



CAPTURING WHAT USERS GO TO CANADA.CA FOR

Alex Chen, Jan Urquico, Linxuan Yang, and Yundong Yao

Mentor: Jungyeul Park, University of British Columbia

Coordinators: Michael Hewlett, Canada Digital Analytics Team

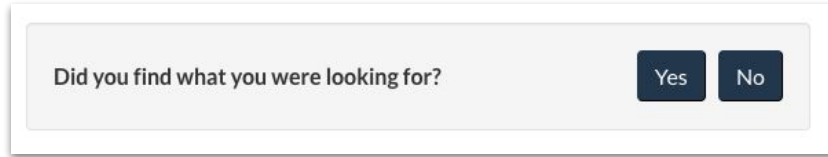
Capture What Users Go To CANADA.CA For

Content

1. Supervised Learning - Domains/Tags Classification
2. Unsupervised Learning - Feedback Clustering
 - a. Hierarchical Clustering
 - b. Topic Model
3. Linguistic Features
4. Conclusion and Future Work

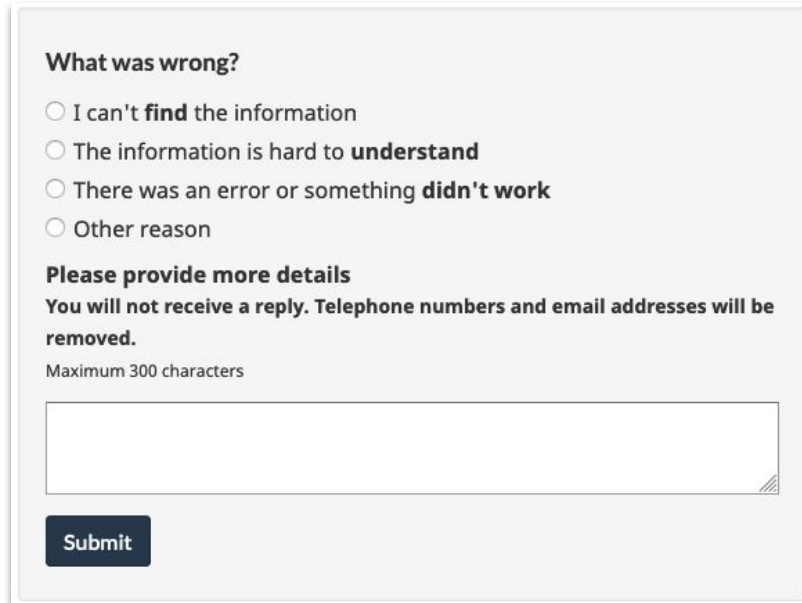
1. Supervised learning - Tags/Domains Classification

A feedback form:



Did you find what you were looking for?

Yes No



What was wrong?

☐ I can't **find** the information

☐ The information is hard to **understand**

☐ There was an error or something **didn't work**

☐ Other reason

Please provide more details

You will not receive a reply. Telephone numbers and email addresses will be removed.

Maximum 300 characters

Submit

Predict/Classify the tasks (tags) from visitors' feedback:

- Domains classification (Vaccine, Travel, Taxes, Benefits)
- Tags classification (Specific labels in each domain)

Feedback examples:

*What is delivery schedule for July Pfizer
moderne AstraZeneca and j&j*

Domain: Vaccine

Tags: Data and tracking vaccines

*Can I travel to the USA while I have a one dose
of vaccine?*

Domain: Travel

Tags: Restrictions or Requirements

1. Supervised learning - Tags/Domains Classification

Data (Feb - May, 2021):

Travel: 12K feedback

Vaccine: 29K feedback

Models:

Linear SVC (Support Vectors Machines): **Statistical model**. Fast and light

BERT: **Neural Network model**. Very accurate

Results (Accuracy):

Domains: **93%**

Vaccine: from 66% to **86%**

Travel: from 55% to **72%**

20% accuracy improvement compared to existing system of CDAT!

	Accuracy in Vaccine	Accuracy in Travel
CDAT Naive Bayes	0.66	0.55
OUR SVM	0.80	0.68
OUR BERT	0.86	0.72

2. Unsupervised learning

What: Without looking at the tags, the machine groups user feedback into clusters by itself, then generates keywords from these clusters.

Why: Predict new tags based on keywords

Examples:

- One group of 36 user feedback in Travel:
 - Keywords: “dog”, “pets”, “walk with”, “quarantine”
 - Example feedback:
“I was wondering if I can walk my dog off of my property during quarantine.”
- Another group of 324 user feedback in Vaccine:
 - Keywords: “second”, “2nd”, “second vaccines”
 - Example feedback:
“What about the second dose of madorna vaccine”

Models:

1. Hierarchical Clustering
2. Topic Modelling

2a. Unsupervised learning - Hierarchical Clustering

Main idea: let the machine group feedback into clusters using a bottom-up approach

Pros: No need to specify the number of clusters

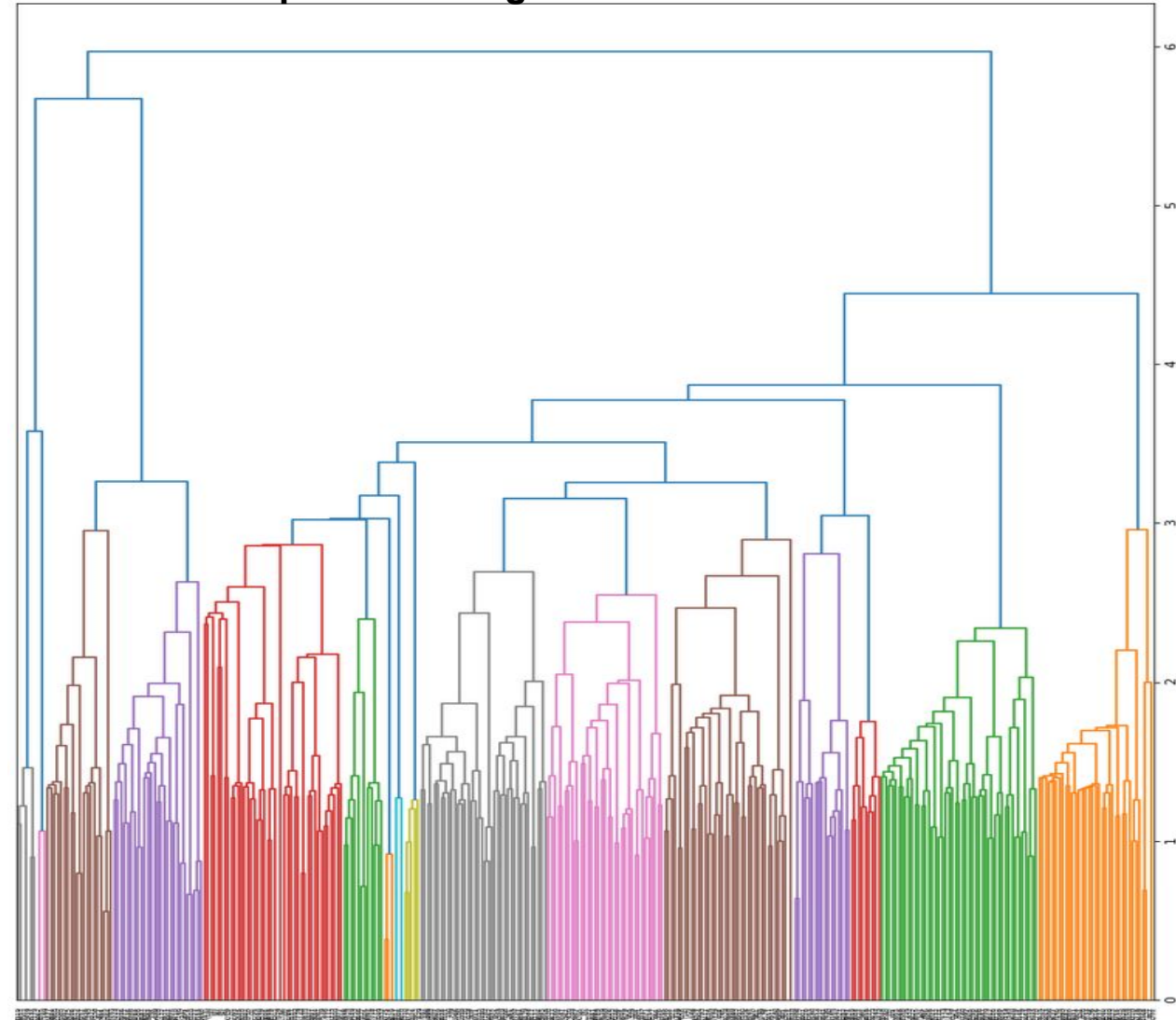
Cons: Computationally expensive

Solution: Work with a reasonably-sized sample instead of the full data set

Procedure:

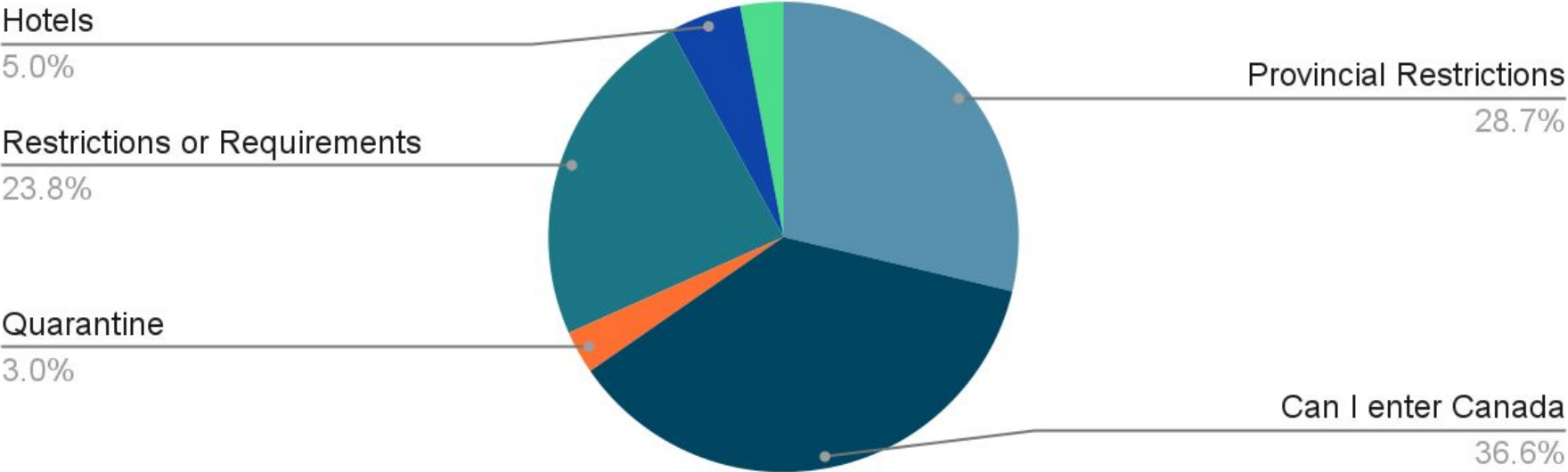
1. Cluster user feedback
2. Determine if cluster is good or bad
3. Predict new tags based on word pair frequency

Sample Clustering from the Travel Dataset



Example: Cluster 1 - Bad Cluster

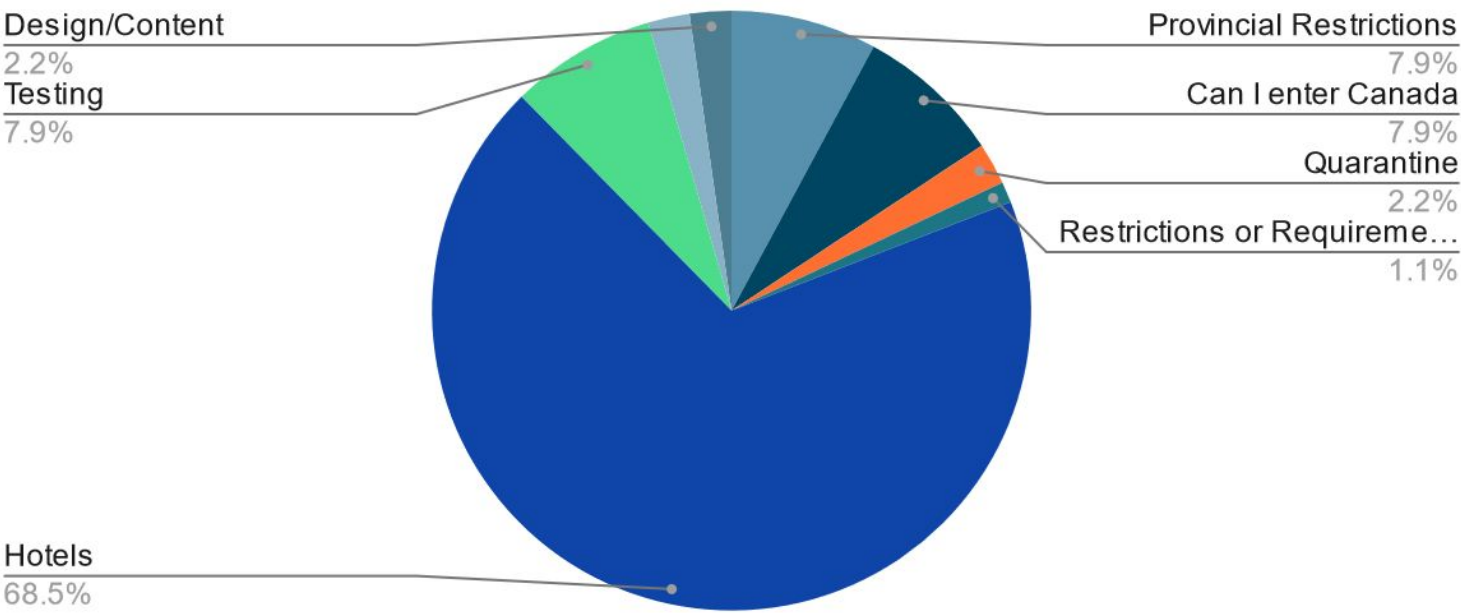
Tag Distribution



Machine-generated keywords	Possible tag	Example feedback
('travel', 'canada'), ('I', 'travel'), ('travel', 'within'), ('within', 'canada'), ('self', 'isolate')	'Traveling within Canada'	<i>travelling within canada</i> <i>travel within canada</i> <i>can a tourist travel though canada to alaska and back this summer ?</i> <i>can i travel through canada to move to alaska ?</i> <i>do i have to self isolate if i travel only in bc</i>

Example: Cluster 2 - Good Cluster

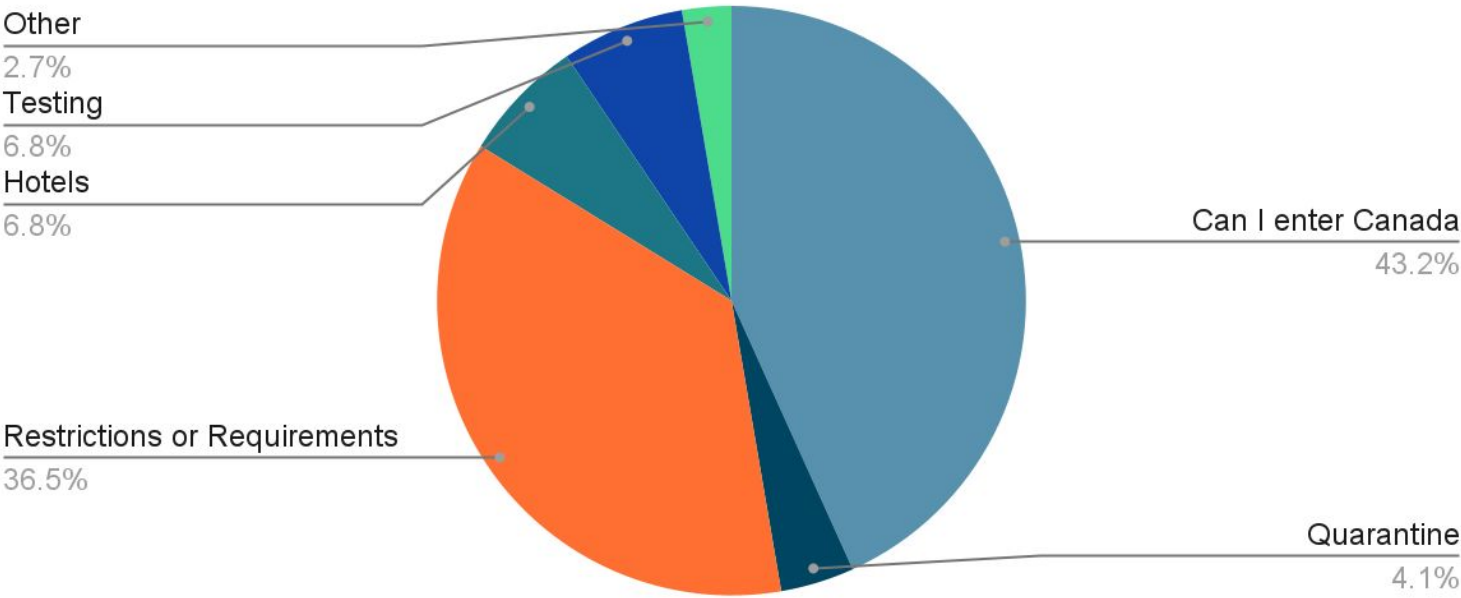
Tag Distribution



Machine-generated keywords	Possible tag	Example feedback
('hotel', 'quarantine'), ('hotel', 'stay'), ('need', 'hotel'), ('3', 'day'), ('?', '?')	'Hotels'	<div>it s not clear if i would need to do the hotel quarantine</div> <div>it doesn t address fully vaccinated travelers</div> <div>whom do i contact to receive an exemption from the hotel stay ? ? ? ? ? ? ?</div> <div>can we travel within our own province of ontario</div> <div>3 day hotel quarantine is a bunch of bull shit quarantine at home is very acceptable</div>

Example: Cluster 3 - Bad Cluster

Tag Distribution



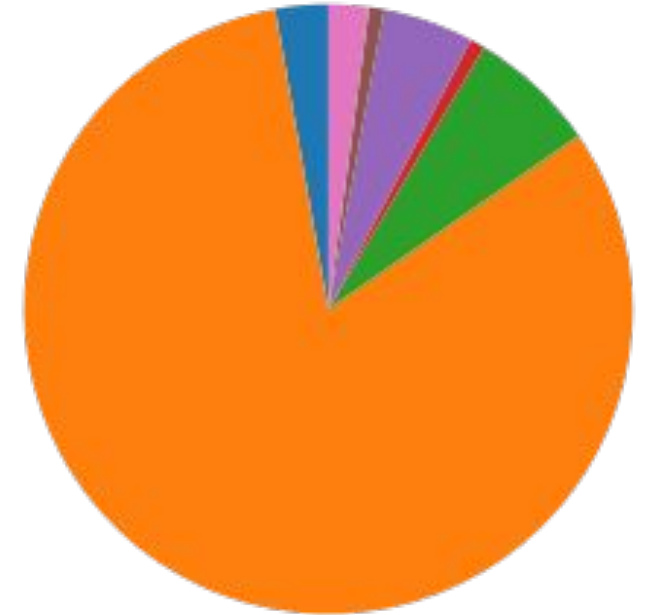
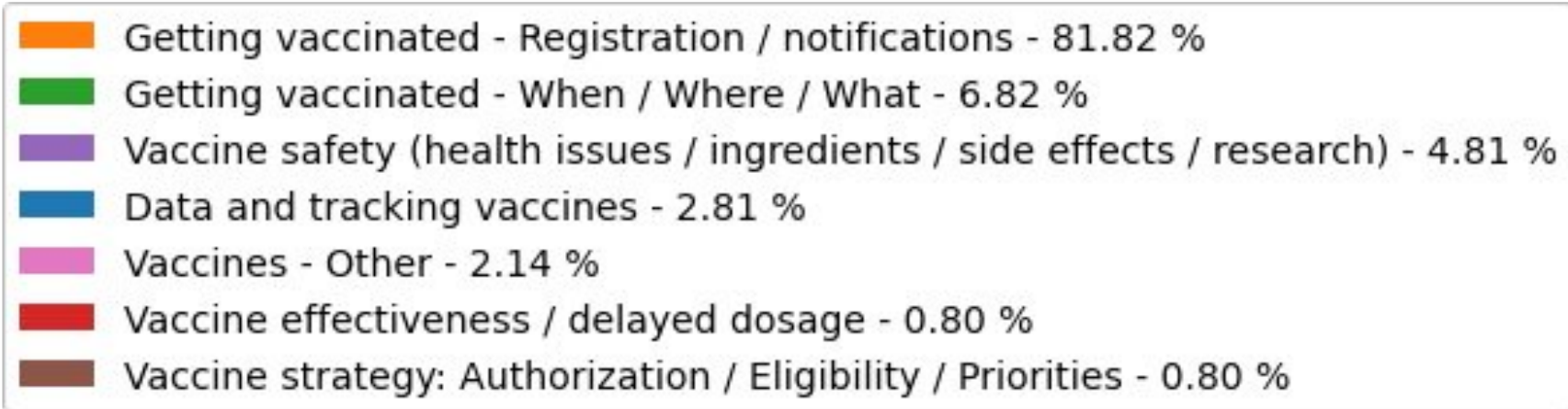
Machine-generated keywords	Possible tag	Example feedback
('fully', 'vaccinated'), ('fully', 'vaccinate'), ('receive', 'covid'), ('covid', '19'), ('I', 'fully'), ('I', 'receive'), ('enter', 'canada'), ('canada', '?'), ('still', 'need'), ('covid', 'vaccine')	'Fully vaccinated persons'	<i>what are the rules if you have been fully vaccinated would like to know if fully vaccinated citizens can enter from us without quarantining</i> <i>is there exemption for fully vaccinated travelers ?</i>

2b. Unsupervised learning - Topic Model: the case of vaccine feedback



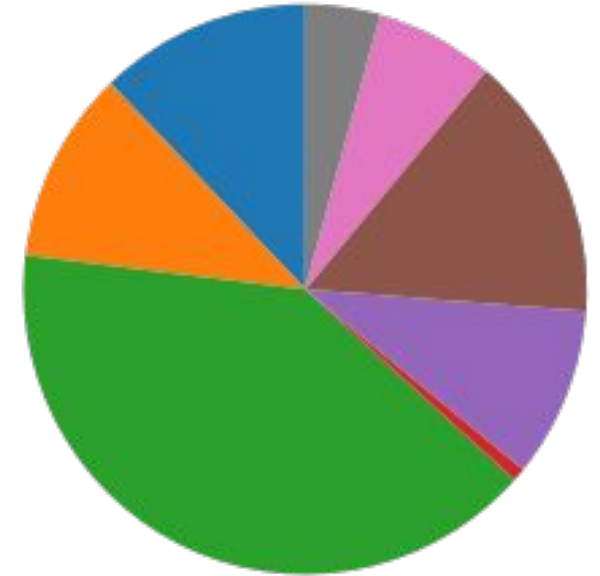
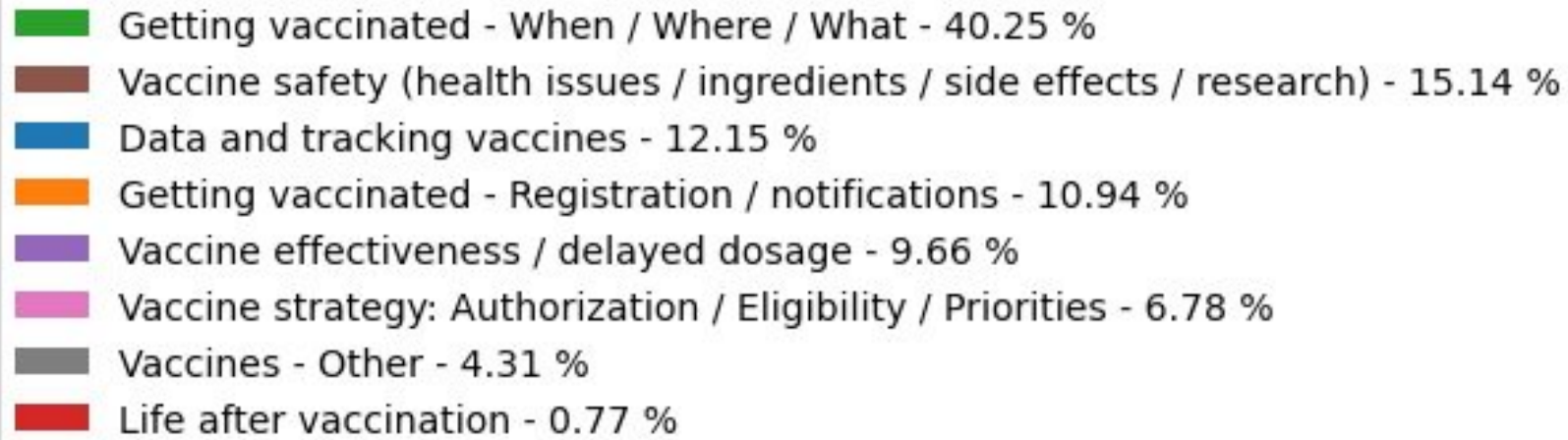
2b. Unsupervised learning - Topic Model

A “good” cluster



2b. Unsupervised learning - Topic Model

A “bad”(surprise) cluster



2b. Unsupervised learning - Topic Model

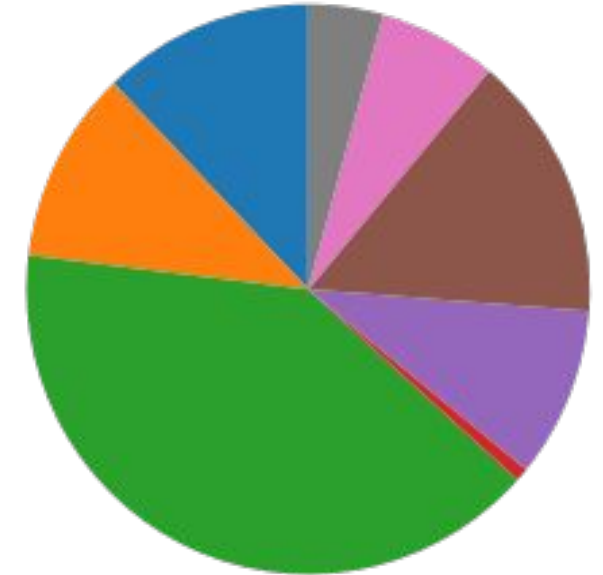
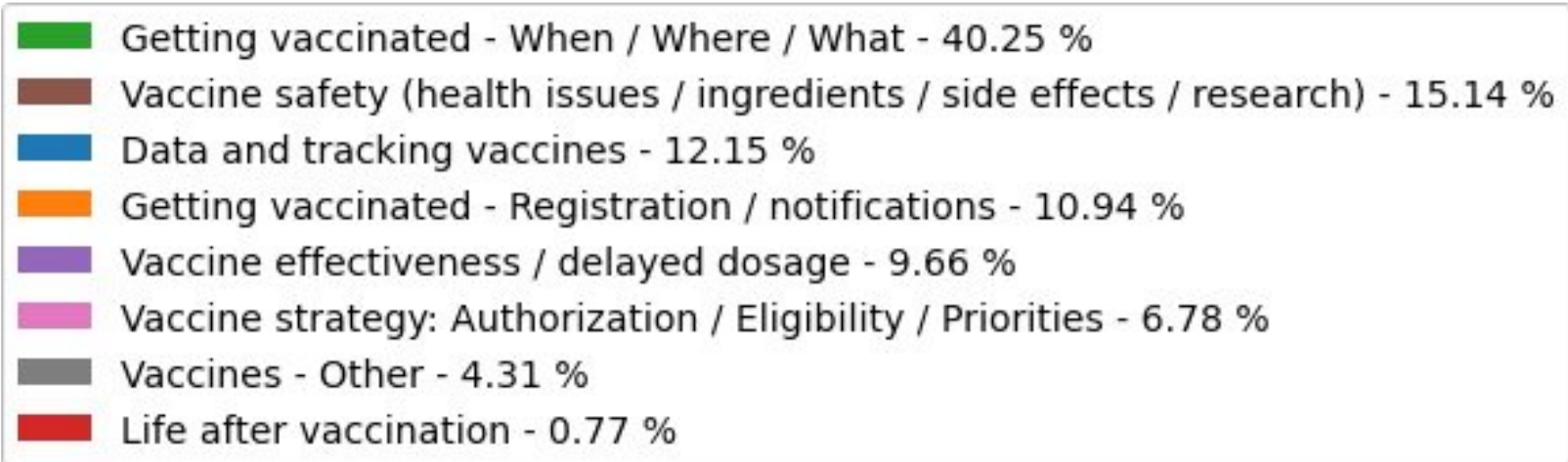
“Bad” clusters are the **useful** ones

Why?

- indicative of new perspectives
- **Useful for generating new tags**

2b. Unsupervised learning - Topic Model

A “Bad” or “surprise” cluster



machine-generated tags:

*receive, avail, people, does,
age_group, pfizer, **want_know**,
age, canada, vaccin*

Possible tagline:

*I want to know when people of my
age can receive pfizer vaccines?*

3. Linguistic Features

What are linguistic features?

Style, topic, tone, sentence structure, words

Why are they important?

We want to understand what people say

We also want the machine to understand what people say

What linguistic features do we have?

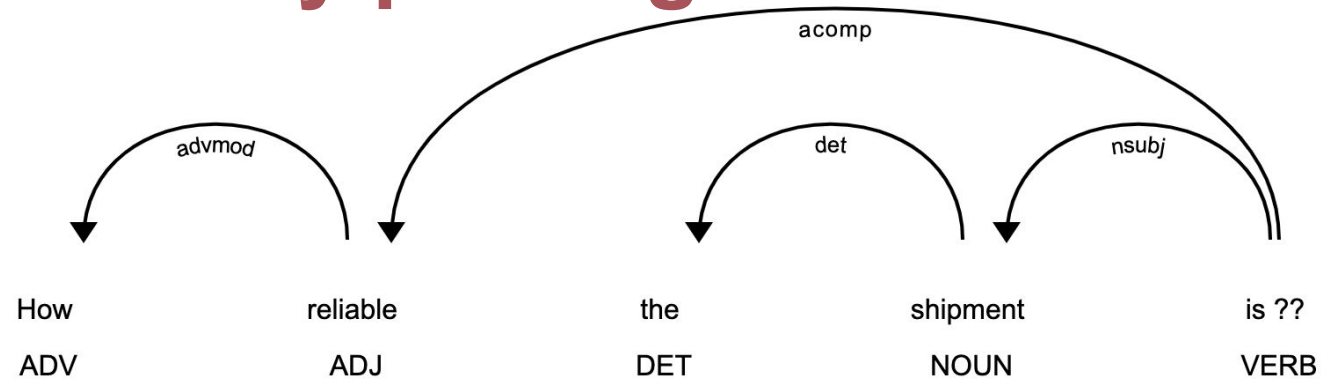
1. Word analysis: proper noun and keywords
2. Syntactic analysis: dependency parsing
3. Semantic analysis: semantic role labeling
4. Sentiment analysis

3. Linguistic Features

Word analysis: proper nouns (NER)

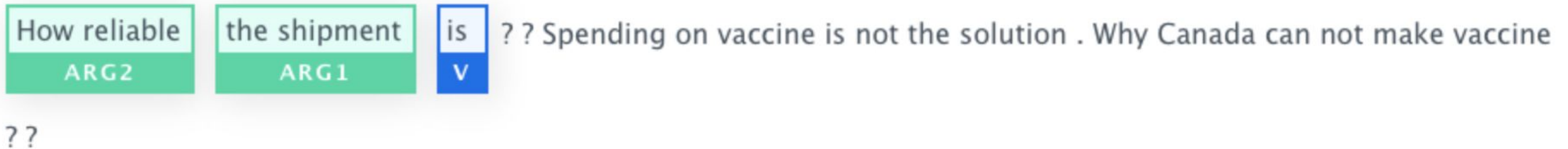
The data on this page seems to be inaccurate and does not line up with the reporting in the news. Also, the presentation and formatting is inconsistent. The Pfizer ORG data shows past distributions and allocations. With the Moderna ORG data, it is not clear whether the 10th-16th DATE allocation happened.

Syntactic analysis: dependency parsing



3. Linguistic Features

Semantic analysis: semantic role labeling



Sentiment analysis

How reliable the shipment is ?? Spending on vaccine is not the solution. Why Canada can not make vaccine ?? Aren't we a develop country ??? No technology at all !! Aren't this TRUDO GOVT. realize that Canada is to vulnerable when is come to safety ?? We need strong leadership. This guy NO GOOD

Sentiment: Negative
(-6.8)

3. Linguistic Features

Example of a feedback:

How reliable the shipment is ?? Spending on vaccine is not the solution. Why Canada can not make vaccine ??

Sentiment analysis: Happy? Not happy

Syntactic analysis: Not happy about what? Shipment!

Word analysis: What is “shipment”? Noun as a keyword

Semantic analysis: Why? Doubt the reliability

Next...
Ready for vectorization!
Ready to use!

4. Conclusion & Future Works

What have we got so far?

- Two classification models with 20% accuracy improvement
- Two unsupervised learning models to generate new tags from unlabelled feedback
- Four types of linguistic features to help machine “understand” text comments

What could we do next?

- Generate more insight from a linguistic perspective
- Develop API to automate routine works
- Design text database to handle growing data
- Real time analysis

Thank you!

Appendix: Comparing Unsupervised Methods

Model	Hierarchical Clustering	Topic Modeling
Pros	No need to specify number of clusters	Fast, interactive (next slide)
Cons	Computationally expensive	Need to specify number of clusters
Solutions	Use a sample instead of full data	Find ideal number of clusters (elbow method)

Appendix: Topic Model Visualization

Selected Topic:

Previous Topic

Next Topic

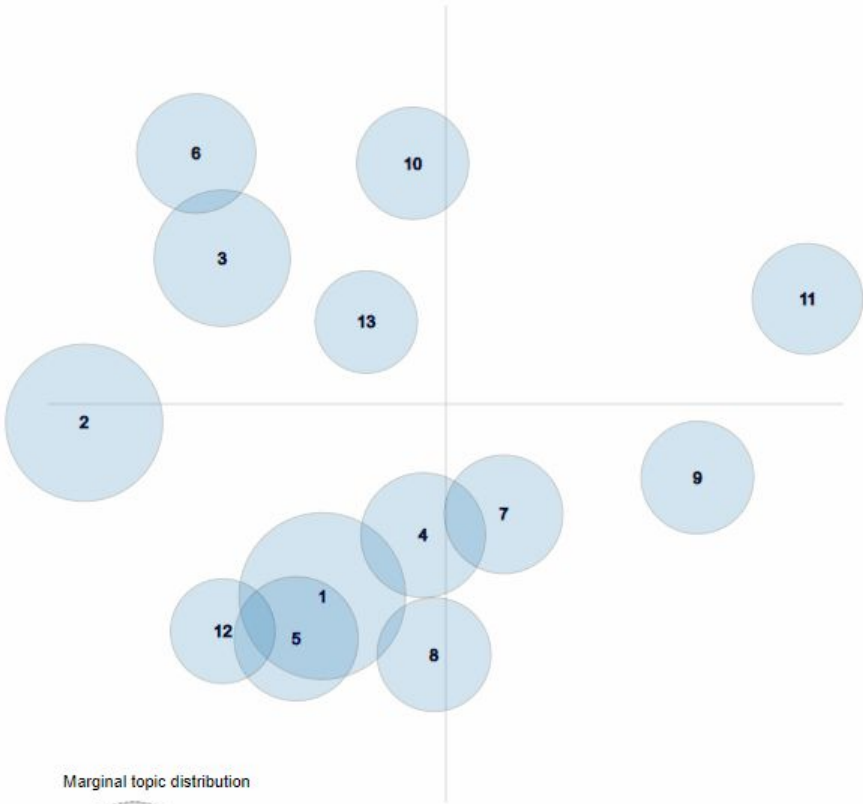
Clear Topic

Slide to adjust relevance metric:⁽²⁾

$\lambda = 1$

0.00.20.40.60.81.0

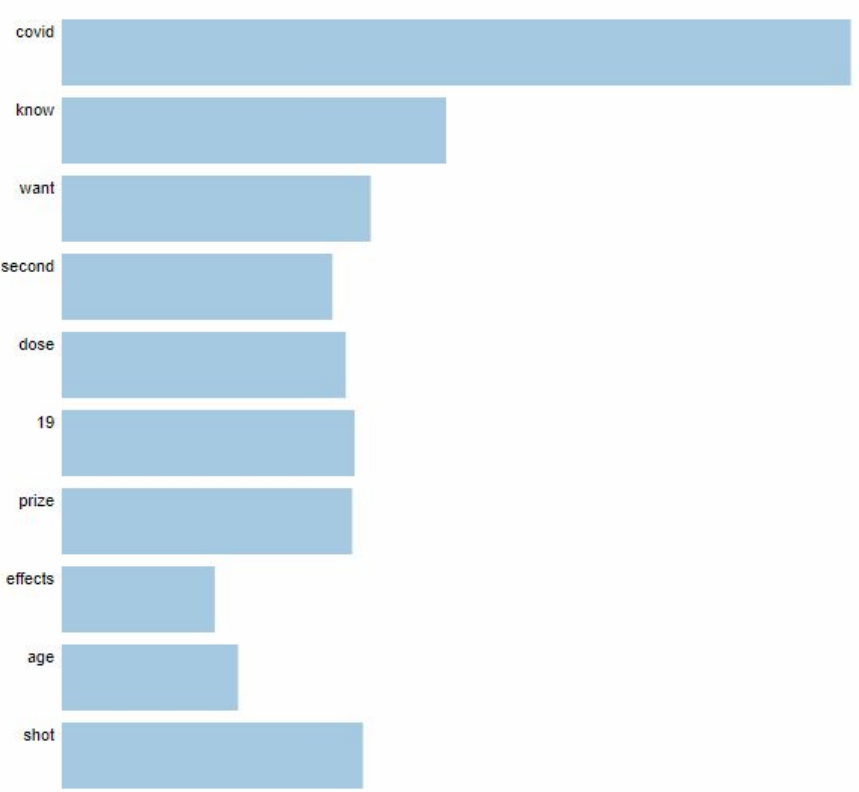
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-10 Most Salient Terms¹



Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t ; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)