NLP in practice

Risto Hinno

risto@feelingstream.com

Chief Feeling Scientist

29.05.2019
What do we do?
Who is the most important client?
Every business is leaking

Customers are leaving
Opportunities are lost
Sales is not executed
Cross channel business leakage monitoring

Signals

- Lost opportunities
- Sales potential
- Churning customers

Customer interactions

- Email conversation
- Phone calls
- CRM comments
- Chat
- Chatbot
- Customer feedback
- Office conversations
- Online meetings
- Support tickets
RESULT:
Routing with high accuracy, business critical cases (urgent queue) were answered 96% faster.
Prioritisation

Contact Centre platform:

Incoming customer case

Case content

Decision

Sales

Regular

Feelingstream Decision Model

Reference case:

Nordea
Customers
How do we do it?
Typical value creation process

1. Business need/problem
2. Data collection
3. Annotating
4. Modelling
5. Live implementation

KEEP CALM AND GO LIVE
Data science tasks

- Text categorization
- Topic modelling
- Text language detection
- Text anonymization:
  - NER
- Voice2text
- Voice feature analysis:
  - cross-talk
  - speed of speaking
  - spoken language detection
- etc.

### Statistical vs business error

<table>
<thead>
<tr>
<th></th>
<th>Statistical error</th>
<th>Business error</th>
</tr>
</thead>
<tbody>
<tr>
<td>real</td>
<td>⅖ = 40%</td>
<td>⅕ = 20%</td>
</tr>
<tr>
<td>predicted</td>
<td>This is a demo sentence</td>
<td>This is aa demos sentence</td>
</tr>
</tbody>
</table>

- Statistical error: ⅖ = 40%
- Business error: ⅕ = 20%
Models used in text mining

- SVM
- FastText
- Latent Dirichlet allocation
- Naive Bayes
- Bagging, stacking, boosting etc.

LDA

Source: https://m-cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext?mobile=true
Transfer-learning


Source: [https://towardsdatascience.com/transfer-learning-using-differential-learning-rates-638455797f00](https://towardsdatascience.com/transfer-learning-using-differential-learning-rates-638455797f00)
Daily challenges
Context Matters

Source: https://adexchanger.com/comic-strip/adexchanger-context-matters/
Business and data context matters

● But you have to figure it out yourself:
  ○ is there something missing
  ○ is there something that client avoids telling about
  ○ meaning of words

Maturity level

Source: https://taraskaduk.com/2018/01/22/4-stages/
Ways to gather more information

- Business analysis techniques
  - interviews, meetings
  - observations
  - mystery-shopping
- Metadata analysis
- Descriptive analysis
- Contacts inside organization
- The Data is the Specification:
  https://dataisspec.github.io/

"Based on my observations, this is how the Business Analysts do their job."
Be quick and accurate

- Simple models
- Software engineering practices
- Platform
  - https://www.feelingstream.com/demo/
- Good quality training data
Human vs machine

- human error rate 30% -> ok
- machine error rate 10% -> not ok

Source: https://cdn-images-1.medium.com/max/1116/1*B_RYlyCm_oOfNXB76zz6Hw.png
Coding quality

Source: https://medium.com/@navdeepsingh_2336/good-code-vs-bad-code-35624b4e91bc
Organizing work

- [https://github.com/audreyr/cookiecutter](https://github.com/audreyr/cookiecutter)
- reviews
- naming conventions

Source: [https://commons.wikimedia.org/wiki/File:Messy_storage_room_with_boxes.jpg](https://commons.wikimedia.org/wiki/File:Messy_storage_room_with_boxes.jpg)
Quality of annotation

The chart shows the distribution of annotations across different categories:
- **Neutral**
  - Neutral: 88
  - Positive: 38
  - Negative: 118
- **Positive**
  - Neutral: 17
  - Positive: 136
  - Negative: 1
- **Negative**
  - Neutral: 228
  - Positive: 2
  - Negative: 445
Annotation matters

![Box plot showing accuracies for different numbers of annotators](image-url)
Annotation task

- Clear rules
- Test task
- Data always surprises
- At least 3 opinions per text
  - where to find people who do it fast
- Minimum number of labels to choose from
Other challenges

- Every day new data
- Language understanding
- Explain complex stuff as simple as possible

Source: https://twitter.com/nijfranck/status/977277874460938242
Data Scientist
# Data Scientist

<table>
<thead>
<tr>
<th>What society thinks I do</th>
<th>What my friends think I do</th>
<th>What my parents think I do</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://source.com" alt="Image" /></td>
<td><img src="https://source.com" alt="Image" /></td>
<td><img src="https://source.com" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What my coworkers think I do</th>
<th>What I think I do</th>
<th>What I really do</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://source.com" alt="Image" /></td>
<td><img src="https://source.com" alt="Image" /></td>
<td><img src="https://source.com" alt="Image" /></td>
</tr>
</tbody>
</table>

Source: [https://medium.com/indeed-engineering/theres-no-such-thing-as-a-data-scientist-8dae923c14e3](https://medium.com/indeed-engineering/theres-no-such-thing-as-a-data-scientist-8dae923c14e3)
What data scientists spend the most time doing

- Cleaning and Organizing Data: 60%
- Collecting Data Sets: 19%
- Refining Algorithms: 4%
- Other: 5%
- Modeling / Machine Learning: 9%
- Building Training Sets: 3%

We spend 80% on prep

Other stuff

● **Business requirements:**
  ○ what is the real problem?
  ○ what creates the most value for customer?
  ○ what is business context?

● **Timeline**

● **Training people**

● **Pitching, presenting results**

Source:
https://medium.com/analytics-vidhya/25-fun-questions-for-a-machine-learning-interview-373b744a4faa
Life is like a stream of emotions.