HW 4 – Exploring Common Topics in Health Research News

Shivangi Mundhra | SUID – 842548148

For the three algorithms, kMeans, LDA and BERTopic, I did not sample the data from the given dataset of 10,000 headlines from health research press releases posted on the EurekAlert! Website. All headlines were included in all training all models to explore the common themes of health research news by applying kMeans, LDA, and BERTopic algorithms.

kMeans

Number of clusters –

I used k = 10 to begin with. It resulted in 10 clusters of which cluster #2 was the largest with 4,322 instances. The labels are as below. After using the elbow method to find the best k, I received k = 9 because the slope decreased at k = 9. I used this with sBERT and received below labelled clusters.

Labels –

cluster #	kMeans (k = 1(	sBERT (k = 9)
1	Coronavirus	Research and study findings
2	Tobacco	Depression and Mental Health
3	Unsure	Obesity
4	Brain	Pregnancy
5	Obesity	Cancer
6	Heart Health	Ifectious disease Pandemic
7	Breast cancer	Cognitive Diseases
8	Hysteria	Heart Health
9	Stroke	Tobacco
10	Heart Health	--

There was a lot of overlap in records between the clusters in kMean (10) algorithm but the sBERT model resulted in precise results. The records included in each cluster centered around a specific topic.

LDA

Number of clusters –

I used number of topics = 15.

Labels –

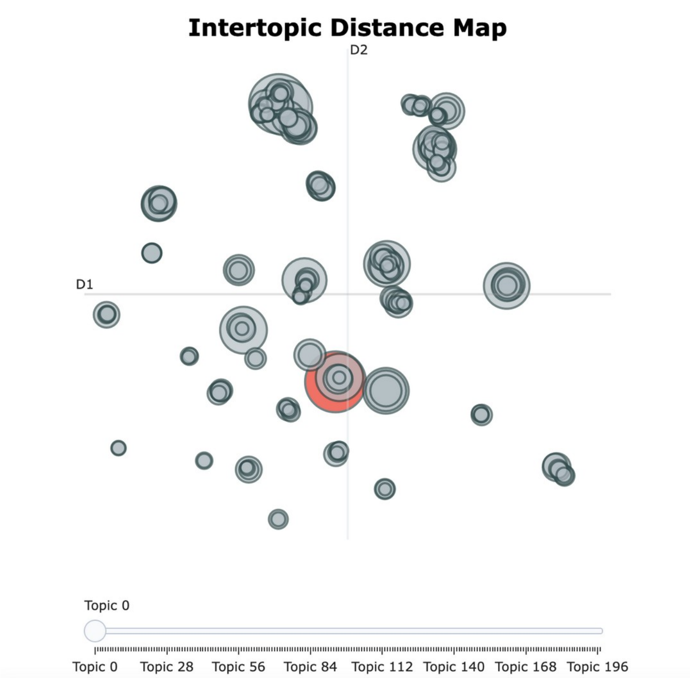
The below table shows the top 10 words for each label in the SVM model.

Topic#	Topic
1	Adult care
2	Research and findings
3	Heart Risks

4	Sickness caused by pollution
5	Orthopedic ailments
6	Cancer
7	STDs
8	Tests and breakthroughs
9	Organs
10	Blood related
11	Parenthood
12	Unsure
13	Covid-19
14	Meals
15	Brain

BERTopic

I received 198 topic using BERTopic. Obviously, it is difficult to label all of these. The below chart shows the intertopic Distance Map obtained from BERTopic.



Below is the table of labels.

topic#	topic
1	breast
2	dementia
3	preterm
4	diabetes
5	tobacco
6	gut
7	alcohol
8	prostate
9	heart
10	opioid
11	eye
12	hiv
13	antibiotic
14	heart
15	liver
16	tumor
17	vitamin
18	stroke
19	colorectal
20	lung
21	obesity
22	asthma
23	hip
24	cannabis
25	covid
26	food
27	brain
28	flu
29	autism
30	pressure
31	malaria
32	sleep
33	schizophreni
34	hepatitis
35	melanoma
36	ovarian
37	bariatric
38	sexual
39	exercise
40	kidney
41	tb
42	allergy
43	parkinson
44	plastic
45	sclerosis
46	sugar
47	suicide
48	pollution
49	cholesterol

topic#	topic
50	adhd
51	older
52	ptsd
53	migraine
54	pancreatic
55	arthritis
56	osteoarthriti
57	epilepsy
58	end
59	mental
60	errors
61	medical
62	trials
63	statins
64	hot
65	stimulation
66	apnea
67	back
68	pregnancy
69	preeclampsi
70	ebola
71	smoking
72	depression
73	fat
74	hpv
75	hearing
76	gun
77	dialysis
78	kidney
79	gestational
80	childhood
81	clots
82	mammograp
83	eating
84	meat
85	road
86	obesity
87	esophageal
88	neck
89	racial
90	ct
91	testosterone
92	readmission
93	delirium
94	dna
95	omega
96	countries
97	concussion
98	tooth

topic#	topic
99	leukemia
100	pulmonary
101	expectancy
102	sepsis
103	bladder
104	kidney
105	preterm
106	media
107	heat
108	cystic
109	breast
110	cervical
111	fibrillation
112	zika
113	arrest
114	traumatic
115	withdrawal
116	fruit
117	reconstructio
118	sleep
119	spinal
120	psoriasis
121	falls
122	breast
123	postpartum
124	exercise
125	pain
126	depressed
127	thyroid
128	hands
129	copd
130	ketamine
131	older
132	salt
133	nurses
134	fitness
135	medicaid
136	stents
137	breastfeedin
138	walking
139	gender
140	financial
141	older
142	work
143	erectile
144	organ
145	anesthesia
146	parasite
147	eczema

topic#	topic
148	transfusions
149	sitting
150	exercise
151	valve
152	violence
153	american
154	aneurysm
155	football
156	radiation
157	celiac
158	toxicity
159	dengue
160	secondhand
161	lymphoma
162	autoimmune
163	cocaine
164	incontinence
165	tamoxifen
166	alcohol
167	coffee
168	dismissing
169	appendicitis
170	hepatitis
171	clock
172	nsclc
173	mindfulness
174	particulate
175	violent
176	smokers
177	hiv
178	burnout
179	thinners
180	antidepressa
181	gout
182	radionuclide
183	knee
184	chocolate
185	muscular
186	fatigue
187	media
188	transplant
189	shoulder
190	coronavirus
191	pet
192	diet
193	abuse
194	pollution
195	body
196	rebuilding
197	obesity
198	depression

Shivangi Mundhra
SUID - 842548148

Ethic Statement

There are certainly overlaps in many of the clusters in probably all these models. We understand that all these health topics are sensitive and sometimes controversial. The idea is to create a best estimate for research and study purposes and not to slight any article about any disease that we may have missed to cast light on.

Conclusion

The effectiveness and precision by which BERTopic classifies all these headlines so fast is impressive. None of the other models produce such a fine distinction between headlines of topics. Plus BERTopic also eliminates the manual effort to guesstimate the number of clusters which makes it more desirable.