Ratemaking, Product and Modeling Seminar and Workshops

March 15–17, 2021
Virtual Conference
Antitrust Notice

• The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

• Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

• It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.
Mining for Gold: Text Analytics in Insurance

Liam McGrath, ACAS, Willis Towers Watson
Yelena Kropivnitskaya, PhD, The Wawanesa Mutual Insurance Company
Agenda

- Natural Language Processing – why should you care?
- How do we structure unstructured data?
- Text mining for feature engineering
- Case studies
  - Underwriting
  - Claims – Independent Medical Examination
  - Claims – At-fault rating
- Conclusions
Sources of unstructured text data

The insurer’s goldmine

- Social media
- Call center voice logs
- Review websites
- Loss control surveys
- Underwriting reports
- Business news
- Claims handler notes
- Accident interviews
- Legal documents

Marketing and Distribution
Customer Management
Underwriting and Risk Management
Pricing
Claims Management
Asset Management
Tapping into these sources allows us to...

Make the most of data we already have
Insurers already collect various types of text data through normal business operations. Text mining ensures it isn’t collecting dust!

Fill gaps in structured data
Structured data has limitations. Text data can provide more nuance to fill in gaps.

Quantify internal knowledge consistently
Machine learning can quantify the implicit knowledge of adjusters and underwriters while also smoothing out the natural variation of human decision makers.
Expected benefits have yet to be realized

### Personal Lines

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Expected for 2019 (in 2017)</th>
<th>Actual for 2019</th>
<th>Expected for 2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured internal claim information</td>
<td>66%</td>
<td>38%</td>
<td>69%</td>
</tr>
<tr>
<td>Unstructured internal underwriting info</td>
<td>50%</td>
<td>18%</td>
<td>67%</td>
</tr>
</tbody>
</table>

### Commercial Lines

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Expected for 2019 (in 2017)</th>
<th>Actual for 2019</th>
<th>Expected for 2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured internal claim information</td>
<td>91%</td>
<td>53%</td>
<td>81%</td>
</tr>
<tr>
<td>Unstructured internal underwriting info</td>
<td>63%</td>
<td>16%</td>
<td>66%</td>
</tr>
</tbody>
</table>

Natural Language Processing will allow us to extract much more out of unstructured data
Example: a claim note

PC to Jane Doe/insd: DOI: 01/01/16 Clmt was carrying drywall up steps with a co-worker. When the co-worker reached the top of the stair steps, he started to walk faster, causing clmt to twist his back and strain his R/shoulder. C/o pain in mid-back & R/shoulder Incident witnessed by co-worker, John Smith Clmt did not report back sprain injury to his supervisor until D/L NLT...

- Problems with unstructured data
  - Junk words, numbers, and formatting
  - Many meanings for a word (polysemy)
  - Many words with the same meaning (synonymy)
  - Negation
  - Abbreviations and acronyms
Text processing

- Lower case transformation
- Tokenization
- Token filtering by length
- Word mappings
- Stop word removal
- Stemming/lemmatization
- N-Gram generation
- Feature generation

Example: Tokenization of "it not cool that ping pong is not included in rio 2016" results in "it not cool that ping pong is not included in rio 2016".
Text processing

Addresses the problem of abbreviations and acronyms

rtw  return to work
Text processing

Lower case transformation → Tokenization → Token filtering by length → Word mappings → Stop word removal → Stemming/lemmatization → N-Gram generation → Feature generation

Addresses the problem of... **Junk words**
Text processing

- Lower case transformation
- Tokenization
- Token filtering by length
- Word mappings
- Stop word removal
- Stemming/lemmatization
- N-Gram generation
- Feature generation

Addresses the problem of...

Synonymy

- adjustable → adjust
- formality → formaliti
- was → (to) be
- better → good
Text processing

Addresses the problem of polysemy:

- twisted back (2-gram)
- back to work (3-gram)
Text processing

- Lower case transformation
- Tokenization
- Token filtering by length
- Word mappings
- Stop word removal
- Stemming/lemmatization
- N-Gram generation

Significant effort
Domain expertise required

Feature generation
Feature engineering

- **Word Indicators**
  - Binary variable representing the presence of a word

- **Sentiment Analysis**
  - Measures the valence of a document including simple lexicon mapping

- **Topic Models**
  - Detects topics (or themes) in text that are composed of multiple words
  - The topics and words are expressed in terms of probabilities

- **Word Embedding**
  - Translates each term or phrase into a vector in a lower dimension space
  - Words with similar context are in close proximity to each other

- **Transformers**
  - Develops embeddings that account for longer term dependencies
  - Useful for various tasks, like classification and named entity recognition

**Complexity**
Case study: Commercial lines underwriting

Goal
• Segment risks using underwriting reports

Features
• Structured fields typically used for rating and underwriting: policy details, exposure information, loss history, 3rd party data
• Underwriting reports, loss descriptions, and loss control surveys
  • Topic Modeling
Topic Models provide context

### Word Indicators

<table>
<thead>
<tr>
<th></th>
<th>surgery</th>
<th>claimant</th>
<th>cactus</th>
<th>back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claim 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Claim 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Weaknesses:
- Many meanings for a word (polysemy)
- Many words with the same meaning (synonymy)

### Topic Models

<table>
<thead>
<tr>
<th></th>
<th>return</th>
<th>duty</th>
<th>work</th>
<th>back</th>
<th>full-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 2</td>
<td>back</td>
<td>strain</td>
<td>disc</td>
<td>neck</td>
<td>sprain</td>
</tr>
</tbody>
</table>
Big picture of topic modeling

• Goal of topic modeling is to discover the **hidden thematic structure** in a large set of documents using posterior inference

• Documents are assumed to exhibit traits from multiple topics with **different topic proportions**, i.e., *mixed-membership model*

• Topic modeling:
  • Automates the annotation of a set of documents
  • Does not require any prior annotation or labeling of documents, i.e., unsupervised

• Topic modeling represents a **core idea with many different versions**
  • Like Regression, different versions include OLS, GLM, Ridge, Lasso, and Elastic Nets
  • Like CART, different versions include Gradient Boosting and Random Forests

*Latent Dirichlet Allocation*, D. Blei et al. 2003
What is a topic?

• A topic is a probability distribution over a fixed vocabulary

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>claim</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>arm</td>
<td>0.30</td>
<td>0.01</td>
</tr>
<tr>
<td>leg</td>
<td>0.01</td>
<td>0.40</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• We can understand a topic by examining its most likely words

<table>
<thead>
<tr>
<th>Topic</th>
<th>laceration</th>
<th>sutures</th>
<th>removal</th>
<th>hospital</th>
<th>stitches</th>
<th>feet</th>
<th>issued</th>
<th>wound</th>
<th>complete</th>
<th>injuring</th>
</tr>
</thead>
</table>

Willis Towers Watson
What do topics tell us about a document?

Example topics:

Topic 1 Top Words:
farm, crops, tractors, harvesting, acres, plant, grow

Topic 2 Top Words:
drug, testing, checks, required, employment, mvr, physical

Topic 3 Top Words:
feed, fertilizer, elevator, bins, farmers, seed, mill

Topic 4 Top Words:
mold, castings, sand, aluminum, foundry, pour, cooling

Topic 5 Top Words:
walls, masonry, structural, retaining, waterproofing, dry, basement

Topic 1: 40%
Topic 2: 30%
Topic 3: 30%
Topic 4: 0%
Topic 5: 0%
Topics are powerful predictors

Feature Importance*

<table>
<thead>
<tr>
<th>Class</th>
<th>Hazard Group</th>
<th>Topic features</th>
<th>Neural sentiment</th>
<th>Geodemographic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* - approximation across multiple models, normalized by Class

Sample topics

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>wear</td>
<td>safety</td>
<td>construction</td>
</tr>
<tr>
<td>required</td>
<td>training</td>
<td>residential</td>
</tr>
<tr>
<td>ppe</td>
<td>safety_program</td>
<td>carpentry</td>
</tr>
<tr>
<td>glasses</td>
<td>safety_meetings</td>
<td>framing</td>
</tr>
<tr>
<td>safety</td>
<td>documented</td>
<td>remodeling</td>
</tr>
<tr>
<td>hats</td>
<td>osha</td>
<td>plumbing</td>
</tr>
<tr>
<td>gloves</td>
<td>formal_safety</td>
<td>renovation</td>
</tr>
<tr>
<td>safety_glasses</td>
<td>safety_training</td>
<td>siding</td>
</tr>
<tr>
<td>boots</td>
<td>written_safety</td>
<td>hvac</td>
</tr>
<tr>
<td>hard_hats</td>
<td>certified</td>
<td>subcontract</td>
</tr>
</tbody>
</table>

Willis Towers Watson
Case Study – Independent Medical Exams

• An independent medical examination (IME) helps us to:
  • Determine the cause of injury associated with the incident
  • Evaluate the claimant condition and medical treatment
  • Mitigate risk of injury deterioration

• IMEs are paid by the insurance company in addition to the medical and rehabilitation costs
  • We want to better understand our current spend
  • We want to optimize IME spend and order only when it is necessary to minimize premium impact for all policyholders
IME- Solutions

Structured Data Analysis → Geographic Analysis → Text Analytics

Internal and External Structured Data → Claim Notes
Word Embeddings – Word2Vec

- Words that have the similar meaning have a similar representation
  - Words > real-valued vectors
  - Predefined vector space: tens or hundreds of dimensions
Embeddings - BlazingText

Learned embeddings

Text classifier

INPUT  PROJECTION  OUTPUT

\[ w(t) \rightarrow w(t-1) \rightarrow w(t-2) \]

\[ w(t+1) \rightarrow w(t+2) \]

Softmax

Hidden Layer (Average Pooling)

Embedding vectors

\[ w_1 \rightarrow w_2 \rightarrow w_3 \rightarrow w_{n-1} \rightarrow w_n \]
Embeddings - Word to Numbers
IME – Text Analytics: Classifying Text

- Low probability of IME
- High probability of IME

Local Interpretable Model-agnostic Explanations (LIME): https://github.com/marcotcr/lime
Case Study – Who’s at Fault?

• In a car accident: who is responsible?
  • We have a structured data column to record fault rating
    • manual entry
    • low quality
  • Can text analytics do better?

• We have subrogation models to predict:
  • Who’s insurer should pay for the damages?
At Fault Rating Discrepancy

ClaimCenter - **structured** fields:

- **Fault**: Insured not at fault
- **Insured's Liability %**: 0

ClaimCenter – **unstructured** notes:

```
02:20 PM

Reviewed claim and the statements from both parties and the estimate advised would appear hit dead center front end therefore TP would of been directly infront of our insured therefore i have gave waive and send for 100% with our insured being held liable TPA advised there insured's vehicle is a TL
```
Bidirectional Encoder Representations from Transformers (BERT)

- State of the art transformer-based Natural Language Processing models
- Used in Google search engine
- Good compromise between performance and complexity
- Considers words in the context of the whole sentence

Example – “bank”:
BlazingText – same representation for “bank deposit”, “river bank”
BERT – representation depends on the entire sentence
Bidirectional Encoder Representations from Transformers (BERT)

Pre-train: large text corpus

Text Classifier: Fine-tune with your data

Had a science lecture at DSMeetUp

BERT– Attention Patterns

“under IBC we will be 100 at fault”

“under IBC we will not be at fault”

BertViz tool https://github.com/jessevig/bertviz
BERT – Challenges

• Lessons learned
  • Volume of data – Spark, Hadoop Distributed File System (HDFS)
  • Length of text - Sliding window
  • Interpretability

• Future work
  • Adding insurance-specific vocabulary
  • Better target label? Semi-supervised?
InterBERTability

Generate a recommended correction that looks like this:

<table>
<thead>
<tr>
<th>Claim Number</th>
<th>Insured at Fault</th>
<th>Recommended Insured at Fault</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>94%</td>
</tr>
</tbody>
</table>

Recommendation Explanation:

unclear of the color of the light the decision was reached in favor with royal with cooperators being 100 at fault left turning vehicle was held at fault with no witnesses to the color of the light ccdoclink 23211544 apd emailed insured going over liability decision and how we will be held 100 liable for the loss included auto liability letter in the email ....

Transformers Library - Huggingface

• Deep interoperability with TensorFlow and PyTorch
• Over 32+ pretrained models in 100+ languages
  • BERT
  • GPT-2
  • RoBERTa
  • XLM
  • DistilBert
  • XLNet
  • CTRL
  • ...

https://github.com/huggingface/transformers
Conclusions

• Variety of techniques
  • Basic to very complex
  • Quickly growing field

• Value of Natural Language Processing
  • Provide insights
  • Augment structured data
  • Make better predictions
Thank you