented classes (only 100 training sentences available)
- Improve iteratively (e.g., with Active Learning)
Demo
Next steps

- tailor-made and targeted advice for investments
- all information about the development potential of the other companies in the same peer group
- service for our costumer

building peer groups

webcrawling news and information from feeds, blogs etc

combine with extracted risk & chances
Thank you for your attention – any questions?

Michael Höhn
Ai Associates
mh@ai.associates

Christine Diegel
2911 Data Analytics
christine.diegel@LBBW.de

Kilian Retter
2911 Data Analytics
kilian.reter@LBBW.de