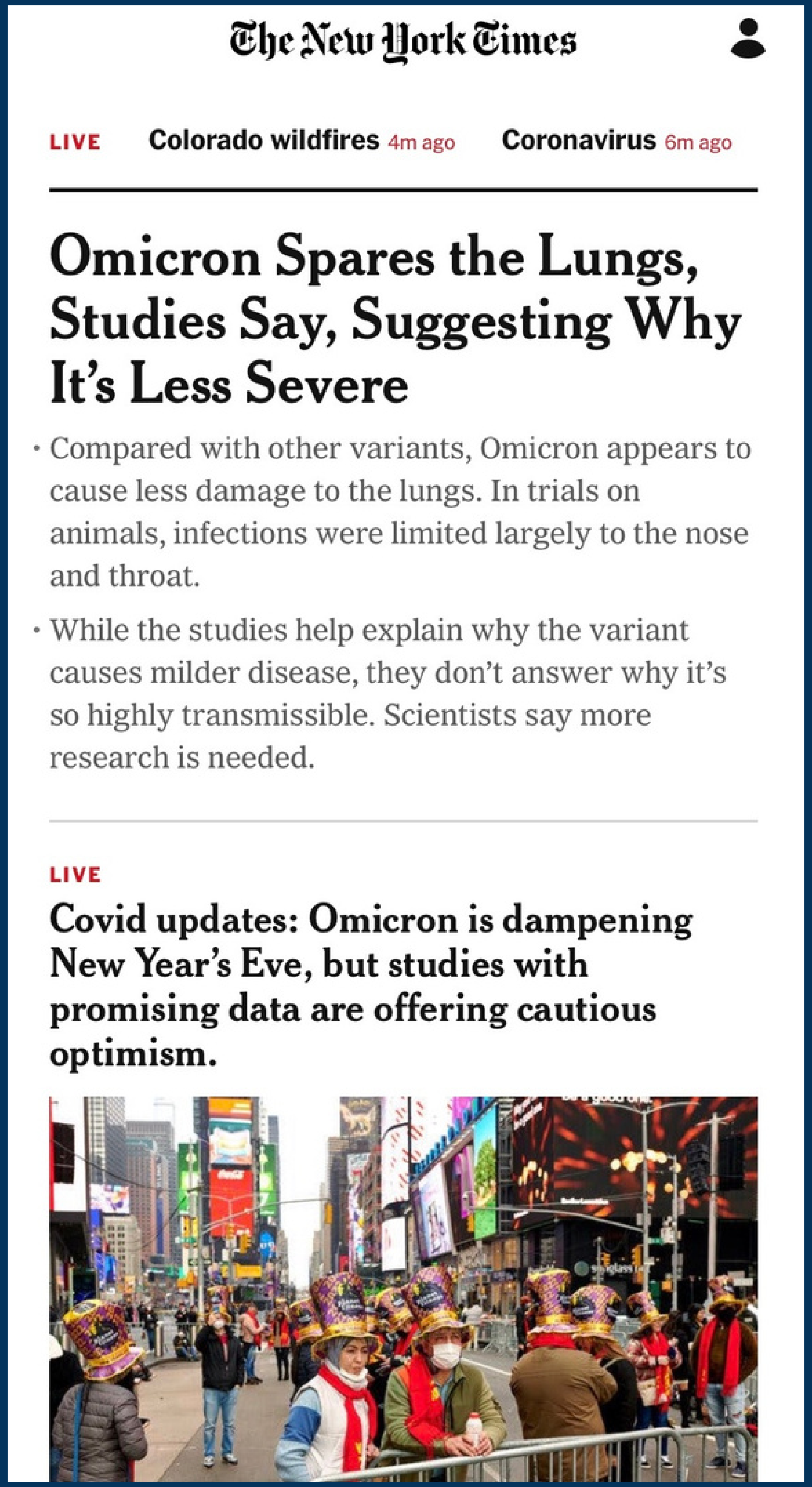
# NEWSGROUP TOPIC MODELING

# THE NEW YORK TIMES



The New York Times

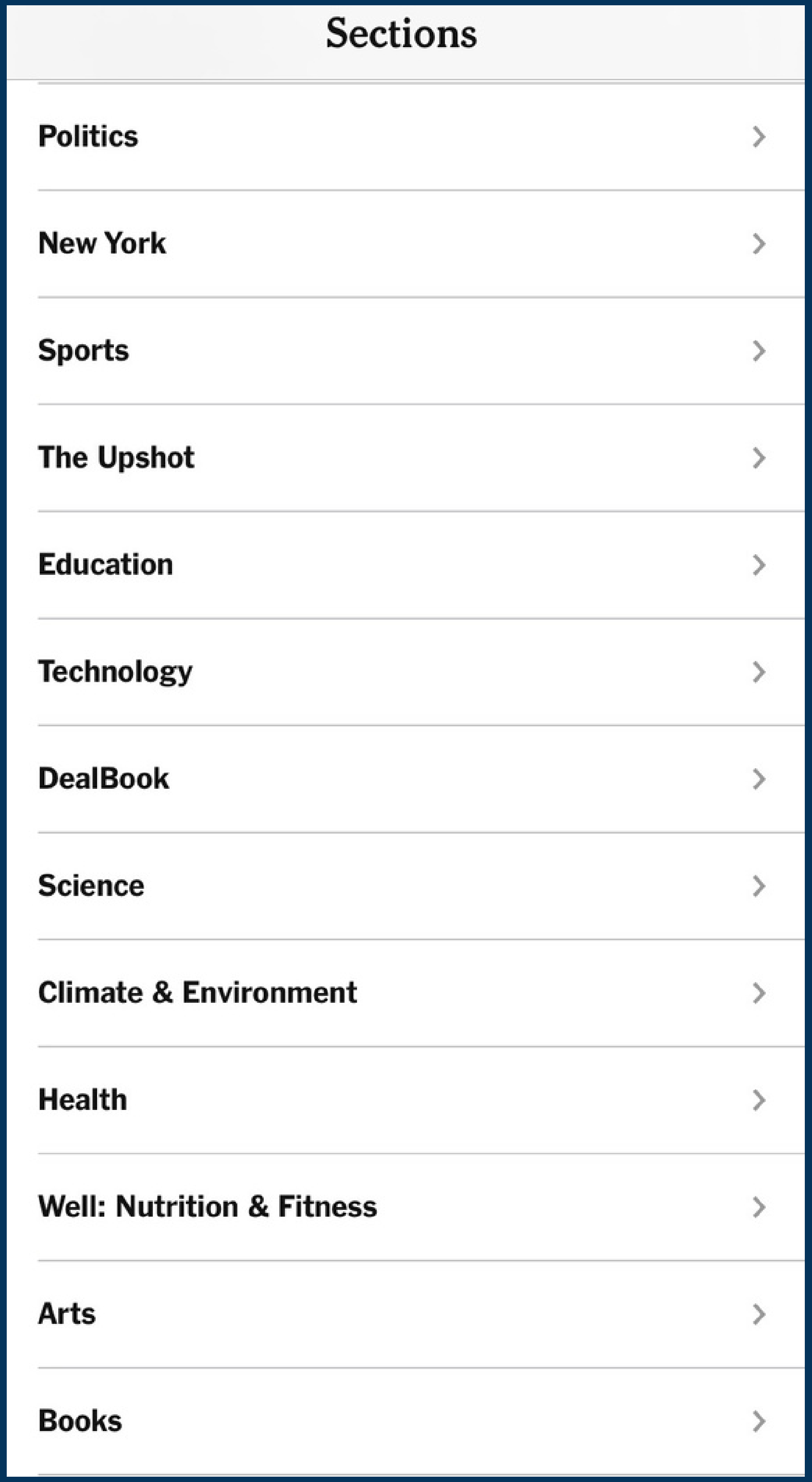
LIVE Colorado wildfires 4m ago Coronavirus 6m ago

## Omicron Spares the Lungs, Studies Say, Suggesting Why It's Less Severe

- Compared with other variants, Omicron appears to cause less damage to the lungs. In trials on animals, infections were limited largely to the nose and throat.
- While the studies help explain why the variant causes milder disease, they don't answer why it's so highly transmissible. Scientists say more research is needed.

LIVE

### Covid updates: Omicron is dampening New Year's Eve, but studies with promising data are offering cautious optimism.



Sections

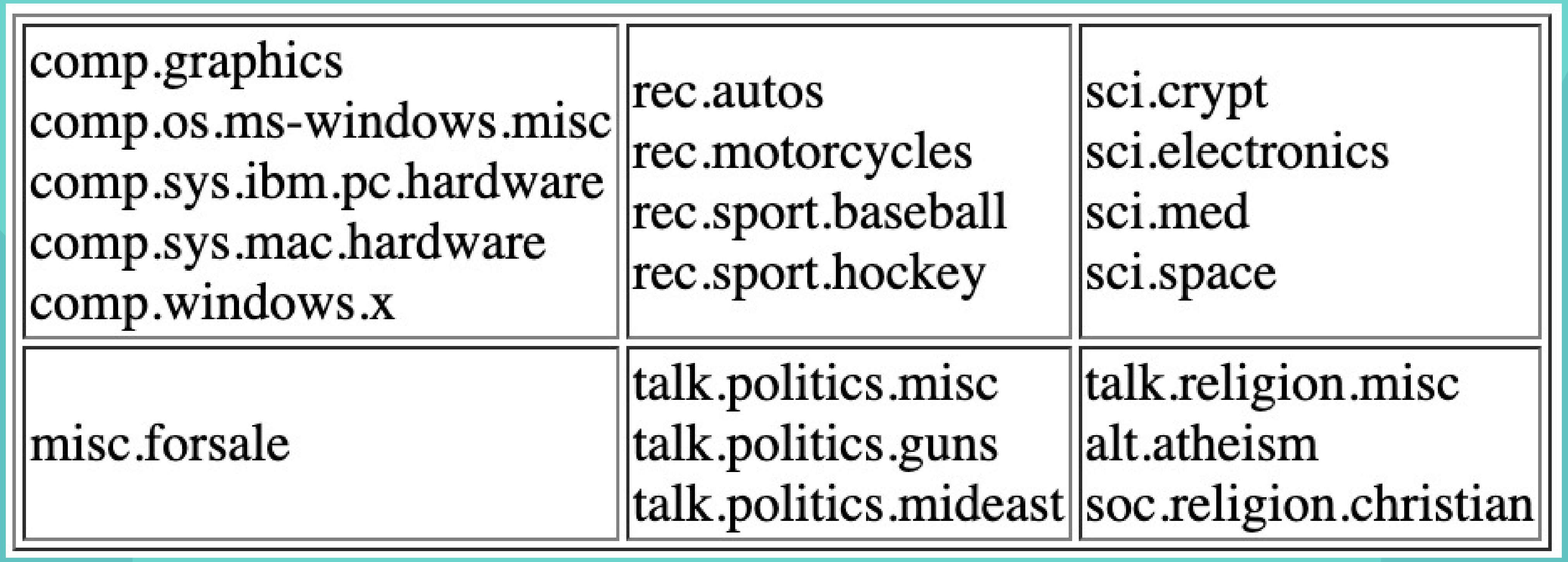
- Politics
- New York
- Sports
- The Upshot
- Education
- Technology
- DealBook
- Science
- Climate & Environment
- Health
- Well: Nutrition & Fitness
- Arts
- Books

# OUR HYPOTHESIS

A model can be derived to input news related language material, so it can reliably organize the material into coherent topics

# The Dataset

18,846 news documents manually labeled into twenty evenly distributed topics



| | | |
|---|---|---|
| comp.graphics<br>comp.os.ms-windows.misc<br>comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>comp.windows.x | rec.autos<br>rec.motorcycles<br>rec.sport.baseball<br>rec.sport.hockey | sci.crypt<br>sci.electronics<br>sci.med<br>sci.space |
| misc.forsale | talk.politics.misc<br>talk.politics.guns<br>talk.politics.mideast | talk.religion.misc<br>alt.atheism<br>soc.religion.christian |

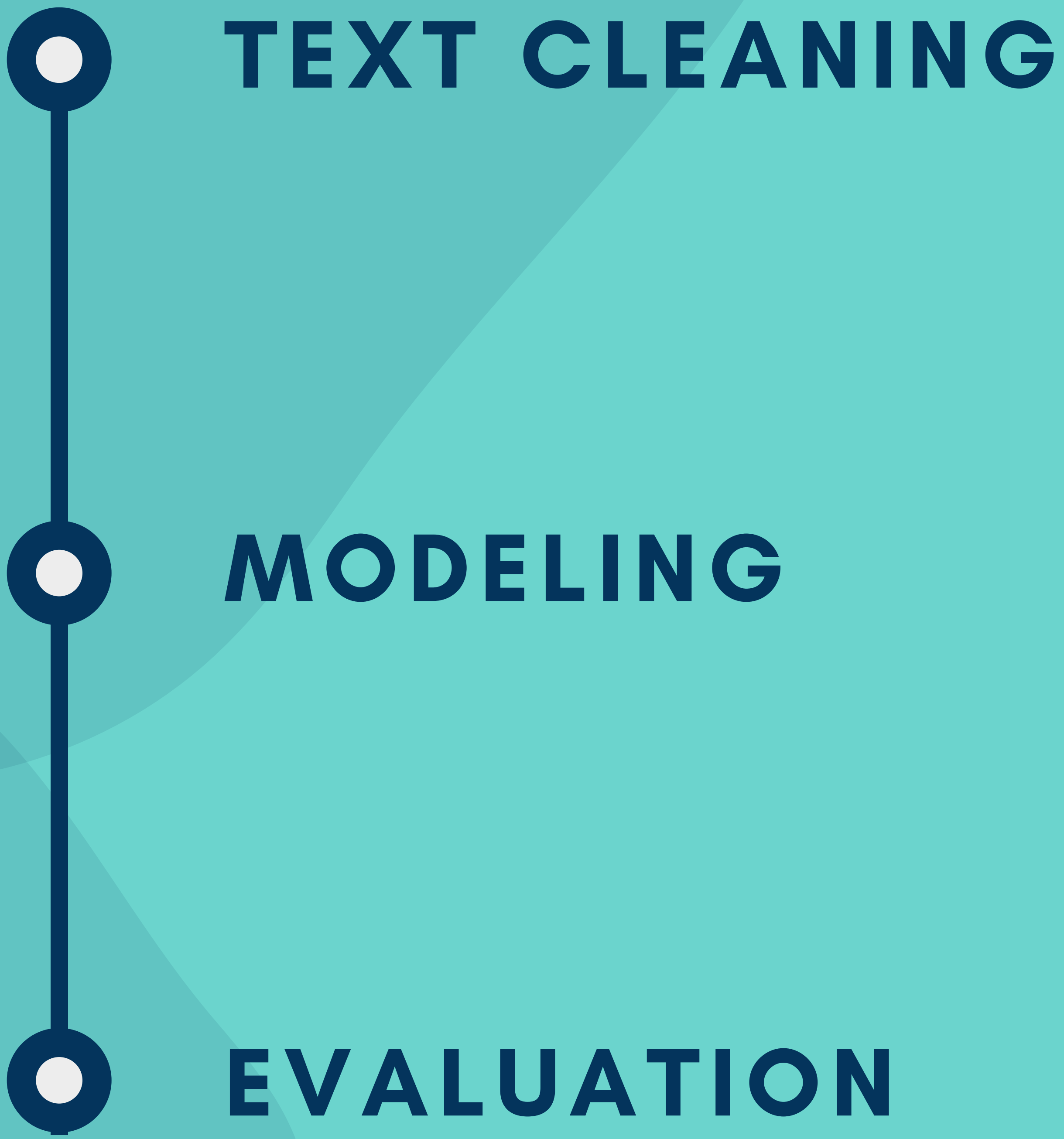
# Example Docs

## comp.sys.ibm.pc.hardware

'MY BROTHER IS IN THE MARKET FOR A HIGH-PERFORMANCE VIDEO CARD THAT SUPPORTS\NVESA LOCAL BUS WITH 1-2MB RAM.  DOES ANYONE HAVE SUGGESTIONS/IDEAS ON:\N\N  - DIAMOND STEALTH PRO LOCAL BUS\N\N  - ORCHID FARENHEIT 1280\N\N  - ATI GRAPHICS ULTRA PRO\N\N  - ANY OTHER HIGH-PERFORMANCE VLB CARD\N\N\NPLEASE POST OR EMAIL.  THANK YOU!\N\N  - MATT\N'

## sci.med

'IT IS NOT TRUE THAT DERMATOLOGISTS GAVE NOT REACHED THE LASER AGE, IN\NFACT, LASERS IN DERMATOLOGICAL SURGERY IS A VERY NEW AND EXCITING FIELD.\N\NIT PROBABLY WON'T BE EFFECTIVE IN TINEA PEDIS BECAUSE THE LASER IS\NUSUALLY A SUPERFICIAL BURN (TO AVOID ANY DEEPER DAMAGE). LIMITED TINEA\NPEDIS CAN BE CURED ALBEIT SOMETIMES SLOWLY BY TOPICAL ANTIFUNGALS AS\NWELL AS SYSTEMIC MEDICATION I.E. TABLETS. FINALLY, A SELF-DIAGNOSIS IS\NNOT ALWAYS RELIABLE, LICHEN SIMPLEX CHRONICUS CAN LOOK LIKE A FUNGAL\NINFECTION AND REQUIRES VERY DIFFERENT TREATMENT.'



TEXT CLEANING

MODELING

EVALUATION

## TEXT CLEANING

- "\n\nIt probably won't be effective in tinea pedis because the laser is\nusually a superficial burn (to avoid any deeper damage)."

## MODELING

- Goal: organize documents into coherent topics
- Method: model analyzes numbers converted from text

## EVALUATION

- 90% success rate in matching the model's topics with original labeled topics

# EVALUATION OF PREDICTED TOPICS

| ORIGINAL | PREDICTED |
|---|---|
| comp.sys.mac.hardware | card, monitor, video, bus, color, mb, bit, video card |
| misc.forsale | sale, price, offer, new, email, please, condition, interested |
| rec.sport.baseball | team, year, player, last, season, last year, league, play |
| sci.space | space, nasa, mission, science, orbit, launch, cost, station |
| soc.religion.christian | god, jesus, christian, bible, church, christ, sin |
| talk.politics.mideast | israeli, israel, jew, arab, palestinian, jewish, state |
| talk.politics.misc | law, government, gun, state, crime, federal, right |

# Original Hypothesis

***A model can be derived to input news related language material, so it can reliably organize the material into coherent topics***

- Assigned a descriptive and coherent topic at 90% rate
- The model discovered key words for topic creation/assignment, e.g.:

Card

Price

Team

Space

God

Law, Israeli

# Thank You!