Data enrichment and model creation using text mining and other unstructured data

Dr. Andreas Nawroth
Head of Artificial Intelligence
German Data Science Days 2019
19. Feb 2019
Agenda for data enrichment and model creation using text mining and other unstructured data

1. Introduction to Analytics & Artificial Intelligence
2. Deep dive into Deep Learning approach
3. Our approach to industrialize text mining
4. Long term vision
5. Question & discussion
Introduction to Analytics & Artificial Intelligence
Munich Re actively shapes the transformation of the (re-)insurance industry

Efficiently running the traditional book while continuously exploring new products and markets

New trends & risks
- Digitisation
- Longevity
- Climate Change

MARKETS

PRODUCTS

ESTABLISHED

NEW

Incremental innovations

Emerging markets

New products for emerging risks

Traditional reinsurance business
Example: Ad-hoc risk identification and quantification
Example: Extraction of timeline of events
In recent years, there has been many applications of text mining
Deep dive into Deep Learning approach
What is Text Mining: unlock the value of data in unstructured files

Text Mining

- Unstructured data as input
  - WWW
  - Email
  - File

- Text Pre-processing
  - OCR
  - Tokenization
  - Word Embeddings

- Identify information
  - Named Entities
  - Relation
  - Sentiment

- Deliver result
  - Structured data in predefined format
  - Risk Indices

Structured data as output

Evolution of approaches

- Look-up methods
- Rule-based approach
- “Traditional” Machine Learning approach
- Deep Learning approach
One most used text mining techniques: Named Entity Recognition (NER), solves “Who exactly, When exactly, What exactly”?

Illustrative example

Munich Re is expecting a loss of €1.4bn due to Hurricanes Harvey, Irma and Maria, for the third-quarter in 2017.

Desired outcome

Who: Munich Re
What: loss of €1.4bn
When: third-quarter in 2017

Data enrichment with additional data source (wiki, etc.)
With Munich Re’s MIRA Digital Suite, life insurers are massively reducing the effort and expense involved in applications and claims. CLARA, for example, halves the average time taken to settle disability claims.

Republican lawmakers still think Google is biased against conservatives, Google still claims that it’s not. The news agency reports.
Fast innovations of text mining enable faster and more efficient info extraction for enhanced data quality and process automation

**Milestones of text mining**

- **2008**: NLP (almost) from scratch, Multi-task learning\(^1\)
- **2013**: Word embedding (faster)
- **2013**: Neural networks for NLP
- **2014**: Sequence-to-sequence framework
- **2015**: Attention
- **2015**: Memory-based networks
- **2018**: Pre-trained language models

---

**Enhanced efficiency**: With this framework, Google in year 2016 started to replace its monolithic phrase-based machine translation (MT) models with neural MT models (Wu et al., 2016), replacing 500,000 lines of phrase-based MT code with a 500-line neural network model.\(^2\)

**Reduced limitation**: Enables learning with significantly less data, only require unlabeled data.

---


---

- Mikolov et al., 2013a, b, c. Efficient Estimation of Word Representations in Vector Space. Distributed Representation of Words and Phrases and their Compositionality. Linguistic regularities in continuous space word representations.
TM framework captures recent AI innovations – text mining processes become more similar to human brain processes

Models evolved from rule-based to those represent human brains*

Deep learning neural network applied in text mining (illustrative)

* Based on illustration on wired.com
TM framework captures recent AI innovations – text mining processes become more similar to human brain processes

Models evolved from rule-based to those represent human brains*

Deep learning neural network applied in text mining (illustrative)

Character-level CNN (Convolutional Neural Network)

Word-level embedding (Glove / Word2vec)

* Based on illustration on wired.com
TM framework captures recent AI innovations – text mining processes become more similar to human brain processes

Models evolved from rule-based to those represent human brains*

Deep learning neural network applied in text mining (illustrative)

+ RNN (Recurrent Neural Network)

* Based on illustration on wired.com
TM framework captures recent AI innovations – text mining processes become more similar to human brain processes

Models evolved from rule-based to those represent human brains*
Deep Learning moves manual work complexity to model complexity

Deep learning offers a flexible framework to approximate complex & abstract tasks by automating the feature generation process.

Move manual work complexity to model complexity

<table>
<thead>
<tr>
<th></th>
<th>Data preparation</th>
<th>Feature engineering</th>
<th>Rules Creation</th>
<th>Output interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical programming</td>
<td>manual</td>
<td></td>
<td>manual</td>
<td></td>
</tr>
<tr>
<td>Machine Learning</td>
<td>manual</td>
<td>automated</td>
<td>manual</td>
<td></td>
</tr>
<tr>
<td>Deep Learning</td>
<td>manual</td>
<td>automated</td>
<td>manual</td>
<td></td>
</tr>
</tbody>
</table>

Universal approximation theorem

A feed-forward network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets of $\mathbb{R}^n$.

See tasks as functions

See complex tasks, involving structured or unstructured data as a function to be learned by the network.

Hierarchical abstraction

Neural network architectures typically process the data by adding complexity at each step of computation. This allows for modular re-use of models sharing similar core task.

Example: image models
Our approach to industrialize text mining
Scalable text mining platform: a flexible 3-tier platform, ensuring the optimal usage by internal and external business

- Clients can run core components as standard functionalities
- For client-specific needs, our text mining team is contracted to provide services (configuration and customization, incl. pipelines with multiple core components and data feeding)
- Clients can run PoC on platforms and evaluate performance and success of various approaches, including self-configured workflows
Annotation tool enables flexible and easy way for information capturing

- Basic info
- Loss dates
- Loss location
- Risk info
- Parties involved
- Experts
- Circumstances of loss
- Legal issue
- Financials
Application example: we use text mining to detect organizational name and supply chain info

Underwriters use those info for business interruption risk for companies.
Data Lake Integration: text mining module extracts four types of info to enrich the existing documents

Info extraction

Text Extractor

4 Entity Extractors
- Company
- Location
- Person
- Date

Table of Entities

Keyword Index
- Company
- Location
- Company & Location

Upload UI

pdf

Data Lake

Feedback UI (check for false-positives)
- Company
- Location
- (Person)
- (Date)

View PDF

Keyword-Search UI
Example: Find all PDF files in the Data Lake having the keywords “Munich” AND “BMW”
4 Long term vision
Demo: Knowledge Graph to better understand organization profile
Our goal: enable access to insurance-specific and the most state-of-the-art text mining technology for both internal & external clients

Text mining at finger tips: everyone (data scientists or not) at Munich Re can run text mining for their everyday work, conduct experiments and PoCs with various AI approaches and models

Augmented underwriting: aided by text mining, underwriters analyze the text data from client on near real-time basis and provide immediate response to client requests

Client could access text mining modules through fast API

Packaged applications or customized use cases for clients
Demo: Deep learning approach applied for images
Question & discussion

Please feel free to contact me for further details!

Dr. Andreas Nawroth (Anawroth @ munichre.com)