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

Marketing and Distribution → Customer Management → Underwriting and Risk Management → Pricing → Claims Management → Asset Management

Marketing and Distribution

- Social media

Customer Management

- Call center voice logs
- Review websites

Underwriting and Risk Management

- Loss control surveys
- Underwriting reports
- Business news

Claims Management

- Claims handler notes
- Accident interviews
- Legal documents

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

	Expected for 2019 (in 2017)	Actual for 2019	Expected for 2021
Unstructured internal claim information	66%	38%	69%
Unstructured internal underwriting information	50%	18%	67%

Commercial Lines

	Expected for 2019 (in 2017)	Actual for 2019	Expected for 2021
Unstructured internal claim information	91%	53%	81%
Unstructured internal underwriting information	63%	16%	66%

Natural Language Processing will allow us to extract much more out of unstructured data

2019/2020 P&C Advanced Analytics Survey, Willis Towers Watson

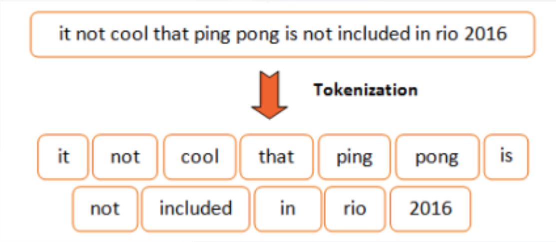
Example: a claim note

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

1. Lower case transformation
2. **Tokenization**
3. Token filtering by length
4. Word mappings
5. Stop word removal
6. Stemming/ lemmatization
7. N-Gram generation
8. Feature generation



Willis Towers Watson

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… **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**

Willis Towers Watson

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

Willis Towers Watson

Text processing

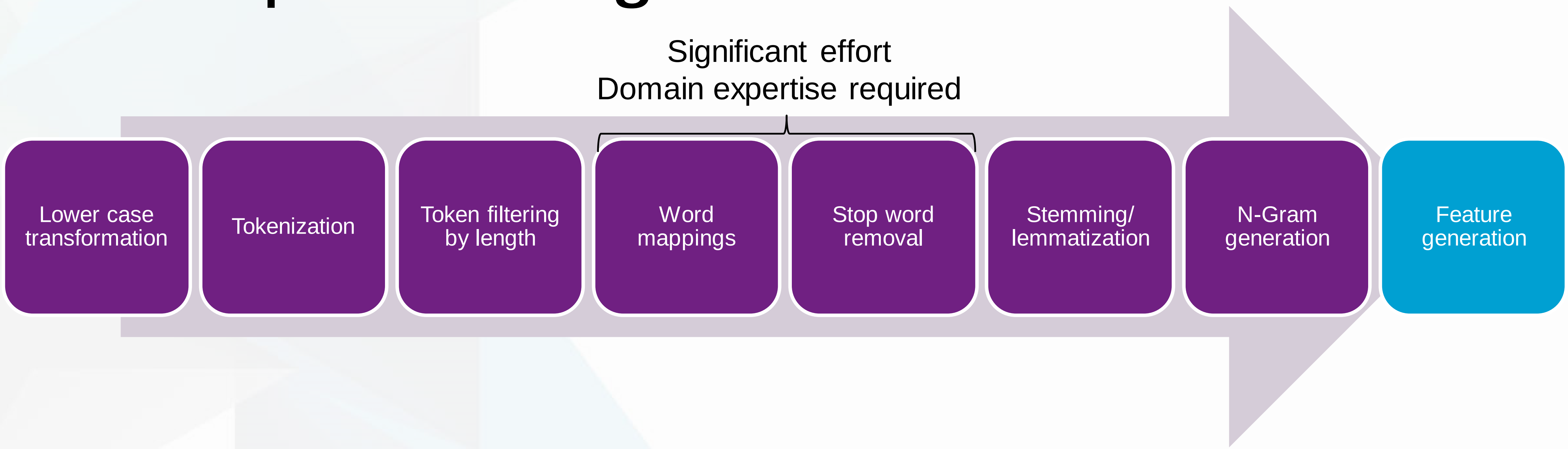
Lower case transformation → Tokenization → Token filtering by length → Word mappings → Stop word removal → Stemming/ lemmatization → **N-Gram generation** → Feature generation

Addresses the problem of… **Polysemy**

twisted back — 2-gram

back to work — 3-gram

Text processing



Feature engineering

Complexity

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

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

	surgery	claimant	cactus	back
Claim 1	0	1	0	1
Claim 2	1	1	0	0

Weaknesses:

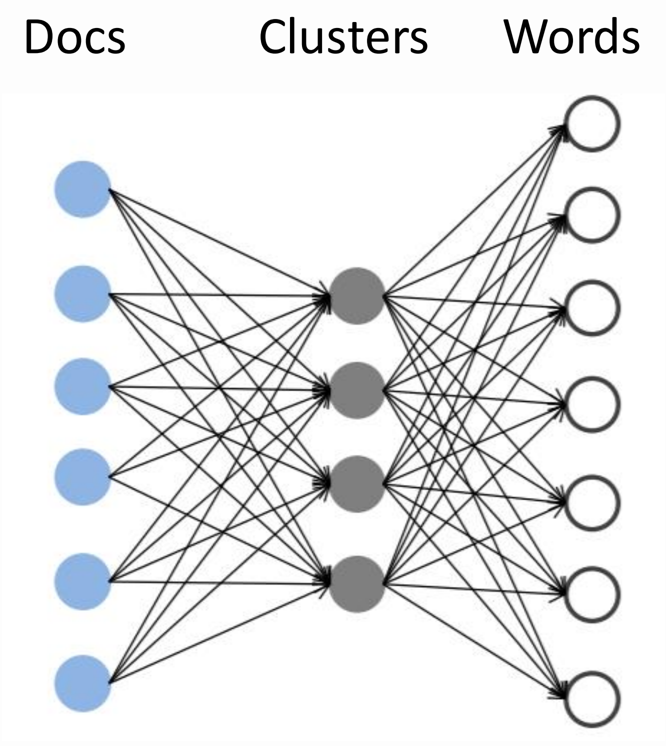
- Many meanings for a word (polysemy)
- Many words with the same meaning (synonymy)

Topic Models

Topic 1	return	duty	work	**back**	full-time
Topic 2	**back**	**strain**	disc	neck	**sprain**

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

	Topic 1	Topic 2
claim	0.05	0.05
arm	0.30	0.01
leg	0.01	0.40
…	…	…

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

Topic	laceration	sutures	removal	hospital	stitches	feet	issued	wound	complete	injuring

What do topics tell us about a document?

Example topics:

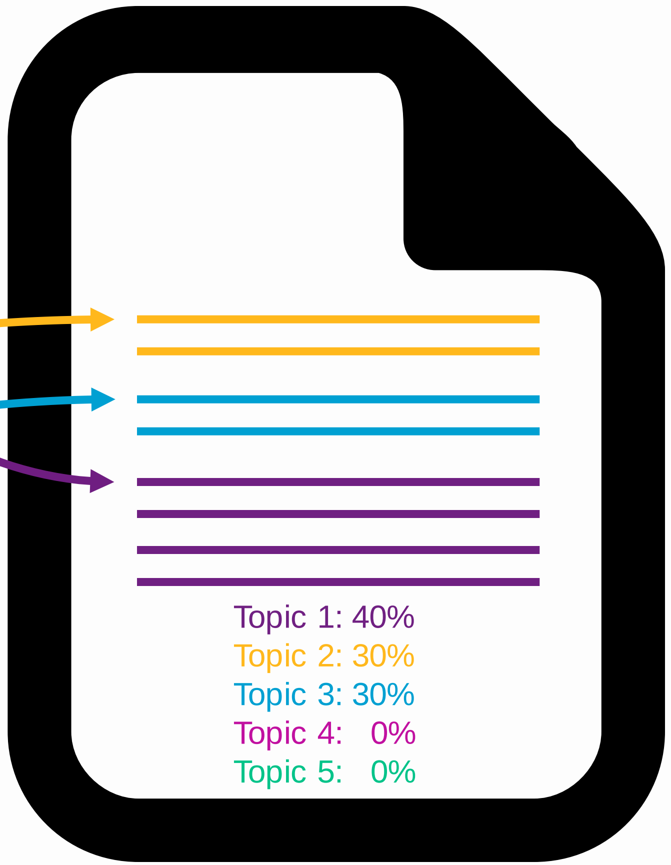
Topic 1 Top Words:
farm, crops, tractors, harvesting, acres, plant, grow

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

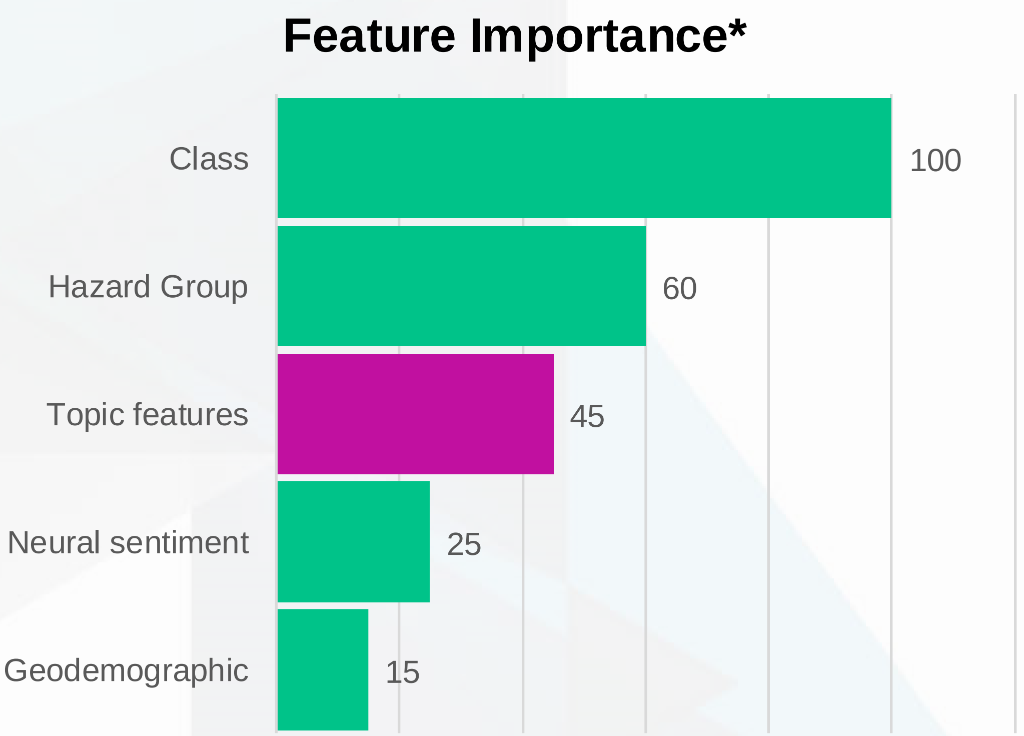
Topic 3 Top Words:
feed, fertilizer, elevator, bins, farmers, seed, mill

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

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



Topics are powerful predictors



Sample topics

Topic 1	Topic 2	Topic 3
wear	safety	construction
required	training	residential
ppe	safety_program	carpentry
glasses	safety_meetings	framing
safety	documented	remodeling
hats	osha	plumbing
gloves	formal_safety	renovation
safety_glasses	safety_training	siding
boots	written_safety	hvac
hard_hats	certified	subcontract

* - approximation across multiple models, normalized by Class

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

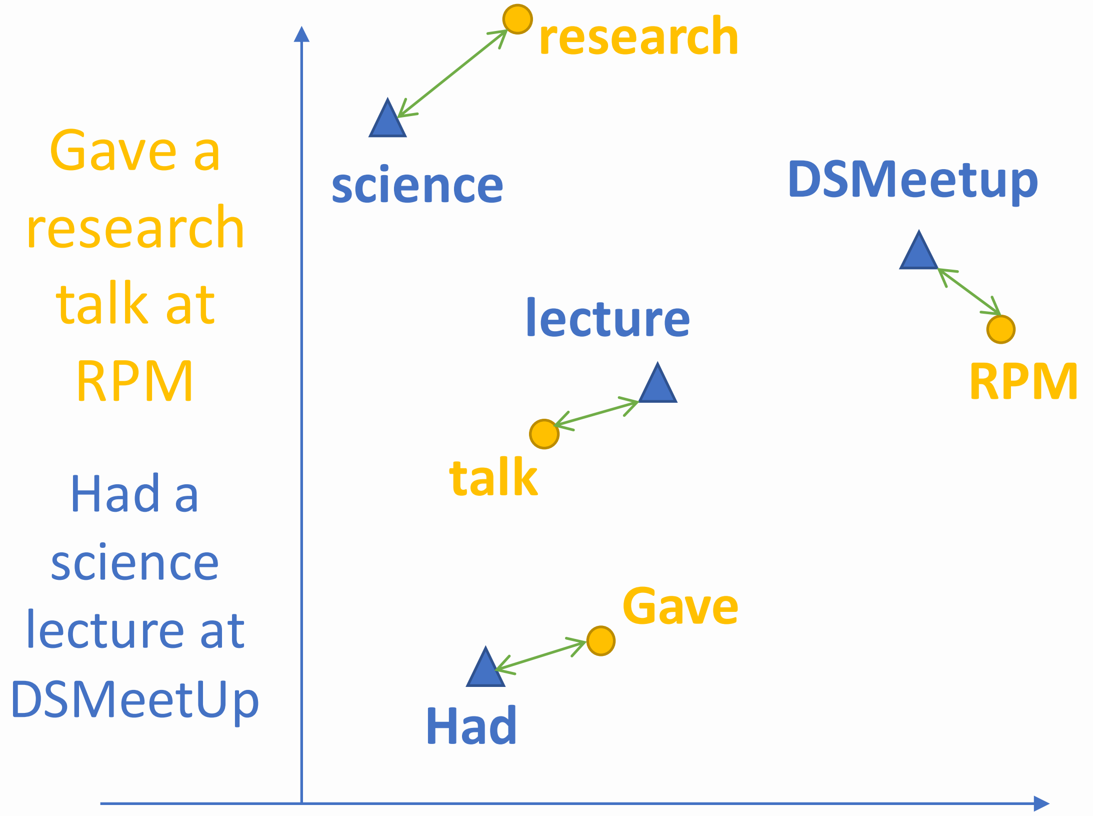
Structured Data Analysis → Geographic Analysis → Text Analytics

Internal and External Structured Data

Claim Notes

Word Embeddings – Word2Vec

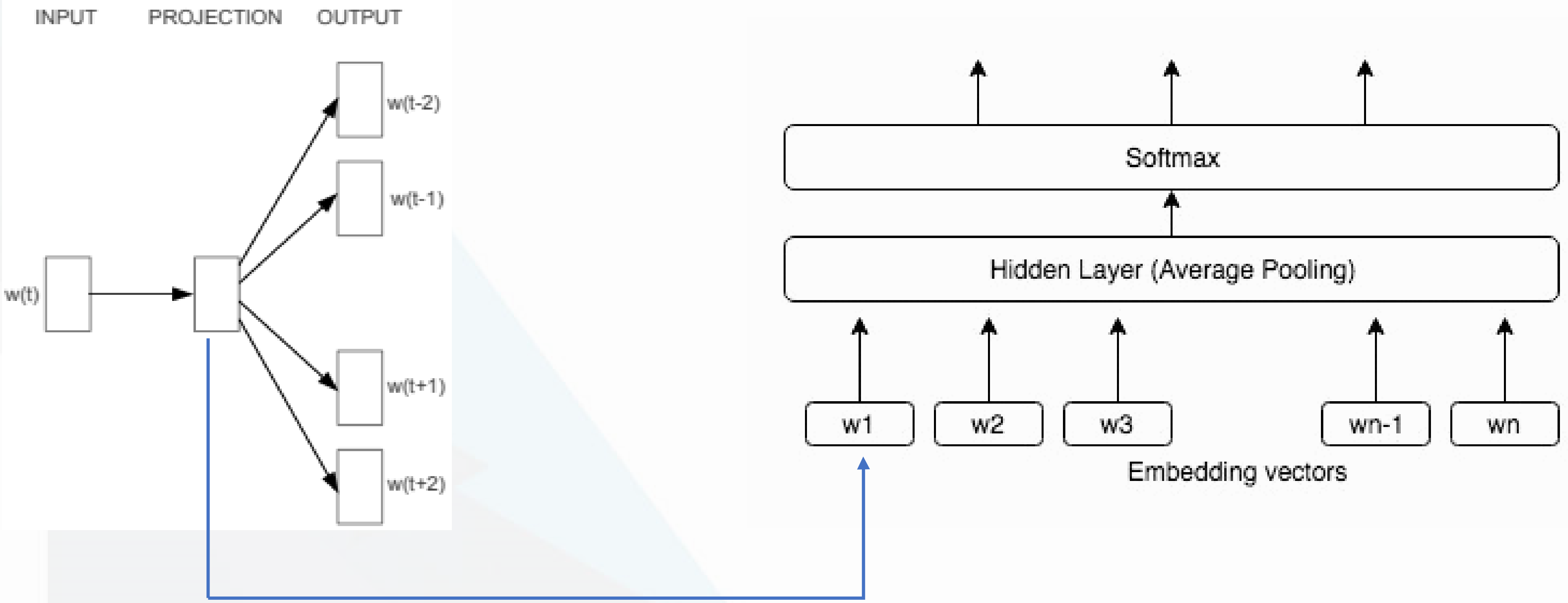
- Words that have the similar meaning have a similar representation
  - Words > real-valued vectors
  - Predefined vector space: tens or hundreds of dimensions



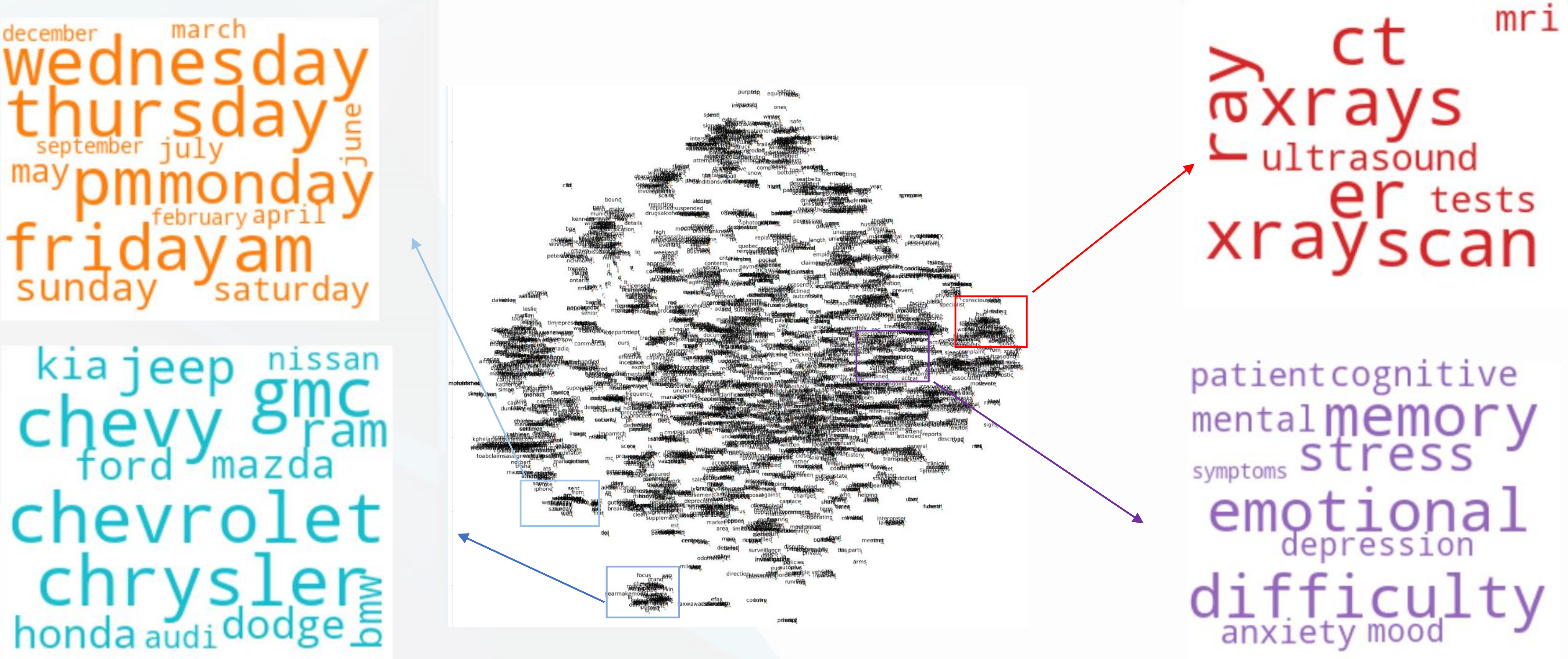
Embeddings - BlazingText

Learned embeddings

Text classifier



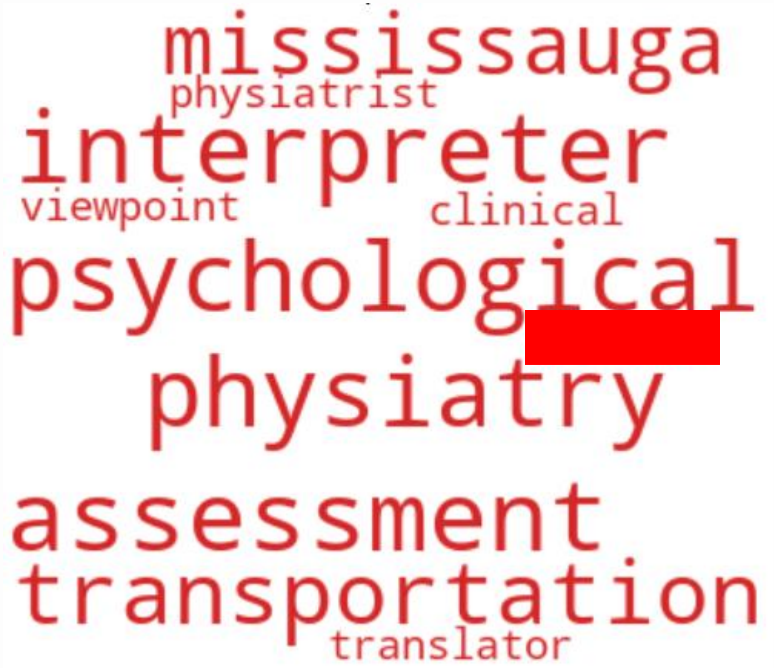
Embeddings - Word to Numbers



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:

General

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

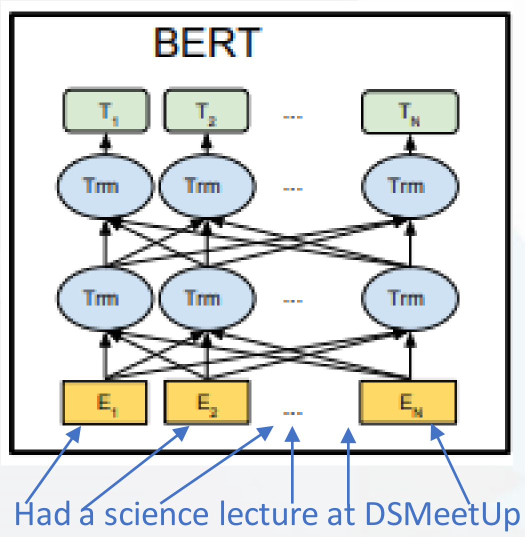
<u>Example – “bank”:</u>

BlazingText – same representation for “bank deposit”, “river bank”

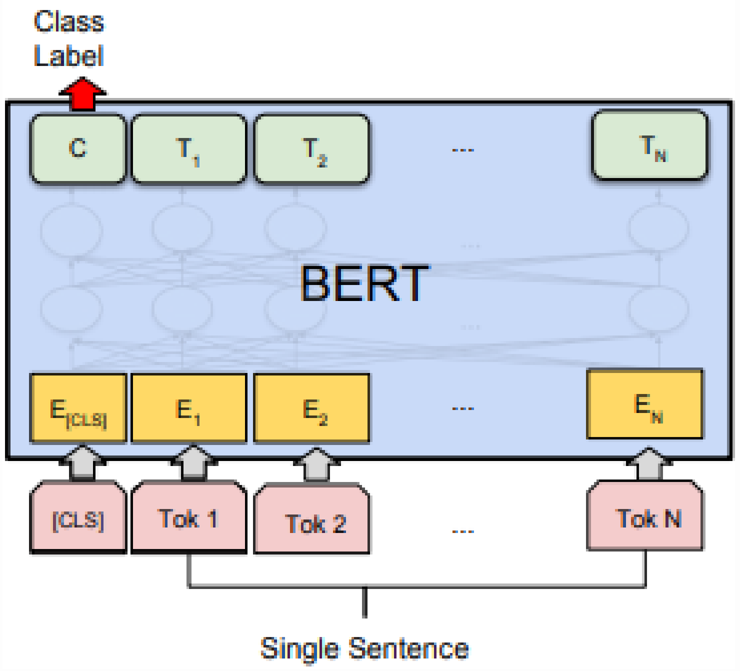
BERT – representation depends on the entire sentence

Bidirectional Encoder Representations from Transformers (BERT)

Pre-train: large text corpus



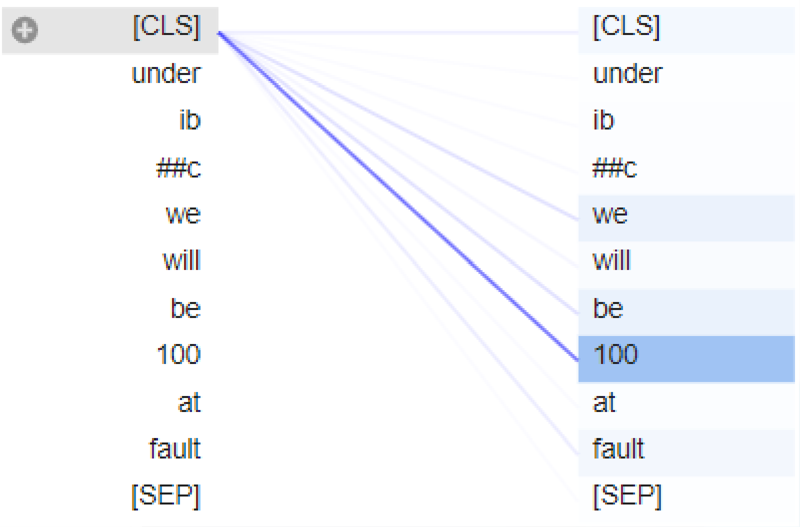
Text Classifier: Fine-tune with your data



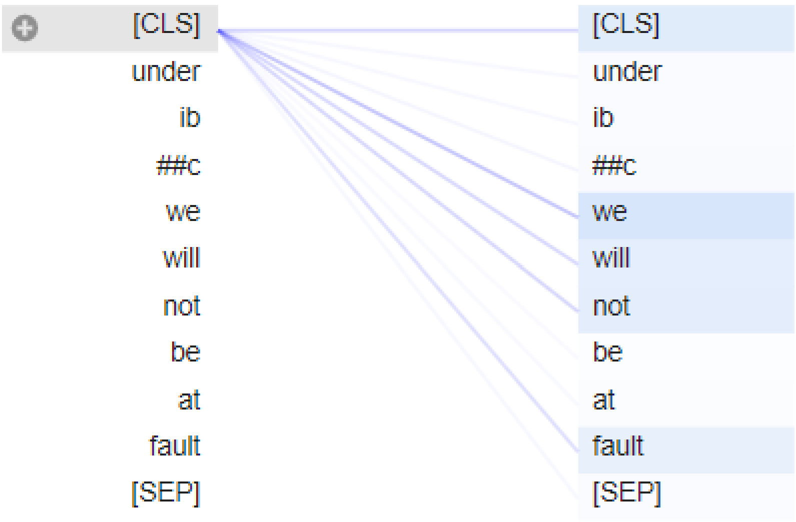
Diagrams from: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et. al, NAACL-HLT 2019

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:

Claim Number	Insured at Fault	Recommended Insured at Fault	Recommendation Explanation
[redacted]	0%	94%	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 ....

Captum library (https://github.com/pytorch/captum).

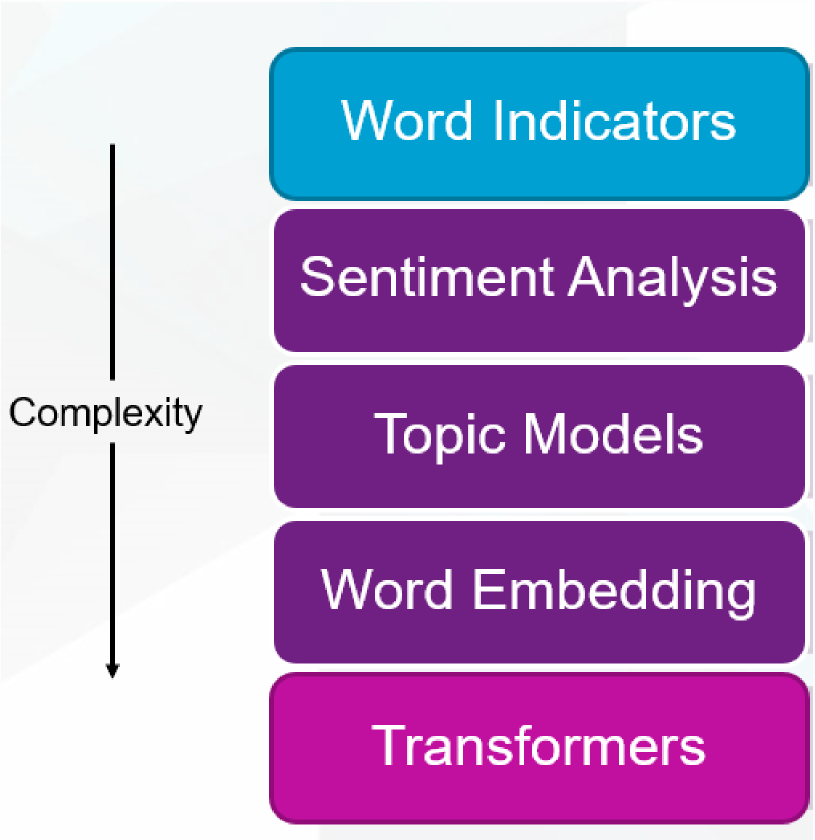
Transformers Library - Huggingface

- 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