BERTopic Modeling on Consumer Product Incident Narratives

UL Standards and Engagement

Data Science Team

Hanjun Wei

Introduction

Throughout history, narratives have been crucial in preserving and disseminating knowledge. In different eras, narratives have served diverse purposes, safeguarding religious foundations, fostering humanistic thoughts, and documenting societal progress. From the ancient Dead Sea Scrolls protecting the Hebrew world's religious legacy to the Renaissance literary narratives like Hamlet and Othello enlightening the Western world's humanistic ideals, narratives have been instrumental in shaping our understanding of the world.

The preference for narratively recording information can be attributed to the flexible structure that allows individuals to express thoughts from various perspectives while preserving ideas and memories to a significant extent. However, this flexibility has gradually transformed into a drawback in the face of today's information explosion. The abundance of narratives makes it challenging to comprehend them comprehensively within limited timeframes due to their inherent ambiguity. As a result, structured data in the form of tables have gained popularity as they provide simpler and more precise information. Yet, despite this shift, a vast majority of the data available to us today still exist in narrative form.

This prevalence of narratives can be attributed to two primary reasons. Firstly, narratives are more cost-effective compared to structured data because they align with the natural language intuition of humans. Anyone proficient in the language can produce or collect narrative data, making it accessible for general data collectors and organizations. Conversely, structured data require specific design and planning, rendering them more intricate and expensive to collect. Secondly, narratives excel at preserving contextual information. While structured data necessitates the pre-definition of attributes to be captured, narratives can encompass more aspects and nuances in natural language. Particularly in scenarios requiring reporting-type information where each individual report holds unique characteristics, designing a data structure capable of capturing all these intricacies becomes challenging. Thus, narratives remain a more feasible option.

Fortunately, the advancement of Data Science has led to the development of various Natural Language Processing (NLP) strategies for working with narrative data. In this article, we will introduce a prominent NLP strategy called Bertopic Modeling and employ this methodology to analyze consumer product incident reports. By comparing the findings with corresponding structured data, we will evaluate the effectiveness and applicability of this approach.

Dataset

The target field of our study focuses on consumer product safety issues in the English-speaking region of North America. To gather data for our analysis, we utilized information obtained from the U.S. Consumer Product Safety Commission (CPSC) and the Canadian Ministry of Health (Health Canada). The CPSC National Injury Information Clearinghouse (Clearinghouse) and Health Canada's consolidated Consumer Product Incidents System served as the primary sources for data collection. The dataset acquired from these sources encompassed incident narratives as well as pre-coded data.

The consolidated database included pre-coded metadata that provided information on the product types involved, severity of injuries, and the origin of each report. To gain a comprehensive understanding of the

targeting issues within these regions, we extracted and combined all incidents recorded between 2013 and 2019 from both sources. Among the various variables available, our primary interest lies in identifying previously undiscovered patterns within the unstructured incident descriptions.

These incident descriptions serve as excellent examples of reporting-type information that cannot be effectively preserved in a structured format. One advantage of this dataset is that it contains both narratives and structured data, allowing for meaningful comparisons between the two.

Topic Modeling

Data scientists have devised effective strategies, such as sentiment analysis, to extract hidden value from narratives. Among these approaches, Topic Modeling is particularly well-suited for our research objective of comprehending the narrative dataset holistically. By utilizing Topic Modeling, we can uncover latent topics within the narratives, providing valuable insights into their content.

For instance, consider the following example:

1. "My cat's name is Bagle."
2. "I like sitting on my sofa."
3. "I enjoy sitting on my sofa and watching my cat."

In this example, each sentence represents a "document," and each term within the sentences acts as a "word." Humans can easily deduce that the first sentence pertains to the topic of cats, the second to the topic of sofas, and the third to both cats and sofas. Similarly, an effective topic model should detect and represent these topics by generating a list of keywords ranked by their importance.

While numerous well-developed topic modeling algorithms exist in data science, our project found that Bidirectional Encoder Representations from Transformers Topic Modeling (BERTopic) outperformed traditional models like Latent Dirichlet Allocation (LDA) and exhibited advanced properties. In the following sections, we will delve into our utilization of BERTopic and discuss its impact on enhancing our understanding of incident descriptions. However, before proceeding with model construction, ensuring the quality of our input data is crucial.

Pre-process

Since our data were collected from two different sources, we initially combined them and selected the five variables that were common to both. Among these variables, the narrative incident description was the only data used for NLP topic modeling, while the remaining attributes were utilized for comparison analysis. To ensure the quality of our model, we conducted comprehensive data preprocessing on the incident descriptions.

Firstly, we cleaned the incident descriptions based on domain knowledge of consumer product incidents and reporting systems. Upon reviewing the dataset, we observed that many meaningless phrases shared

similar formats, which we then removed based on those formats. For example, phrases like "*Form filled and submitted by <redacted>*" were eliminated.

Secondly, we performed text cleaning by removing special symbols, characters, and unnecessary spaces, as they hold no relevance within the scope of our study. Symbols such as "*", ",", ".", "©," and "?" were eliminated.

Considering that Canada has two official languages, text extracts from Health Canada could be recorded in either English or French. To maintain consistency, we filtered out all observations that were recorded in French.

Next, we tokenized the string text narratives into a list of words and lemmatized the narratives. Lemmatization involved converting words back to their common root form or lemma. Additionally, we used Bigram and Trigram techniques to preserve meaningful phrase combinations. This approach reduced the possibility of the algorithm treating different lemmas with the same meaning as distinct objects and improved data quality by avoiding words that only carried partial meaning.

Lastly, we removed stop words, such as "I," "a," or "the," which do not carry important meaning within the scope of this investigation. This step accelerated the preprocessing process and streamlined subsequent analysis.

Here is an example illustrating the comparison between the original text data and the cleaned text data:

Original Text Data: "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."

Cleaned Text Data: ['photo', 'literally', 'explode', 'cold_weather', 'pose', 'expose', 'freeze', 'condition']

After implementing these procedures, we obtained a set of cleaned text data ready for further use. In the subsequent sections, we will introduce BERTopic and provide insights into its construction and utilization.

BERTopic Model

BERTopic is a topic modeling method that leverages Bidirectional Encoder Representations from Transformers (BERT) to map words into numeric vectors, enabling computer processing of textual data through word and document embeddings. Additionally, it utilizes Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) to reduce the dimensionality of the dataset, ensuring the quality of subsequent clustering. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is employed for grouping documents that share similar meanings, effectively identifying topics through document clustering. Furthermore, BERTopic employs class-based term frequency-inverse document frequency (c-TF-IDF) to represent each topic, ranking words by their importance.

Unlike traditional topic modeling techniques such as Latent Dirichlet Allocation (LDA), which treat documents as unordered bags of words, the document embeddings in BERTopic preserve semantic relationships by generating contextual representations. While other embedding methods like Doc2Vec exist, they often adopt a centroid-based clustering and topic interpretation perspective that may not hold true in real-world scenarios. Centroid-based methods assume clusters to be Gaussian spheres, whereas HDBSCAN, a density-based clustering approach, avoids such parametric assumptions by relying solely on data density. This flexibility allows clusters to exhibit various shapes, and the c-TF-IDF technique assists in identifying the most relevant words associated with a given topic or cluster.

Additionally, HDBSCAN provides a hierarchical view of the clusters, while the soft-clustering approach allows for the classification of outliers as noise. These properties contribute to a more precise understanding of latent topics and their relationships.

The optimal number of Topics

In our project, we employed the Normalized Pointwise Mutual Information (NPMI) coherence score, the same metric used by the creator of BERTopic, to assess the model's performance. This score has demonstrated a satisfactory level of performance in approximating human judgment. It ranges from -1 to 1, with 1 indicating perfect coherence.

To determine the optimal number of topics, we initially evaluated the NPMI coherence score by iterating through all possible topic numbers from 20 to 200 in increments of 10. The NPMI score was calculated at each step, and the results were averaged across three runs for each step. Based on this analysis, we narrowed down the range of optimal topic numbers to between 40 and 60.

Next, we conducted a more precise search within the range of 40 to 60 topics, using increments of 1. At each step, we calculated the NPMI score and averaged the results across three runs. The analysis revealed that the optimal number of topics was determined to be 42, with the highest NPMI coherence score recorded at 0.21.

Performance Comparison

To evaluate the performance of BERTopic, we conducted a comparison with other popular topic models in the field, using a similar methodology. The models chosen for comparison were Latent Dirichlet Allocation (LDA), a well-established topic modeling technique, and the Biterm Topic Model (BTM), specifically designed for short text topic modeling.

We performed a search for the optimal number of topics ranging from 20 to 80, with a step size of 1. For both LDA and BTM, we calculated the NPMI coherence score at each step and averaged the results across three runs. Our analysis revealed that the optimal topic number for LDA was 37, with an NPMI coherence score of 0.129. On the other hand, for BTM, the optimal topic number was 48, with an NPMI coherence score of 0.127.

Comparing the optimal results across the three strategies, BERTopic demonstrated superior performance, as it achieved a higher NPMI coherence score of 0.21. Based on this comparison, we have selected BERTopic as our preferred topic modeling method for generating and analyzing the latent topic list.

Topic Analysis

This graph illustrates the hierarchical relationships among the 42 topics generated by the BERTopic model. Each topic is represented by the three most important words calculated using c-TF-IDF, and users have the flexibility to select the desired number of words for topic representation.

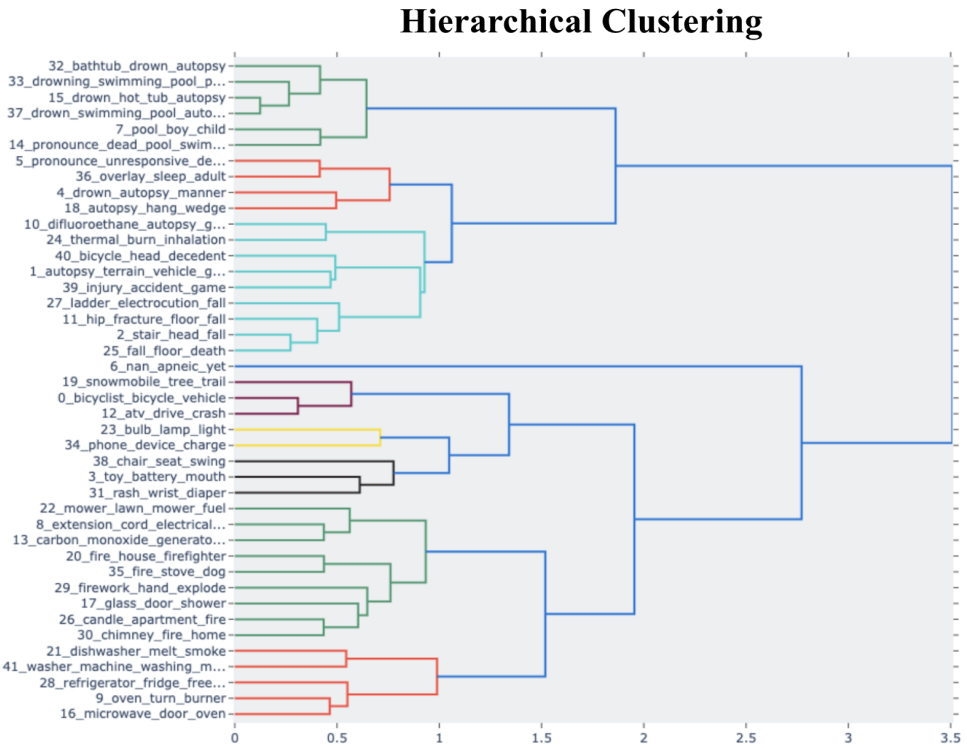

Figure 1. Illustrating the hierarchical relationship among topics. Topics with same color branches are belong to the same main topic.

To determine the optimal number of words for meaningful topic representation, we analyzed the marginal value contributed by an additional word, as shown in Figure 2.

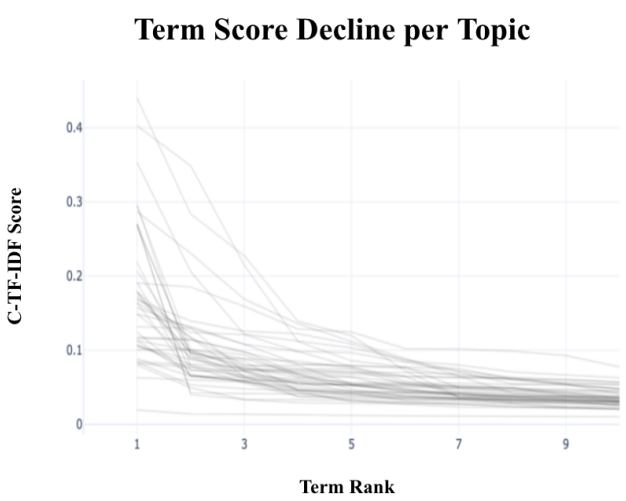

Figure 2. Illustrating the marginal value brings by the additional word for topic representation.

The plot consists of a total of 42 lines, each representing the marginal importance (c-TF-IDF score) of an additional word in topic representation. The negative correlation between the c-TF-IDF score and the Term Rank indicates that the value added by each additional word decreases. Notably, the c-TF-IDF score for the 9th word in each topic representation is consistently lower than 0.1. Hence, we confidently conclude that nine terms are sufficient to explain every topic adequately.

Observing the hierarchical clustering plot (Figure 3), we can identify branches connecting topics with similar meanings, indicated by shared colors. This similarity is further depicted in the Similarity Matrix, where the shade reflects the level of similarity between two topics. Darker shades represent higher similarity. The diagonal line in dark blue across the matrix corresponds to each topic's similarity to itself.

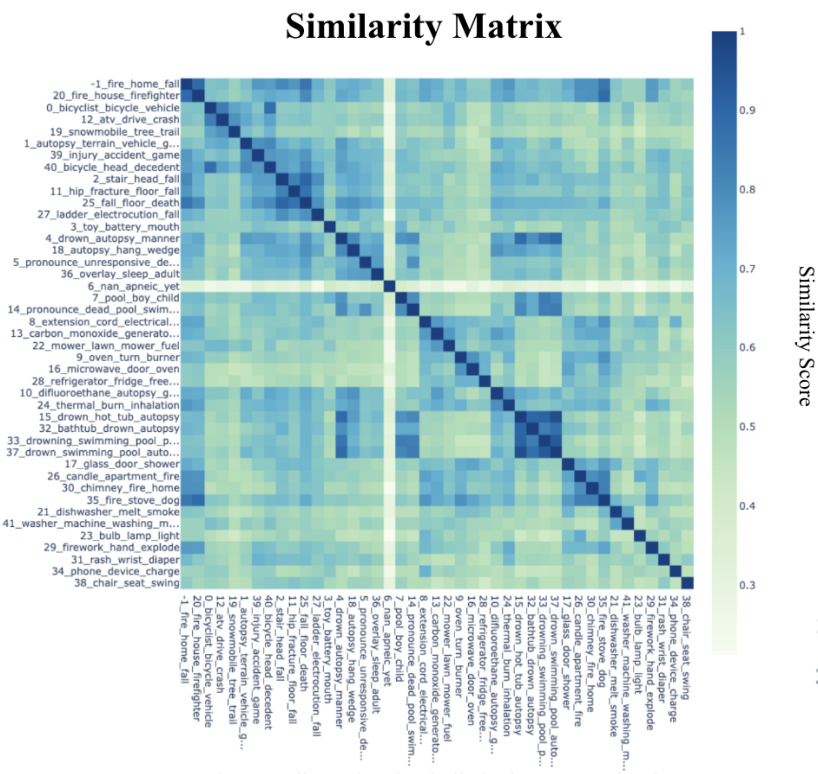

Figure 3. Illustrating the similarity between each pair of topics. Darker shade indicating high similarity and lighter shade indicates low similarity.

Additionally, a slightly darker area is visible around topics 15, 32, 33, and 37 in the bottom right of the matrix, suggesting their similarity. Specifically, there is an even darker square between topics 33 and 37. These minor similarities align with the hierarchical topic graph in Figure 4, as more similar topics are grouped together. Such hierarchical relationships contribute to BERTopic's determination of topic relationships.

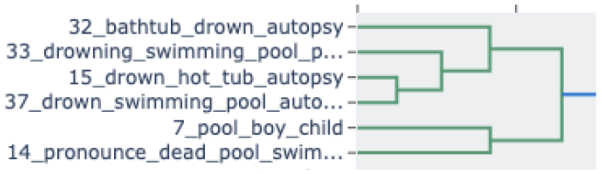

Figure 4. An example of a main topic "drown" with its corresponding sub-topics

The similarity matrix also reveals a vertical and horizontal white line across topic 6, indicating that all other topics significantly differ from topic 6. Further analysis revealed that this grouping occurred because BERTopic had categorized all stop words, which were not removed during preprocessing, into this specific topic.

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

Figure 5. The top 9 representation words for each individual topics Related to the main topic "Drowning".

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

Figure 6. The individual topic names, sub topic name, and main topic name we assigned based on the top 9 representation words and intertopic similarity.

We subsequently named and distinguished these topics and their hierarchical groups based on the top 9 words and their relationships. For instance, topics 32, 33, 15, 37, 7, and 14 were categorized together. Figure 5 provides an example of how BERTopic represents topics, displaying a list of keywords ordered by their importance. To better understand these topics, we manually assigned names based on the keyword list, their order, and their differentiation from similar topics.

Upon examining the keywords in Figure 5, we observe that all of them pertain to drowning. This aligns with their grouping due to high similarity in Figure 4. However, each topic still possesses unique characteristics enabling their distinction. For instance, in topic 7, the second representation word "boy" and the third representation word "child" indicate a drowning topic associated with young individuals, leading us to name it "Young Drowning." Similarly, topic 32, with its first representation word "bathtub" and seventh representation word "tub," suggests a drowning topic related to bathrooms, thus named "Bath Drowning." Topic 33, represented by the second word "swimming_pool," indicates a drowning topic connected to swimming pools, earning the name "Swimming Pool Drowning."

Topics 15 and 37 are also related to pool drowning. However, the ninth representation word "home" in topic 15 and the sixth and ninth representation words "hotel" and "motel" in topic 37 differentiate them. Topic 15 is associated with drowning incidents occurring at home, while topic 37 involves drowning incidents in hotels or motels. Hence, we named them "Home Pool Drowning" and "Non-Home Pool Drowning," respectively.

Lastly, topic 14 lacks unique keywords for differentiation. However, the order of the keywords indicates that "pronounced_death" is the most important feature, leading us to name it "Death Drowning."

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	baby	back	base
3	toy	battery	mouth	pacifier	child	baby	choke	piece	son
31	rash	wrist	diaper	fitbit	wear	skin	blister	watch	band

Figure 7. The top 9 representation words for each individual topics Related to the “unname” main topic.

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

Figure 8. The individual topic names, sub topic name we assigned based on the top 9 representation words and intertopic similarity.

After naming individual topics, we determined the names for the main topic and sub-topic groups indicated by the branches in the hierarchical clustering plot. However, it is important to note that similarity among topics does not guarantee their categorization under a single main topic. For instance, in Figure 7, topics 3, topic 28, and 31 are somewhat related to babies, where topics such as baby chair are a subset of "Chair and Seat" (topic 38), babies' choking hazards are a subset of "Small Part Hazard" (topic 3), and diaper-related skin rashes are a subset of "Skin Rash" (topic 31). However, this does not imply that all three topics belong to the overarching topic of "baby." Thus, not every group of similar topics will

have a common main topic, and the hierarchical structure serves as a guide to understand topic relationships rather than a rigid blueprint for topic naming.

Following a thorough analysis, we assigned names to all 42 topics based on their keyword representations and the order of the keywords.

Dynamic BERTopic

BERTopic offers the advantage of analyzing topics from a dynamic perspective. To achieve this, we first fit the entire corpus to BERTopic, as done previously. We then create a local representation of each topic by multiplying the term frequency of documents in a specific year with the global IDF values, assuming that the yearly topic representation is independent of the global topic representation. This approach allows us to gain a quantitative and qualitative understanding of how each topic evolves over the years.

From a quantitative standpoint, consider a topic related to structural fires. If we observe a gradual increase in the number of documents associated with this topic, it may indicate that the problem has become more severe and requires heightened attention. On the other hand, qualitative evolution can be exemplified by a topic such as "vehicles." In the year 2000, the top three keywords might have been "car," "drive," and "Toyota." However, in 2020, the top three keywords could have shifted to "EV," "self-drive," and "Tesla." Studying the keyword representation allows us to comprehend how the meaning of a topic has changed over time.

During our dynamic analysis, we discovered that only three topics exhibited an overall increasing trend among the 42 topics: "Drowning Asphyxia," "Mobility," and "Stop Words." Additionally, most topics displayed consistency in keyword representation over the years, suggesting their meaning remains stable. However, certain topics indicated the potential for meaning change. To illustrate this, we will focus on two topics as examples of topic evolution.

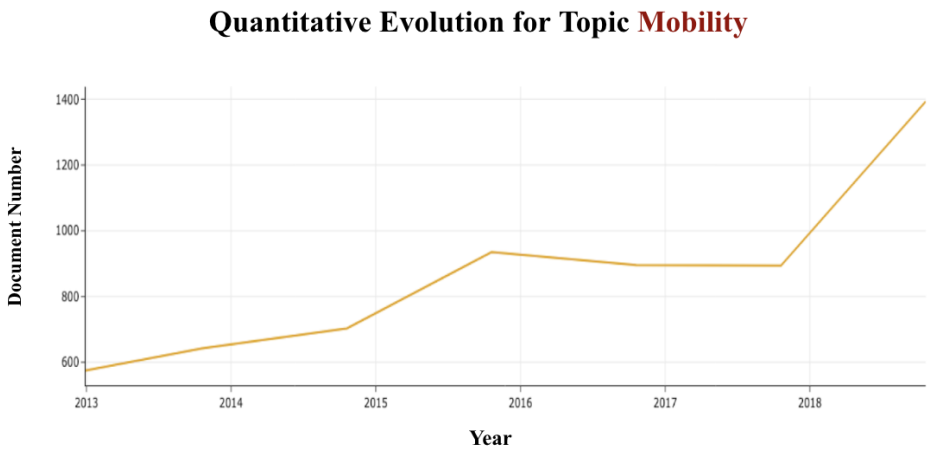

Figure 9. The change of document number related to topic Mobility between 2013 and 2019.

The first example pertains to topic 0, which we named "Mobility" based on its global keyword representations. We observed an overall increasing trend between 2013 and 2019 (Figure 9), indicating a

rise in incidents related to mobility. Further examination of the yearly topic keyword representations revealed that the keyword "scooter" began appearing among the top five keywords from 2014 to 2017. This observation raises intriguing questions, such as why scooters experienced a surge in popularity during that period. Further research unveils that Scoot Networks, the first American scooter share system app, was launched in 2012. Does the increased importance of scooters in Topic 0 reflect the growth of the sharing e-scooter industry?

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

Figure 10. illustrating the evolution of topic meaning for "Mobility" over years. Here, you can see scooter become more important since 2014 and suddenly disappear after 2017

As another example, we noticed that "Fitbit" and "watch" suddenly became crucial keywords in topic 31 (Skin Rash) in 2014. Fitbit is a well-known wearable device brand in the US, and it is worth noting that the first generation of the Apple Watch was launched in 2015. The sudden rise in smartwatches' popularity during those years could have led to increased demand and supply, potentially resulting in poor quality control. This example highlights how the meaning of a topic evolves in line with technological advancements.

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

Figure 11. illustrating the evolution of topic meaning for "Skin Rash" over years. Here, you can see fitbit and watch become more important since 2014

These examples demonstrate the dynamic perspective offered by BERTopic and how it enhances our understanding of topics and narratives by providing a different analytical angle.

Topic Prediction

Up until this point, we have provided an overview of Topic Modeling, discussed the reasons for using BERTopic, and explained how we constructed both a BERTopic and dynamic topic model. In the next step, we will utilize our model to predict the primary topic for each narrative.

The process involves the topic model predicting a topic probability distribution for every narrative. For instance, consider the following example:

- *"On vacation, I used the banana boat product sunscreen, and four days later, I broke out in a rash that turned into blisters."*

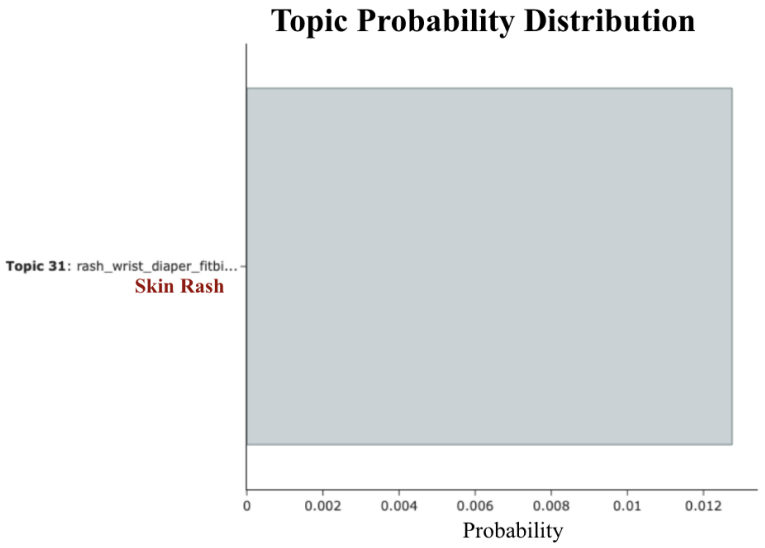

Figure 12. This probability distribution of this text belong to each topic. For this narrative, only Topic 31(Skin Rash) has a non-zero probability.

From a human perspective, we can summarize this narrative as "skin rash caused by sunscreen." However, for BERTopic, the summary is in the form of a topic probability distribution.

In this specific narrative, topic 31 (skin rash) is the only topic with a non-zero probability. As a result, the topic of skin rash will be assigned as the primary topic, indicating that the narrative predominantly relates to a sunscreen rash.

However, not all narratives have such straightforward topic predictions. Longer narratives tend to have more complex topic probability distributions. For example:

- *"We (my husband and I) had our baby strapped into the 'First Years Mi Swivel 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."*

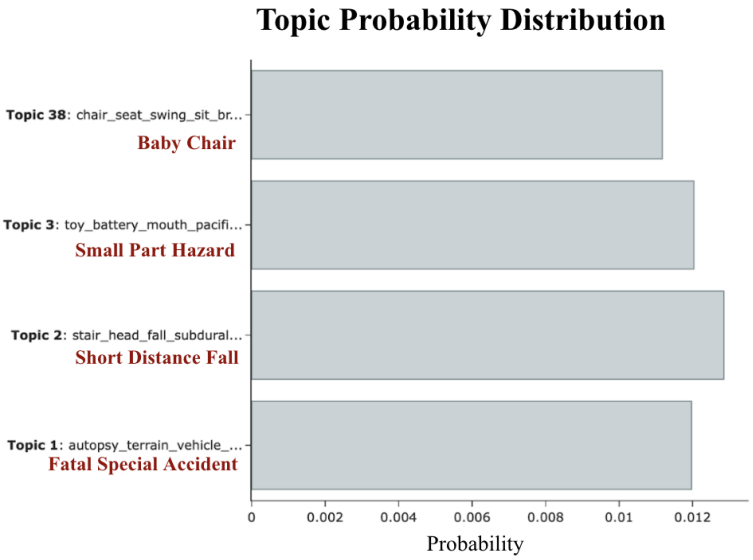

Figure 13. This probability distribution of this text belong to each topic. For this narrative, there are 4 topics with non-zero probability.

By reading this narrative, we can understand that it involves a baby falling from a chair. The topic probability distribution reveals that the topic model suggests multiple topics are involved in this narrative.

Among these topics, topic 2 (Fall) and topic 38 (Baby Chair) align with human understanding. However, topic 1 (Fatal Special Vehicle Accident) and topic 3 (Small Part Hazard) seem unrelated. The reason our model associates this narrative with the topics "Fatal Special Vehicle Accident" and "Small Part Hazard" may be due to common words such as "strapped," "detached," and "terrible" shared by both the "Baby Chair" and "Fatal Special Vehicle Accident" topics, as well as the word "baby" shared by both the "Baby Chair" and "Small Part Hazard" topics.

To maintain a conservative approach, we assign the topic with the highest probability as the primary topic for a narrative. However, it is essential to recognize that this approach may not fully capture the complete meaning of a given narrative. In the example above, we assign the topic "Fall" as the primary topic while leaving the topic "Baby Chair" unrecorded. With this consideration in mind, we proceed to predict and assign primary topic labels for the remaining narratives.

Comparison Analysis

Lastly, we will evaluate the quality of our topic modeling analysis result by comparing the structured attributes with the primary topics we assigned to each narrative.

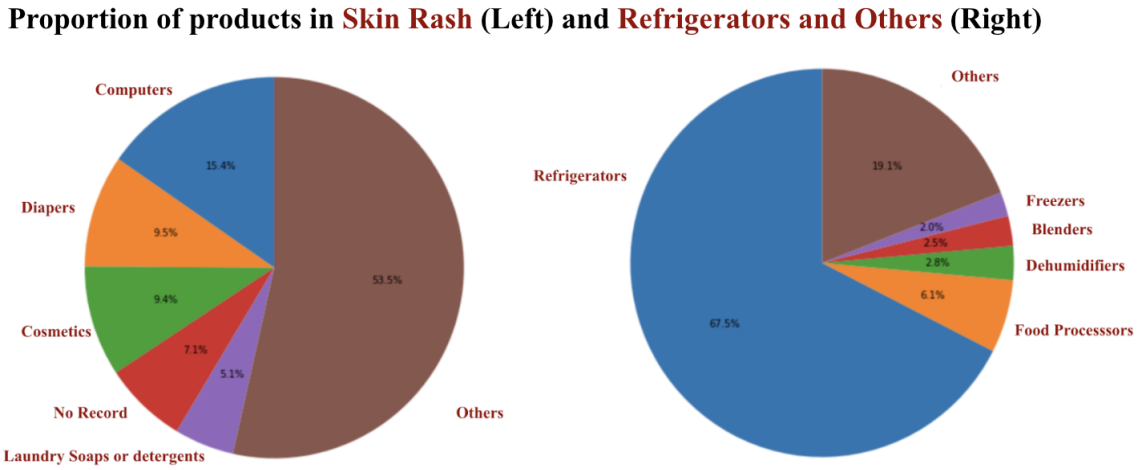

Figure 14. The left pie chart shows the proportion of products (top tive) in the topic Skin Rash; The right pie chart shows the proportion of products (top five) in the topic Refrigerators and Others.

With this comparison, we found that there is a topic type spectrum where the left is the pure product topic and the right is the pure condition topic, all 42 topics fall in between those two boundaries. For topics more similar to the condition type such as the topic "Skin Rash", the proportion of each product involved in this topic is spread more evenly (Figure 14). On the other hand, for topics more similar to product types such as the topic "refrigerators and others", one product will have a large proportion than the others (Figure 14). It seems to be impossible for us to obtain a pure product type product. Intuitively, this is because topics are determined by words and the same word can appear in different topics, thus, it is very hard to find a clear cut to differentiate similar topics since similar products might be categorized into the same topic due to their high similarity. As you can see, other products in the topic "refrigerator and others" such as food processors and blenders are very similar to refrigerators. Thus, we should be aware of this impurity when analyzing those topics. On the other hand, this impurity might help us unlock potential insights when we are dealing with structured data as it can group data based on that latent similarity that we might not be aware of. Hence, we don't need to go over thousands of product codes and manually check their potential similarity.

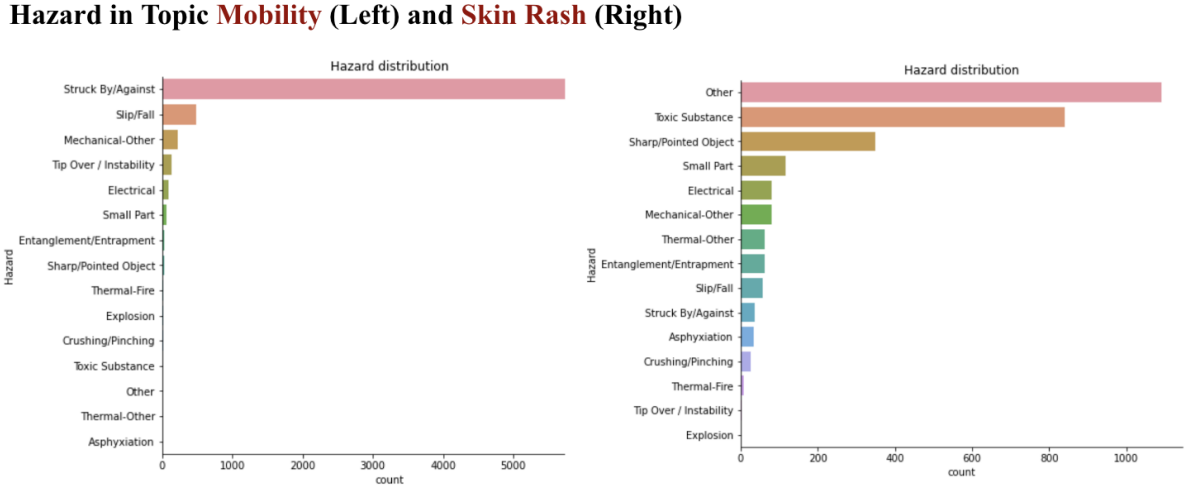

Figure 15. The bar plot on the ***left*** shows the hazard conditions involved in Mobility ordered by their frequency. The bar plot on the ***right*** shows the hazard conditions involved in Skin Rash.

Similarly, variables such as hazards and injury types also align with the expectation of topics. For example, in the topic “mobility”, the most frequent hazard is struck By/Against, this is a typical hazard of mobility accidents. Whereas in the topic “skin rash”, the most frequent hazard besides other is toxic substances.

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

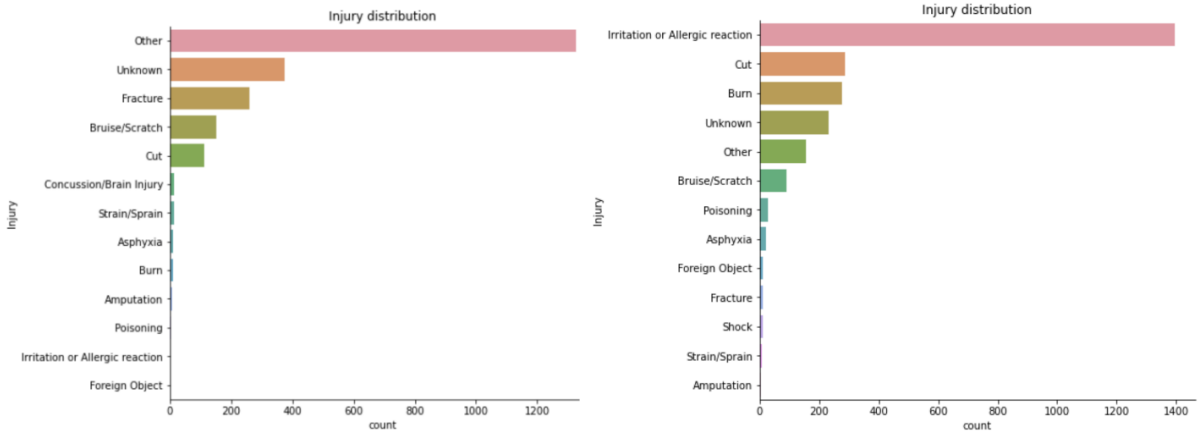

Figure 16. The bar plot on the ***left*** shows the injury type involved in Mobility ordered by their frequency. The bar plot on the ***right*** shows the injury type involved in Skin Rash.

In terms of injury types, “fracture” is the most frequent type besides “other” and “unknown”, which is a common injury type caused by mobility accidents in common sense. In the topic “skin rash”, the most frequent injury type is “irritation or Allergic reaction” which is also aligns with the meaning of skin rash.

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

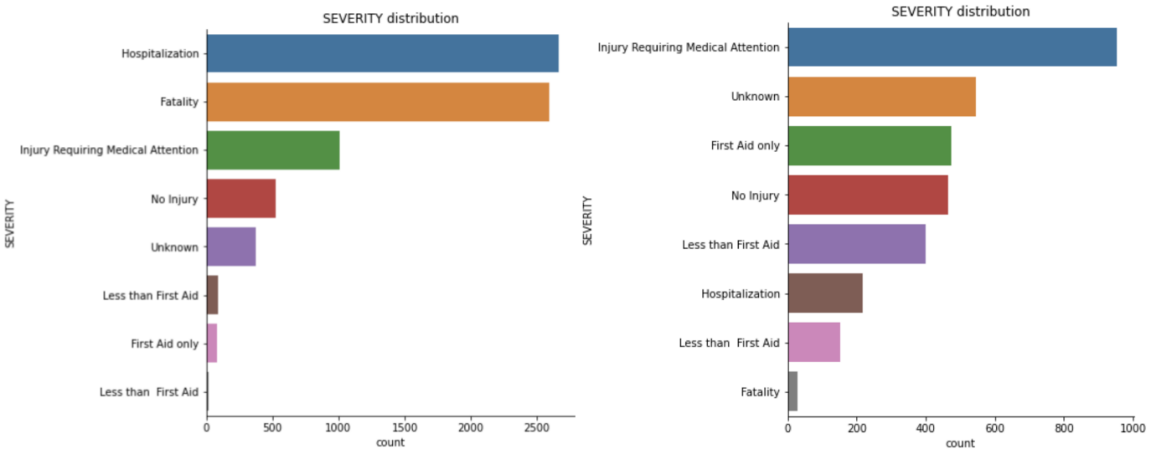

Figure 17. The bar plot on the ***left*** shows the severity involved in Mobility ordered by their frequency. The bar plot on the ***right*** shows the severity involved in Skin Rash.

Those structured variables such as severity can also help us better understand the topic. From this graph, we can see that “mobility” incident is commonly resulting in severe consequences. Whereas in the topic “skin rash”, consequences are milder.

Meanwhile, the structure data shows a similar trend as we found in the dynamic topic modeling. When we introduced the dynamic model, we found the keyword “Fitbit” and “watch” in the topic “skin rash” suddenly became very important around 2014, this has been validated by the structured data.

As you can see, there was a dramatic increase in “wrist” from 2013 to 2014 and it gradually decreased as the year passed. This is a shred of evidence that proves the dynamic model does capture the evolution of topic meaning within a topic. They both imply that there might be a massive product defect for wrist-related products since 2014.

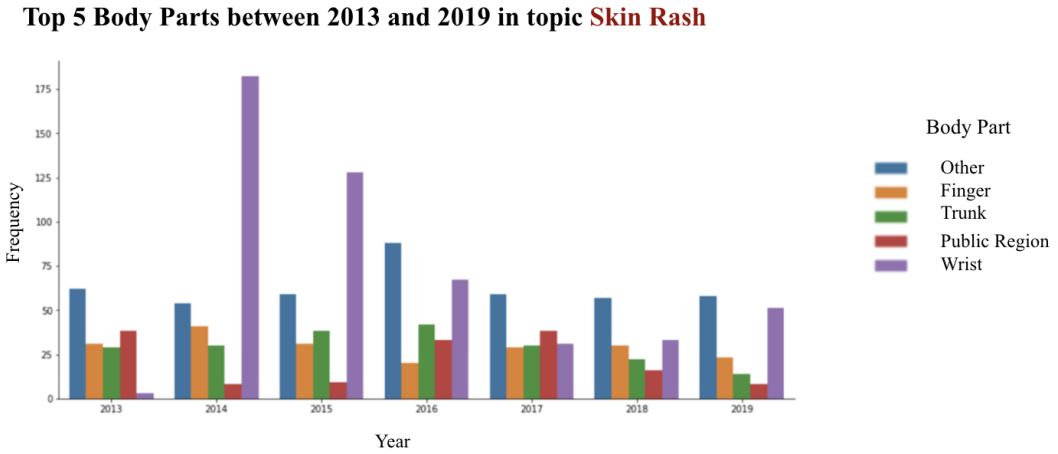

Figure 18. The top 5 body parts involved in topic Skin Rash between 2013 and 2019. Wrist suddenly increased in 2014.

Conclusion and discussion

In this article, we have presented the topic modeling procedures and demonstrated the effectiveness of BERTopic in analyzing narrative data. By applying the BERTopic modeling method, we have discovered 42 underlying topics within the 187,166 incident descriptions. Compared to manual investigation of each narrative, the identification of latent topics provides a more efficient way to comprehend the content of these narratives. Moreover, the dynamic perspective enables us to gain a comprehensive understanding of topic evolution over time. Through the comparison analysis, we have validated that the identified topics capture the latent meaning within the narratives and demonstrate their added value when working with structured data. However, we also acknowledge certain considerations that need to be addressed when conducting topic analysis.

In today's data-driven world, where data is continuously generated in various formats, traditional recording methods like narratives remain indispensable for many industries and human activities. Instead of seeking a more efficient substitute, machine learning methods such as BERTopic enable us to unlock the knowledge embedded within these narratives and preserve this valuable heritage amidst the data explosion era. By decoding this fundamental and primitive form of human record-keeping, we can harness the potential of machine learning to derive meaningful insights and perpetuate the legacy of human understanding.