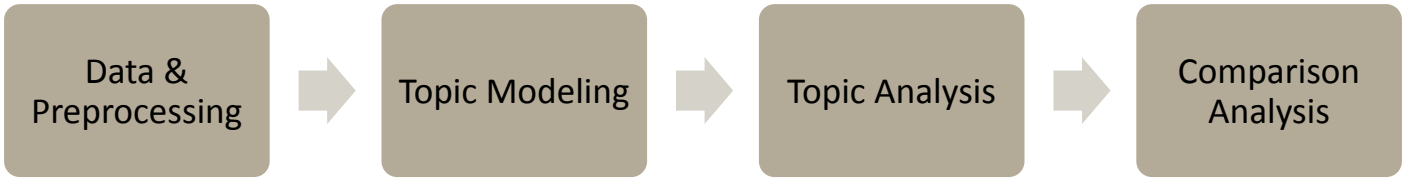
Standards & Engagement

BERTopic Modeling on Narratives

Data Science Team
Hanjun Wei

Introduction: Why Narratives?

- Why are narratives important?
  - Intuitive: cheap to produce and collect
  - Flexible: better preserve information in context
- Drawback of narratives in the era of data explosion
  - Hard to analyze in a large scale
- Introducing BERTopic and analyzing its effectiveness

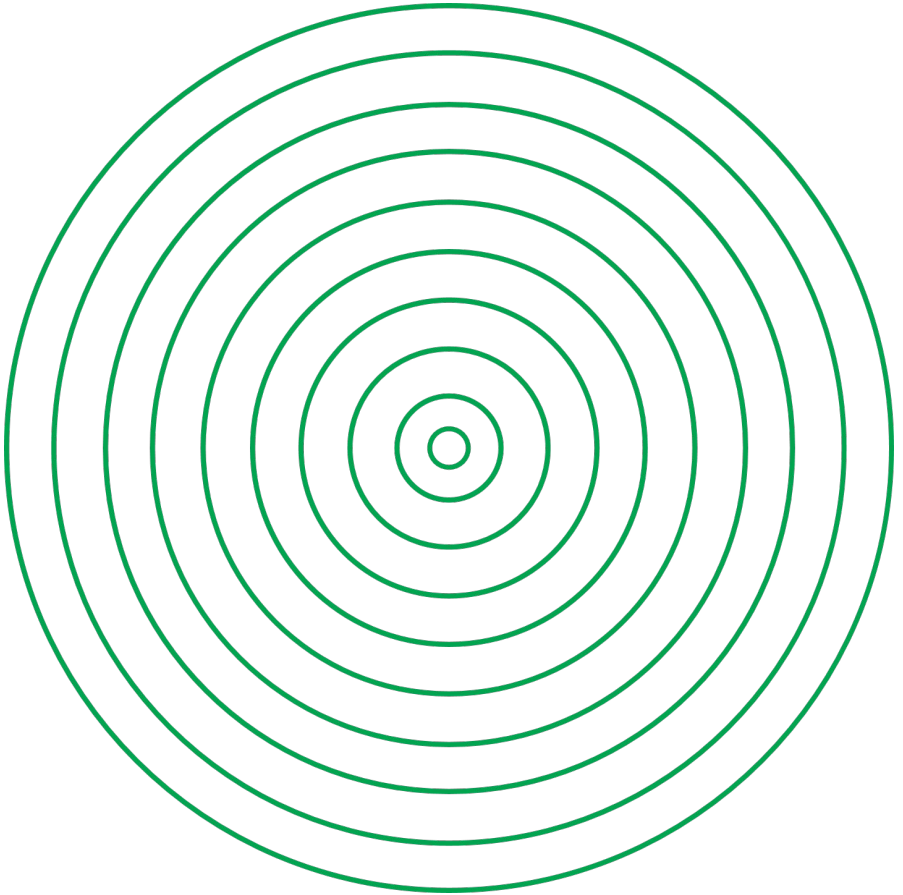


Data & Preprocessing

Topic Modeling

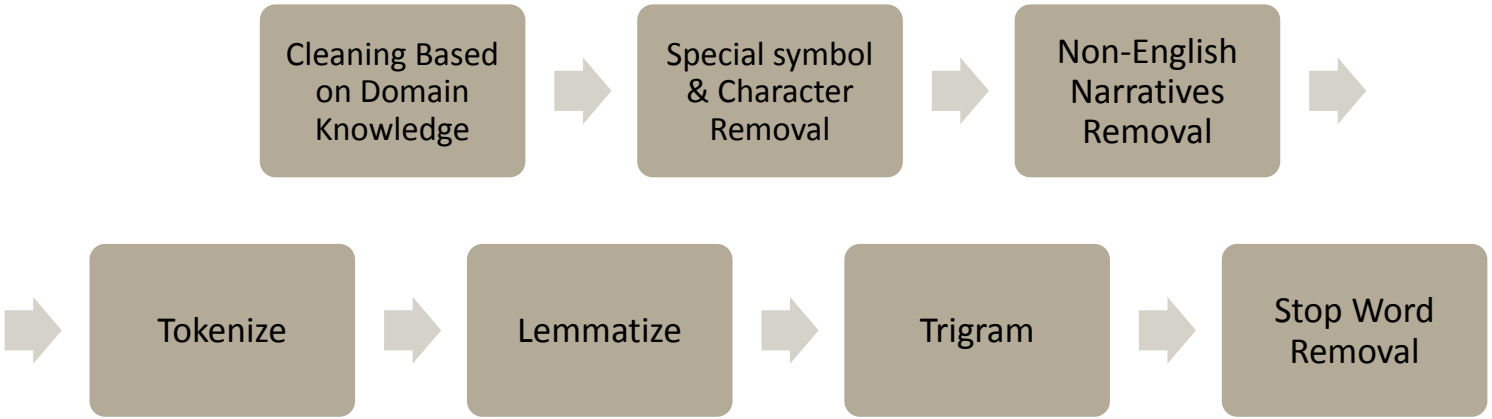
Topic Analysis

Comparison Analysis



Data and background

- **Data**: Consumer product incidents data from Clearing House and Health Canada
  - **187,166** cases and **6** common attributes (**1** narrative description).
  - Year from **2013** to **2019**
- **Field of Interest**: Consumer product safety issue in the English Speaking Region of North America
- **Data Preprocessing**:



Data Preprocessing

- **Cleaning based on domain knowledge**

  `***Form filled and submitted by <redacted>*** …`

- **Special symbol, character removal**

  `* , . © ? …`

- **Non-English records removal**

  `D but d'incendie au niveau d'un thermostat âlectrique dans une chambre coucher (maison unifamiliale)`

Data Preprocessing

- **Original Text**

```
CONSUMER HAS A SAFETY ISSUE WITH THREE PHOTOS OF A PRODUCT WHICH HAS LITERALLY
EXPLODED WITH THE COLD WEATHER. CONSUMER BELIEVES THIS PRODUCT POSES A HAZARD
WHEN EXPOSED TO FREEZING CONDITIONS.
```

- **Tokenize**

```
['consumer', 'has', 'safety', 'issue', 'with', 'three', 'photos', 'of',
'product', 'which', 'has', 'literally', 'exploded', 'with', 'the', 'cold',
'weather', 'consumer', 'believes', 'this', 'product', 'poses', 'hazard',
'when', 'exposed', 'to', 'freezing', 'conditions']
```

- **Lemmatize**

```
['consumer', 'safety', 'issue', 'photo', 'product', 'literally', 'explode',
'cold', 'weather', 'consumer', 'believe', 'product', 'pose', 'hazard',
'expose', 'freeze', 'condition']
```

Data Preprocessing

- **Trigram**

```
['consumer', 'safety', 'issue', 'photo', 'product', 'literally', 'explode',
'cold_weather', 'consumer', 'believe', 'product', 'pose', 'hazard', 'expose',
'freeze', 'condition']
```

- **Additional Stop words removal**

```
['safety', 'issue', 'photo', 'product', 'literally', 'explode',
'cold_weather', 'believe', 'product', 'pose', 'hazard', 'expose', 'freeze']
```

- **Original Text**

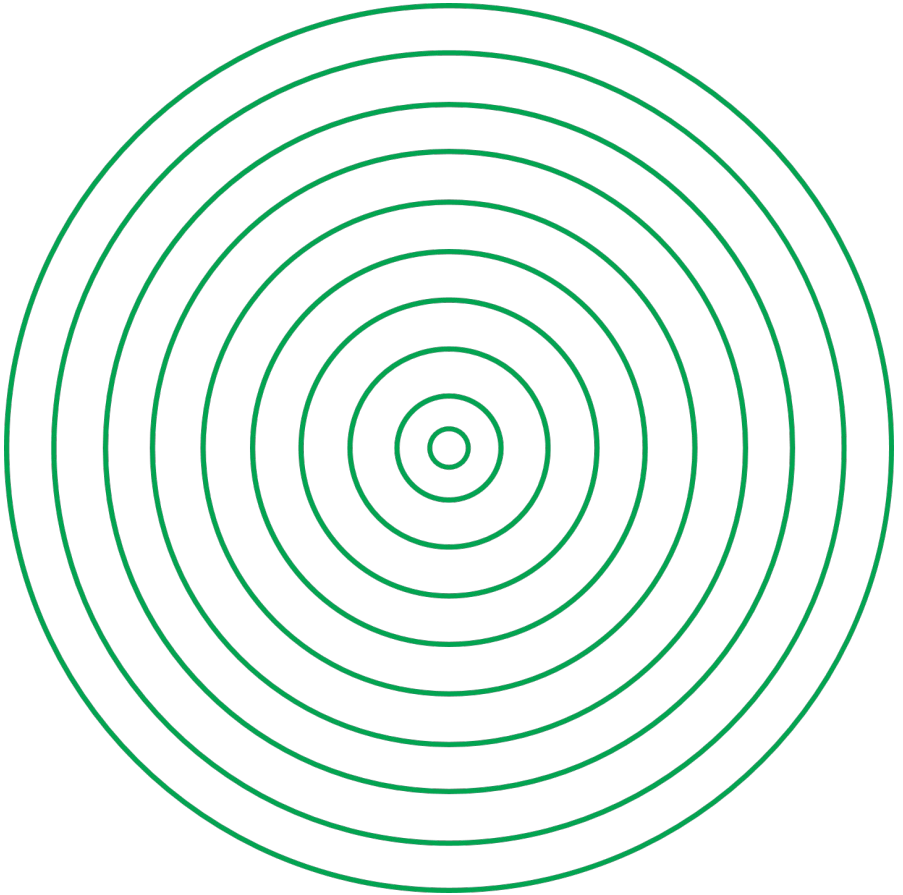
```
CONSUMER HAS A SAFETY ISSUE WITH THREE PHOTOS OF A PRODUCT WHICH HAS LITERALLY
EXPLODED WITH THE COLD WEATHER. CONSUMER BELIEVES THIS PRODUCT POSES A HAZARD
WHEN EXPOSED TO FREEZING CONDITIONS.
```

Data & Preprocessing

Topic Modeling

Topic Analysis

Comparison Analysis



Topic Modeling

- Various Machine Learning strategies for narrative analysis
  - Sentiment Analysis, Language Translation…
- Topic Modeling can enhance our understanding by revealing latent topics among narratives
  - Document
  - Word
  - Topic (assumption)

Sentence	Topic
“My dog’s name is Tony”	Dog
“I like the Beach”	Beach
“I like to walk my dog on the beach”	Dog & Beach

BERTopic Model

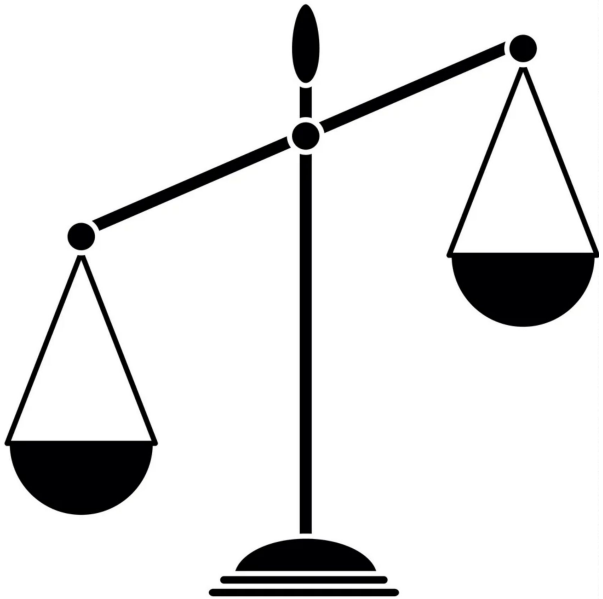
- Maarten Grootendorst, in 2020
- Components of BERTopic
  - **BERT**: word and document embedding
  - **UMAP**: dimension reduction
  - **HDBSCAN**: document clustering
  - **c-TF-IDF**: document representation
- Strengthens of BERTopic:
  - Context preserved
  - Hierarchical structure
  - Soft Clustering - Noises can be classified as outliers

Optimal Number of Topics

- The number of topics is a hyperparameter
- Metric: Normalized Pointwise Mutual Information (**NPMI**)
  - Imitate human judgment
  - Ranges from **-1** to **1**
- Searching Process
  - Initial Search
    - Iterating from **20** to **200** with steps of **10**
    - NPMI was calculated and averaged across **3** runs for each step
  - Precise Search
    - Shrink the scope into a range between **40** and **60** with step of **1**
    - NPMI was calculated and averaged across **3** runs for each step
- Optimal Result
  - Optimal number of topics: **42**
  - Optimal NPMI score **0.21**

Performance Comparison

- Candidates for performance comparison
  - Latent Dirichlet Allocation (**LDA**)
  - Biterm Topic model (**BTM**)
- Metric: Normalized Pointwise Mutual Information (**NPMI**)
- Optimal result for LDA
  - Optimal topic number: **37**
  - Optimal NPMI score: **0.129**
- Optimal result for BTM
  - Optimal topic number: **48**
  - Optimal NPMI score: **0.127**
- Recall optimal NPMI score for BERTopic is **0.21**

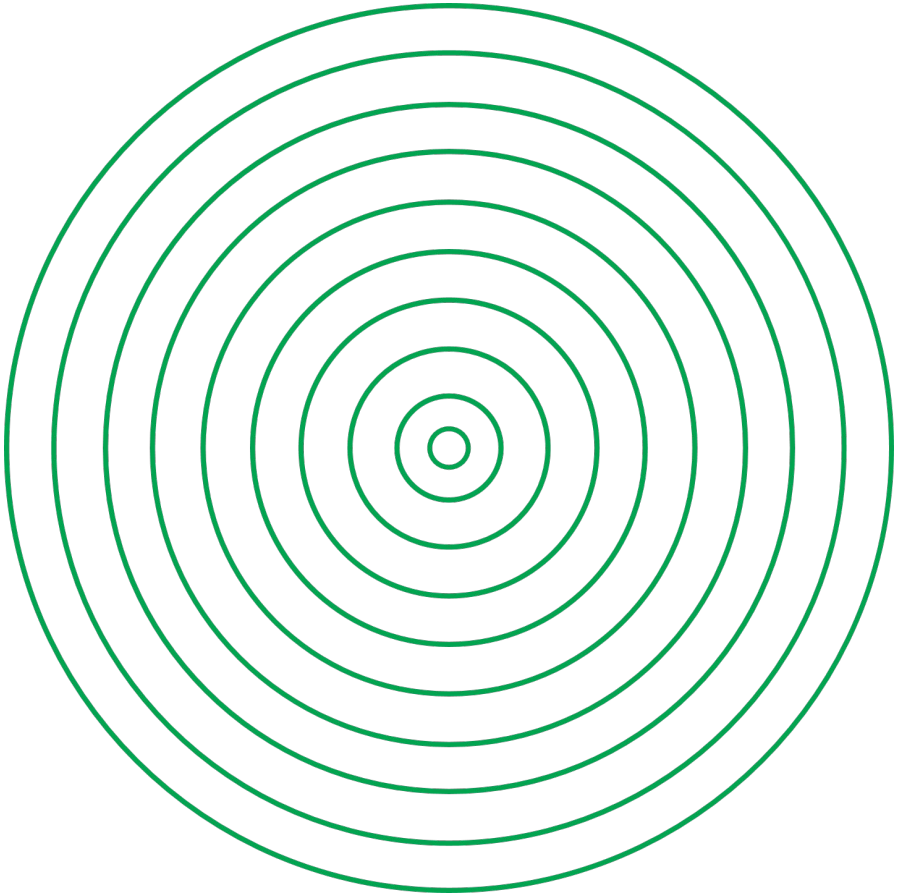


Data & Preprocessing

Topic Modeling

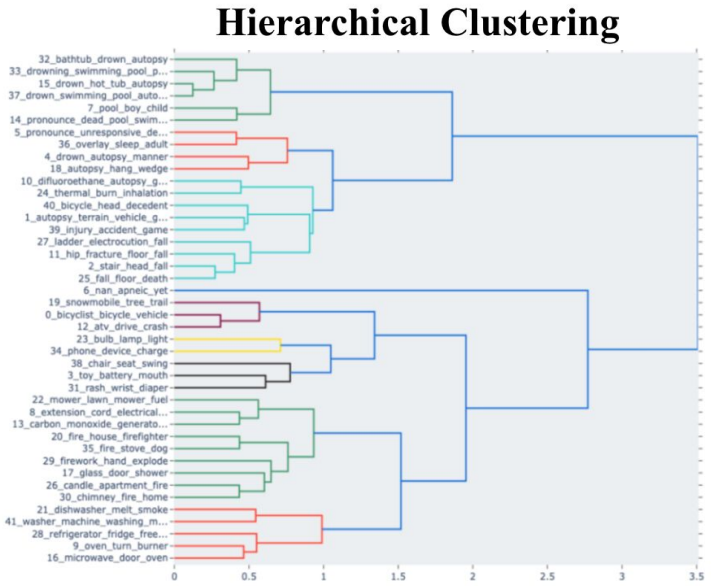
Topic Analysis

Comparison Analysis

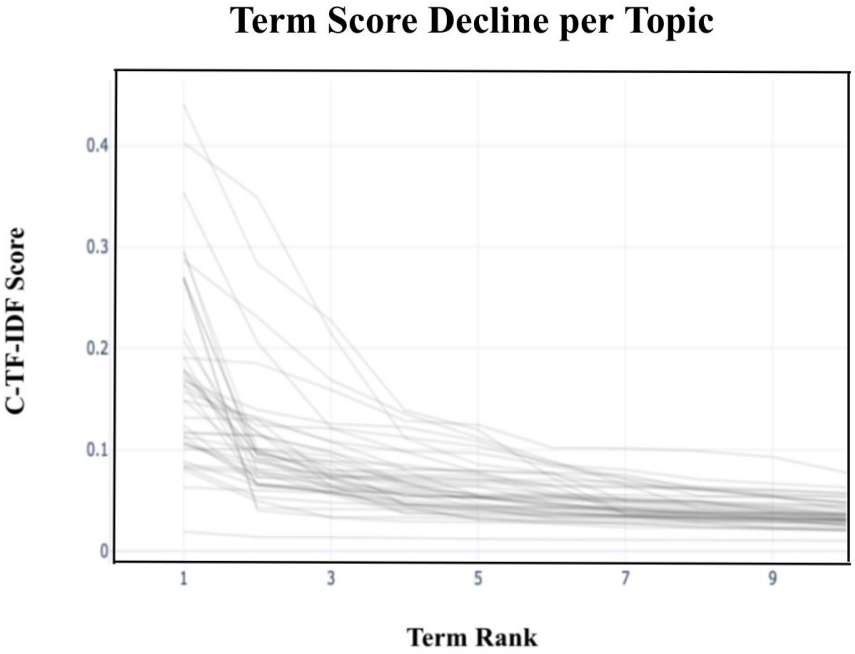


Topic Representation

- **42** topics with Hierarchical Structure

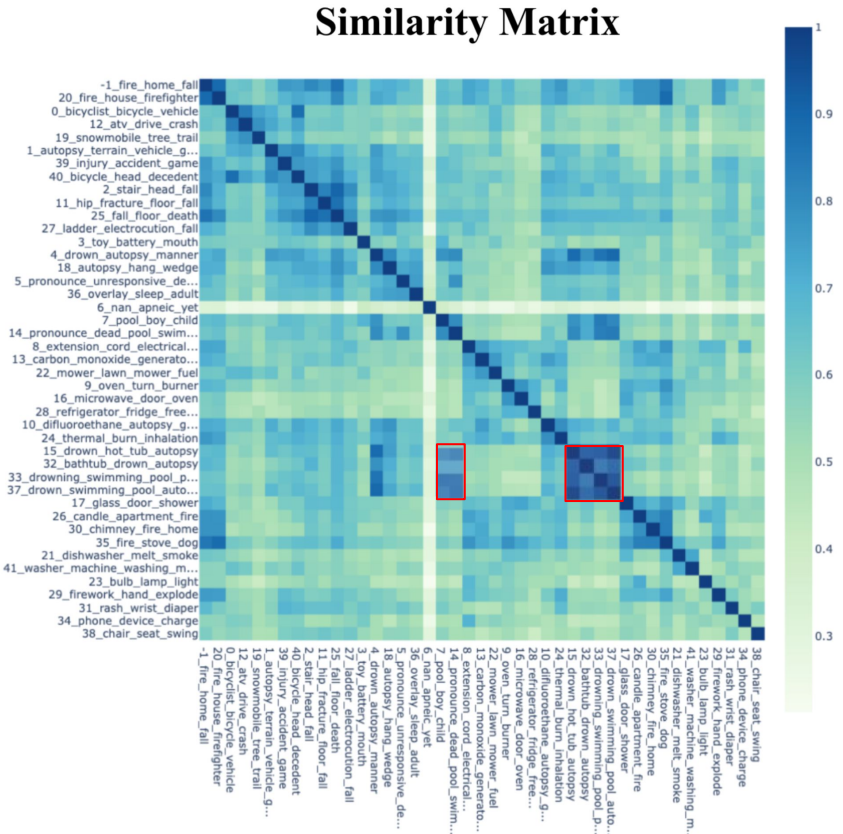


- Topic Representation
  - List of words ranked by their importance
  - **9** words can reliable represent each topic



Topic Similarity

- Some topics are similar to each other
  - A dark diagonal across the matrix
  - A vertical and horizontal white line across topic **6**
  - A slightly darker area around topics **15**, **32, 33**, and **37**



Topic Analysis (Drowning)

- Group topics based on their similarity
- Naming topics based on the **9** words representation list

Topic Representation

Topic_Number	Word_1	Word_2	Word_3	Word_4	Word_5	Word_6	Word_7	Word_8	Word_9
32	bathtub	drown	autopsy	seizure	drowning	decedent	tub	water	residence
33	drowning	swimming_pool	pool	drown	decedent	death	male	swim	face
15	drown	hot_tub	autopsy	pool	decedent	drowning	residential	submerge	home
37	drown	swimming_pool	autopsy	decedent	residence	hotel	drowning	pool	motel
7	pool	boy	child	drown	swimming_pool	hospital	pull	nearly	girl
14	pronounce_dead	pool	swimming_pool	pronounce	hospital	drown	backyard	drowning	unresponsive

Topic and Group Name

Main_Topic_Name	Sub_Topic_Name			Individual_Topic_Name	Topic_Number
Drowning	Location			Bath Tub Drowning	32
		Pool Drowning		Swimming Pool Drowning	33
			Home / Non-Home	Home Pool Drowning	15
				Non-Home Pool Drowning	37
	Other			Young Drowning	7
				Pronounce Dead Drowning	14

Topic Analysis (Unknown)

- Similarity does not guarantee existing of main topic
- Named all **42** topics based on the words list and similarity

Topic Representation

Topic_Number	Word_1	Word_2	Word_3	Word_4	Word_5	Word_6	Word_7	Word_8	Word_9
38	chair	seat	swing	sit	break	leg	stool	back	base
3	toy	battery	mouth	pacifier	child	baby	choke	piece	son
31	rash	wrist	diaper	fitbit	wear	skin	blister	watch	band

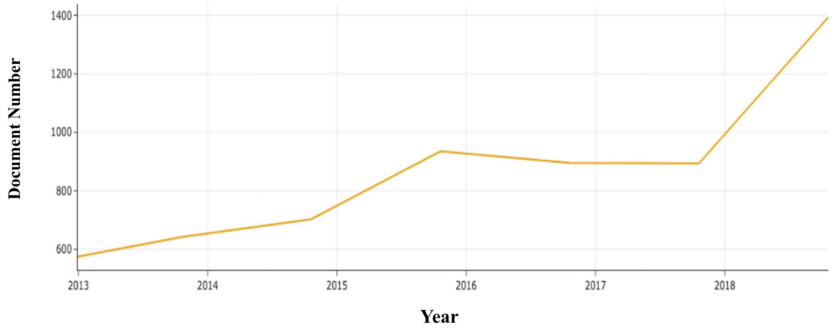
Topic and Group Name

Main_Topic_Name	Sub_Topic_Name	Individual_Topic_Name	Topic_Number
		Baby Chair	38
	Hazard	Small Part Hazard	3
		Skin Rash	31

Dynamic BERTopic

- Analyze topics from a dynamic perspective:
  - Quantitatively: popularity of a topic
  - Qualitatively: Meaning change of a topic
- Topic Mobility
  - Become popular over the year
  - Scooter appeared between **2014** to **2017**
- Topic Skin Rash
  - Fitbit and watch suddenly appeared since 2014

Quantitative Evolution for Topic Mobility



Quanlitative Evolution for Topic Skin Rash

Topic 31 *	Year	Top Five Words
Skin Rash	2013	diaper, rash, pamper, blister, wear
	2014	rash, wrist, fitbit, wear, fitbit_force
	2015	rash, wrist, wear, fitbit, diaper
	2016	diaper, rash, wrist, fitbit, wear
	2017	rash, diaper, sunscreen, wrist, skin
	2018	rash, wrist, fitbit, diaper, watch
	2019	watch, wrist, rash, wear, fitbit

Quanlitative Evolution for Topic Mobility

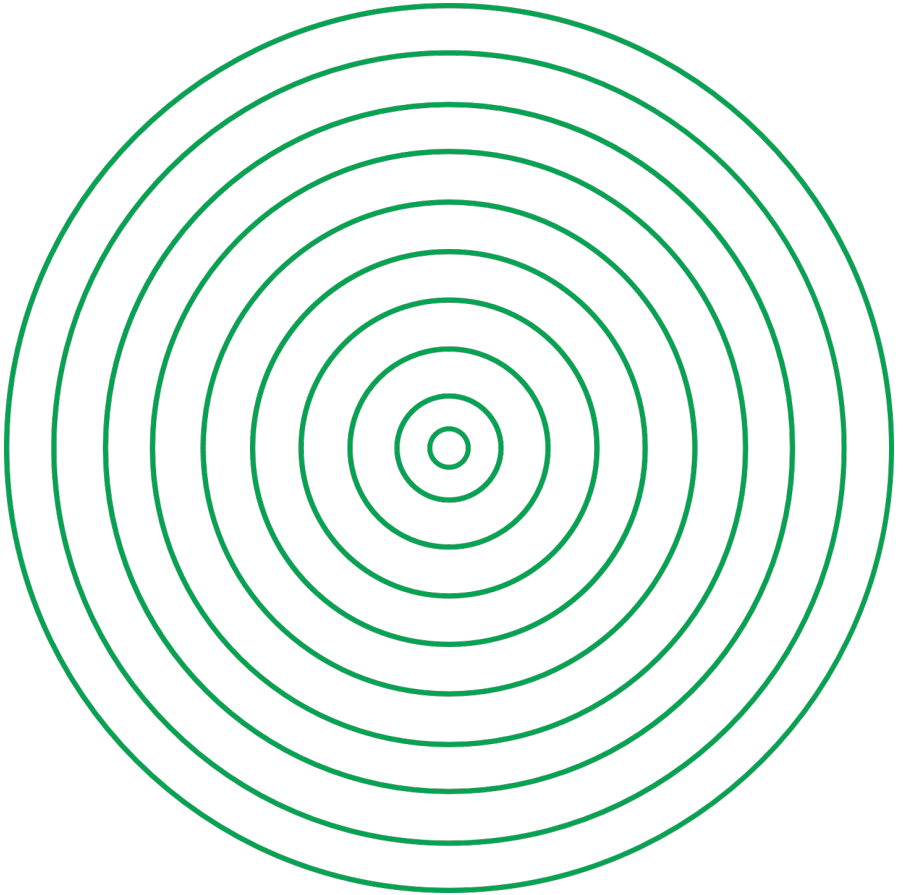
Topic 0 *	Year	Top Five Words
Mobility	2013	bicycle, bicyclist, ride, car, strike
	2014	bicycle, bicyclist, car, ride, scooter
	2015	bicyclist, bicycle, car, scooter, ride
	2016	bicyclist, bicycle, scooter, car, vehicle
	2017	bicyclist, bicycle, scooter, car, vehicle
	2018	bicyclist, bicycle, vehicle, car, ride
	2019	bicyclist, bicycle, vehicle, ride, strike

Data & Preprocessing

Topic Modeling

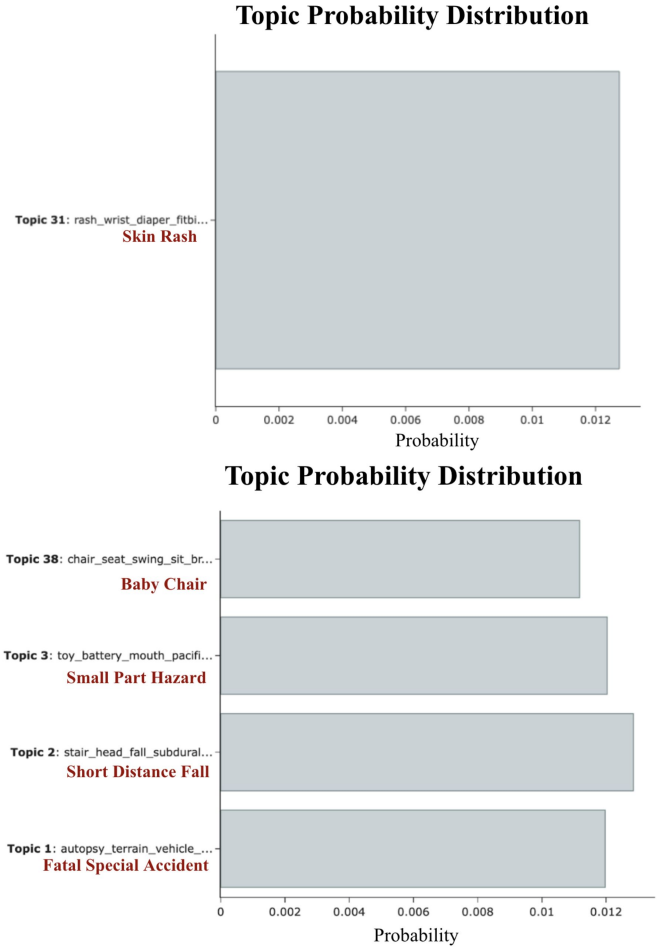
Topic Analysis

Comparison Analysis



Topic Prediction

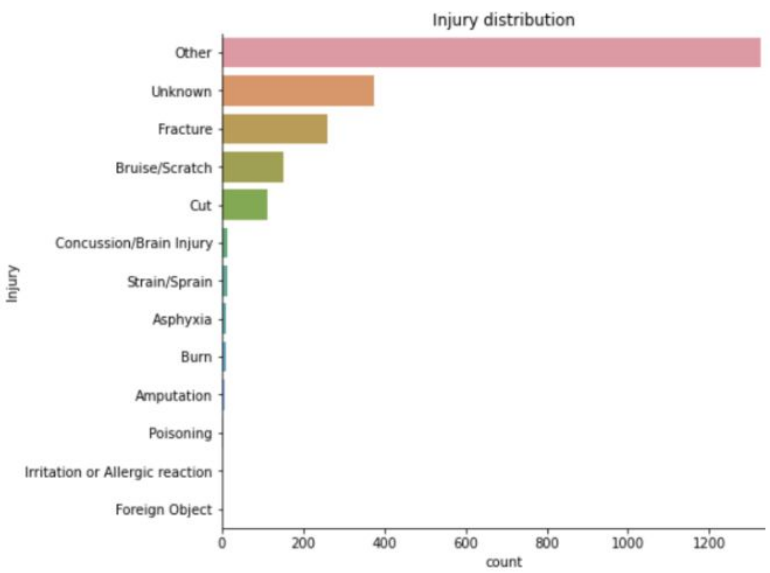
- Topic model summarize a narrative as probability distribution:
  - **[Simple Senorio]** On vacation I use the banana boat product sunscreen and 4 days later he broke out on rash that turned into blister like.
  - **[Complex Senorio]** We (my husband and I) had our baby strapped into the "Feeding Chair" in the kitchen with us. It was not attached to a kitchen chair at the time as we did not have one. I picked him up in the chair to move him to the living room. The back of the chair became detached, it turned upside down and I wasn't able to hold on (as the back came loose) and the chair and baby fell to the floor. It was terrible.
- Only captured primary topic meaning in complex senorio (caution)

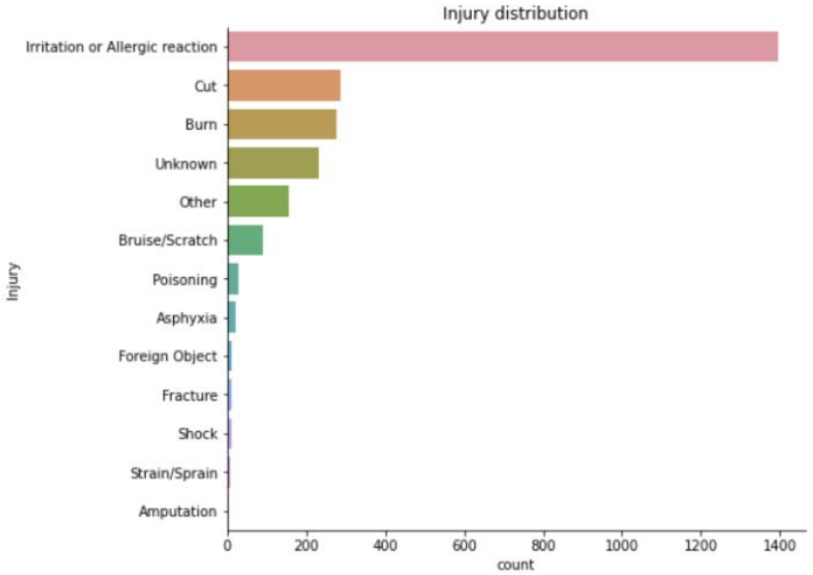


Comparison Analysis (Injury)

- Injury type align with the expectation of topics

Injury in Topic Mobility (Left) and Skin Rash (Right)

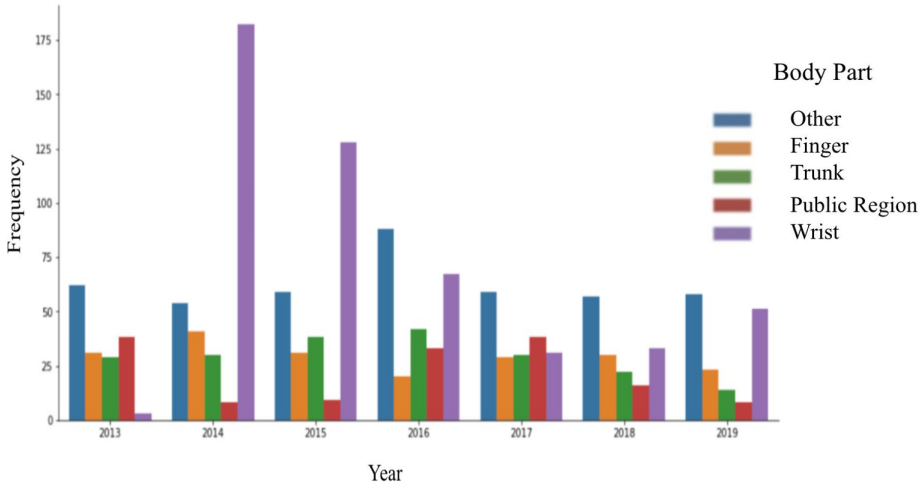




Comparison Analysis

- Wrist suddenly Increases since 2014
- Matches the Dynamic BERTopic Analysis

Top 5 Body Parts between 2013 and 2019 in topic Skin Rash



Quanlitative Evolution for Topic Skin Rash

Topic 31 *	Year	Top Five Words
Skin Rash	2013	diaper, rash, pamper, blister, wear
	2014	rash, wrist, fitbit, wear, fitbit_force
	2015	rash, wrist, wear, fitbit, diaper
	2016	diaper, rash, wrist, fitbit, wear
	2017	rash, diaper, sunscreen, wrist, skin
	2018	rash, wrist, fitbit, diaper, watch
	2019	watch, wrist, rash, wear, fitbit

Conclusion

- Topic Modeling: A strategy to extract latent meaning (topic) underneath narratives.
- BERTopic
  - General Perspective
  - Dynamic Perspective
- Insights from comparison analysis
  - Structured data are not available
  - Structured data are available

Thank you.