Technical Language Processing: 
Putting technical data to work in industrial applications

Sarah Lukens, Ph.D.
GE Digital

15 September 2021
Large amount of text data which cannot be directly used by most analytics

How can we bridge the gap and dig out useful information from industrial text data?
Technical language is characterized by…

- domain-specific jargon
- context
- implied knowledge
- non-standardized language
- missing/inaccurate/ambiguous data
- lack of labels

“Small data”

<table>
<thead>
<tr>
<th>#</th>
<th>Work Request Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CL3-012 bfp not working also bad seal</td>
</tr>
<tr>
<td>2</td>
<td>block valve leaking</td>
</tr>
<tr>
<td>3</td>
<td>Monthly inspection by Joe</td>
</tr>
</tbody>
</table>

Explicit Knowledge
Codified or digitalized knowledge found in data, documents, records or files

Tacit Knowledge
Intuitive knowledge & ”know-how” rooted in experience, context, practice, values

How do models pre-trained on Wikipedia, news articles, or Tweets behave when applied to work order descriptions?
Core NLP concepts – NLP pipelines and out-of-the-box challenges

Raw data → Pre-processing → Representation → Analysis task

Pre-processing:
- Cleans up text such as lowercase, punctuation removal, etc.

Representation:
- Converts text to form which can be used by a numerical algorithm

Analysis task:
- Numerical machine learning algorithm such as neural networks or random forest

Examples:

<table>
<thead>
<tr>
<th>Description</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pump not working</td>
<td>Pre-processing: Stop word removal - OOTB tools may remove the word “not”</td>
</tr>
<tr>
<td>leakage in the CO2 vlv</td>
<td>Pre-processing: Character removal may remove information such as technical abbreviations</td>
</tr>
<tr>
<td>leakage in the CO2 vlv</td>
<td>Semantic meaning: OOTB may not link technical concepts such as “valve” and ”vlv”</td>
</tr>
<tr>
<td>Pmp-01 Work Request per proc 343</td>
<td>Missing context: meaning dependent on knowing procedure 343</td>
</tr>
</tbody>
</table>
Making TLP a reality

A. Use case is explicitly considered as an input
B. Use NLP resources where it makes sense
C. Use computer tools to alleviate burden on domain expert
D. Collaboration between analyst and domain expert
E. TLP resources – such as fortuitous data

Making TLP a reality

A. Use case is explicitly considered as an input

B. Use NLP resources where it makes sense

C. Use computer tools to alleviate burden on domain expert

D. Collaboration between analyst and domain expert

E. TLP resources – such as fortuitous data

Use case #1: Enables consistent reliability metrics

Before: Inability to calculate Mean Time Between Failure (MTBF)

After: Benchmarking comparison of MTBF is possible

### Comparison of MTBF (days)

- **Company 1**: 470 days
- **Company 2**: 314 days

<table>
<thead>
<tr>
<th>Description</th>
<th>Before: Breakdown Indicator</th>
<th>After: Is a Failure event?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seal is leaking badly</td>
<td>FALSE</td>
<td>True</td>
</tr>
<tr>
<td>Block valve is broken open and inoperable</td>
<td>FALSE</td>
<td>True</td>
</tr>
<tr>
<td>00120-Pump 1 work request</td>
<td>FALSE</td>
<td>False</td>
</tr>
<tr>
<td>Check impeller size</td>
<td>FALSE</td>
<td>False</td>
</tr>
</tbody>
</table>

Use case #2: Finding hidden bad actors

Root cause analysis revealed no proper procedures in place for motor maintenance. Once remedied, failures ceased with estimating annual savings of $54,000 for the one pump alone.

Bad Actors List – by total maintenance cost over 2 years

Pump-17 has 9 repairs from the motor starter failing, all causing a break in work.

<table>
<thead>
<tr>
<th>Date</th>
<th>Description</th>
<th>Failure mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 14</td>
<td>M-17 won’t start</td>
<td>Unknown</td>
</tr>
<tr>
<td>Sept 18</td>
<td>Motor won’t start</td>
<td>Unknown</td>
</tr>
<tr>
<td>Nov 6</td>
<td>Starter button not working</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Failure mode characterization can be used for reliability-data based survival models such as Weibull analysis.

Use case #3: Enables reliability distribution fitting

Dog-leg shape – mixed failure mode information inappropriate for statistical modeling.

Weibull distribution for valve leakage failures on compressors.
Reliability distribution fitting can be used for decision making

Repair or replace? Simulation model:

Model 1 is more reliability but costs more than Model 2

Putting industrial data to work with TLP

TLP enables industrial companies to make decisions using their historical data. The value of TLP comes with integrating the data into business processes in ways that create value.

Thank you!

Contact me: sarah.lukens@ge.com
Making TLP a reality

Data Representations & Feature Engineering
• Need for use case-driven representations
• Need for computational support tools to assist in annotation of most-used concepts

Entity Types Definitions and Dictionaries
• No wide-spread community consensus or adoption exists for agreed on entity sets or hierarchies in maintenance.

Raw and Annotated datasets
• Industry perception that datasets are valuable intellectual property.

Verification & Validation approaches
• High fidelity requirements in industrial applications