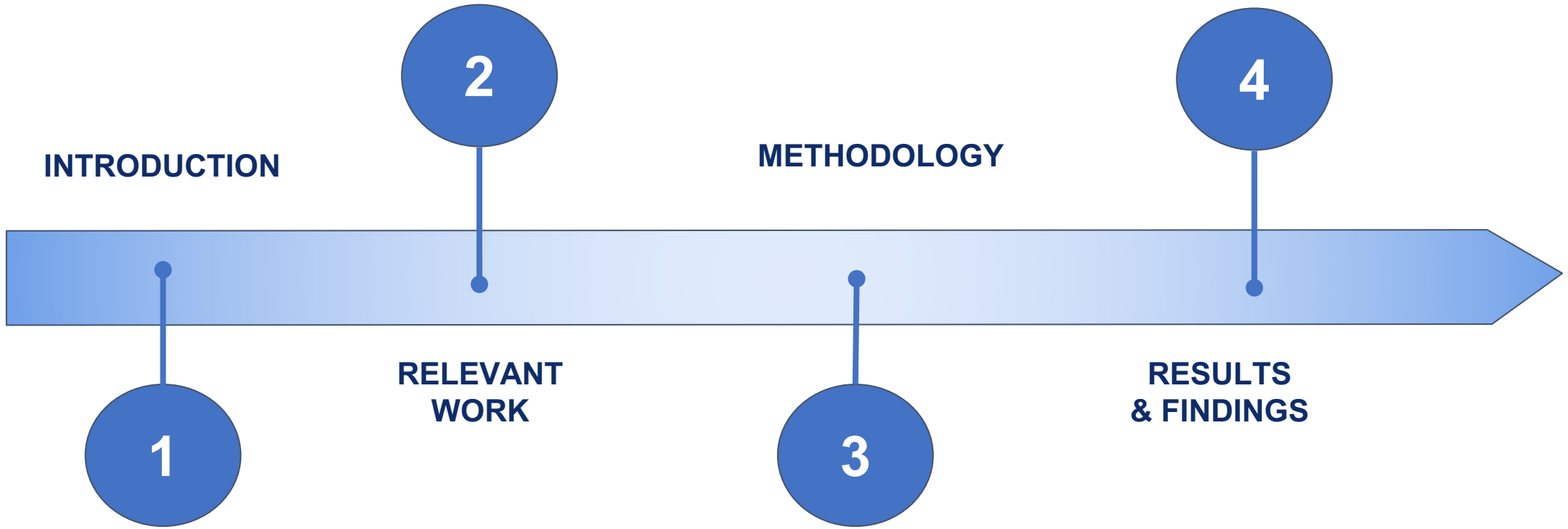
Faculty of Computer Science

Data Science and Business Analytics

Moscow 2023

Application of Machine Learning Models for the Analysis and Prediction of Socio-Behavioural Trends for Business

Research paper
Prepared by Dzhkha Anika
Supervised by Dimova Elena


1
INTRODUCTION
2
RELEVANT WORK
3
METHODOLOGY
4
RESULTS & FINDINGS

1 INTRODUCTION

What are Socio-Behavioural Trends?

How to effectively apply **machine learning models** to **analyze** and **predict socio-behavioural trends** for **business**?

Relevance & significance

- New **data-driven method** for predicting socio-behavioural trends
- The approach equips businesses with the ability to anticipate **customer preferences**, forecast **market shifts**, and optimize **decision-making**

2

OVERVIEW OF RELEVANT WORK

Books on Analyzing Socio-Behavioral Trends

Following books valuable insights into **understanding evolving consumer behavior**, leveraging social influence, and adapting marketing approaches in the digital age:

- "Marketing in the Moment" (Tasner, 2011): Real-time marketing, Web 3.0, importance of machine learning
- "Contagious" (Berger, 2013): Viral content, social influence, word-of-mouth marketing, STEPPS
- "Thinking, Fast and Slow" (Kahneman, 2011): Cognitive biases, decision-making, consumer behavior
- "Consumer Behavior" (Solomon and Armstrong, 2019): Psychological factors, emerging trends, marketing effectiveness
- "The Power of Habit" (Duhigg, 2012): influence of habits on human behaviour

"A Topic Modelling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts" (Egger, R., Yu, 2022)

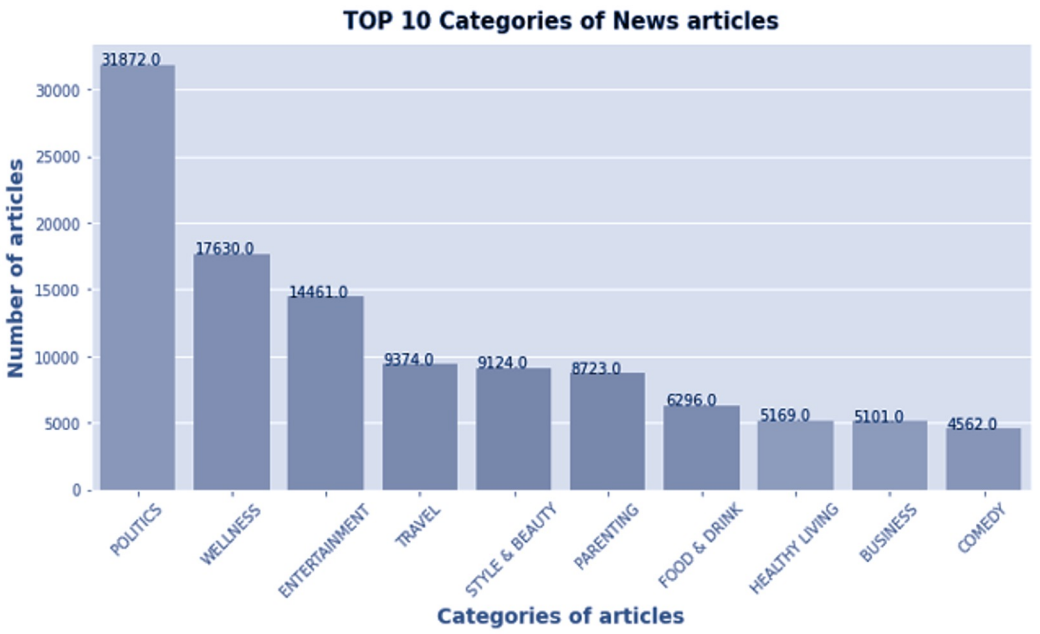
Method	Advantages	Disadvantages
LDA	• No prior domain knowledge required • Finds coherent topics with tuning • Easy interpretation • Full generative models	• Detailed assumptions required • Hyperparameters tuning needed • Overlapping topics • User-defined number of topics • Assumes independent topics, ignores word correlations
NMF	• No prior domain knowledge required • Can use TF-IDF weighting • Computationally efficient • Easy implementation	• Frequently delivers incoherent topics • User-defined number of topics • Implicit generative models

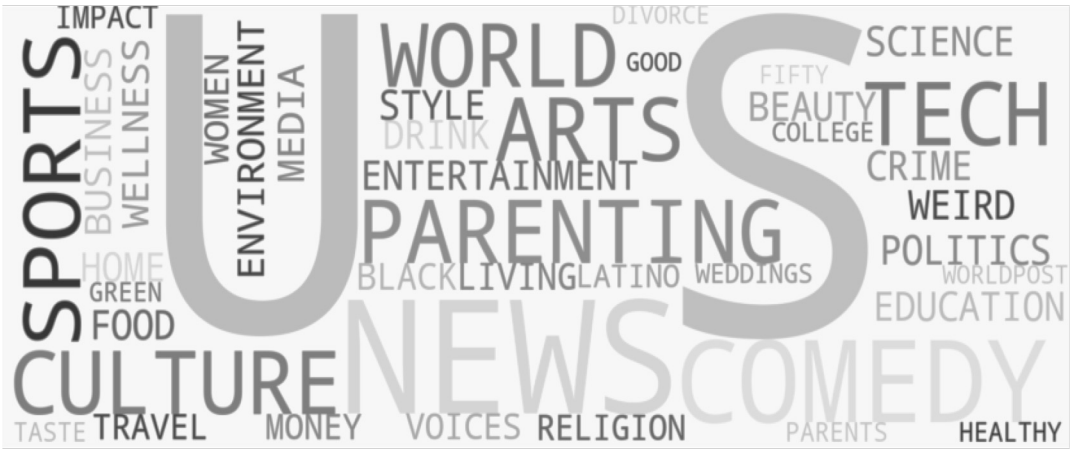
Method	Advantages	Disadvantages
Top2Vec	• Supports hierarchical topic reduction • Multilingual analysis • Automatic number of topics • Built-in search functions • Scalable on large datasets • No preprocessing of original data	• Potential for too many topics • Generates outliers • Not suitable for small datasets • Each document assigned to one topic
BERTopic	• Advantages of Top2Vec • Versatile and stable across domains • Supports topic modeling variations • Built-in search functions • Broader support of embedding models	• Potential for too many topics • Generates outliers • No topic distributions within a document

3 METHODOLOGY

Dataset Overview

- The **News Category Dataset** obtained from Kaggle, containing approximately 210,000 news in English headlines from 2012 to 2022.
- One of the largest comprehensive news datasets available.
- Covering various categories, including politics, sports, entertainment, technology, and more.
- Contains dates, which facilitates better analysis.
- 41 distinct categories of news articles.
- The total amount of news from 2019 to 2022 is significantly lower than from 2012 to 2019.
- Politics is the most common category of news in the dataset. Wellness and Entertainment are taking 2nd and 3d places.





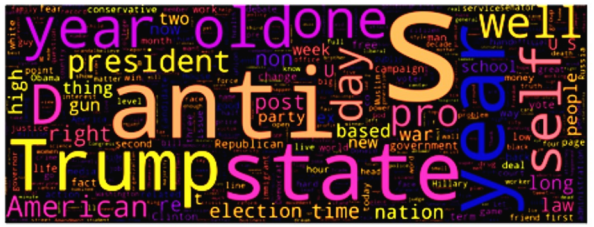
Dataset Overview: sample

	Headline	Category	Short description	Date
0	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	23.09.2022
1	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	23.09.2022
2	23 Of The Funniest Tweets About Cats And Dogs .	COMEDY	"Until you have a dog you don't understand wha...	23.09.2022
3	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	23.09.2022

The 4 variables that are involved in research shown above as a sample.

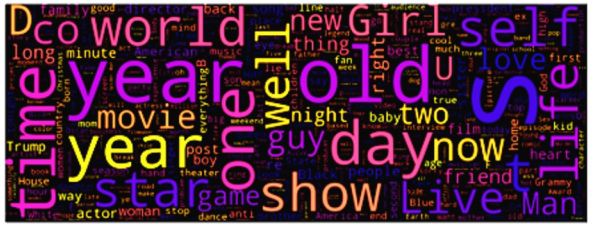
Dataset Overview: “wordclouds” for popular categories, highlighting highly used words

Wordcloud for news topic POLITICS



Wordcloud for news topic WELLNESS

Wordcloud for news topic ENTERTAINMENT



Wordcloud for news topic TRAVEL

Wordcloud for news topic STYLE & BEAUTY

Wordcloud for news topic PARENTING

Wordcloud for news topic FOOD & DRINK

Wordcloud for news topic HEALTHY LIVING

Wordcloud for news topic BUSINESS

Wordcloud for news topic COMEDY

Methodology

NLP technique to identify such trends in an effective way - **topic modelling.**

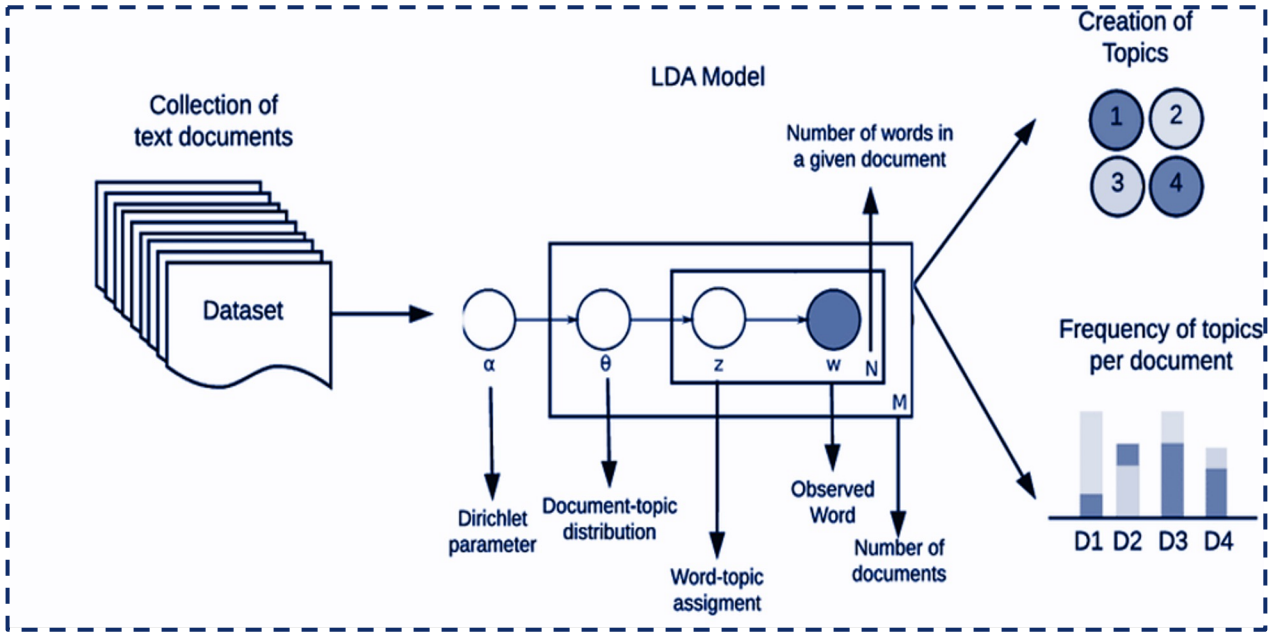
- **Preprocessing** of textual data was done, including processes of unnecessary words removal, lemmatization.
- Latent Dirichlet Allocation serves as the **baseline method** for traditional probabilistic topic modelling.
- BERTopic algorithm as the **primary method** with a proven efficiency over other Topic models.
- Both models have their strengths and weaknesses, and the findings require careful qualitative interpretation.
- The model effectiveness evaluated using **coherence scores** to assess the interpretability of the generated topics.
- Insights could be gained from both topics and graphs.

Preprocessing

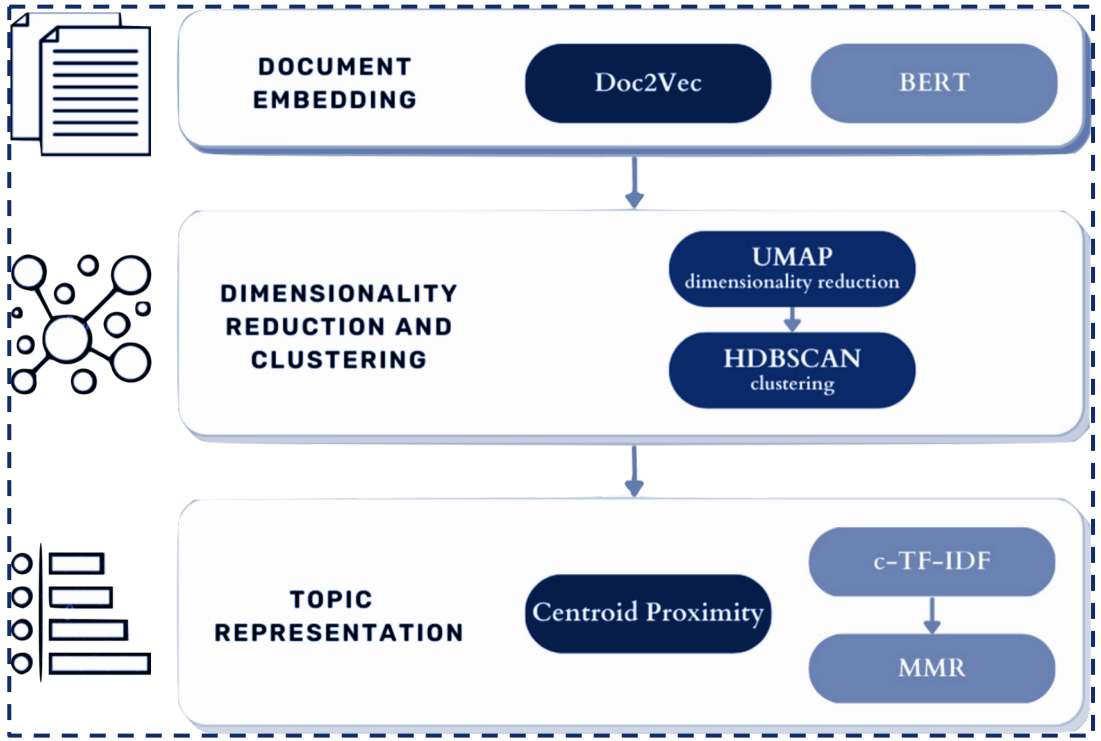
BEFORE	AFTER
Health experts said it is too early to predict whether demand would match up with the 171 million doses of the new boosters the U.S. ordered for the fall.	health expert say early predict whether demand would match million dose new booster order fall
He was subdued by passengers and crew when he fled to the back of the aircraft after the confrontation, according to the U.S. attorney's office in Los Angeles.	subdued passenger crew flee back aircraft confrontation accord attorney office los angeles
"Until you have a dog you don't understand what could be eaten."	dog understand could eaten

LDA & BERT

Latent Dirichlet Allocation



Top2Vec **BERTopic**



LDA & BERT

Requirements for LDA include:

- A bag-of-words representation
- The assumption of a Dirichlet prior
- The need to specify the number of topics in advance

BERTopic requires:

- a pre-trained BERT model
- works with document embeddings
- utilizes clustering and embedding techniques

TF-IDF formula: $W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{tf_t}\right)$

Evaluation metric

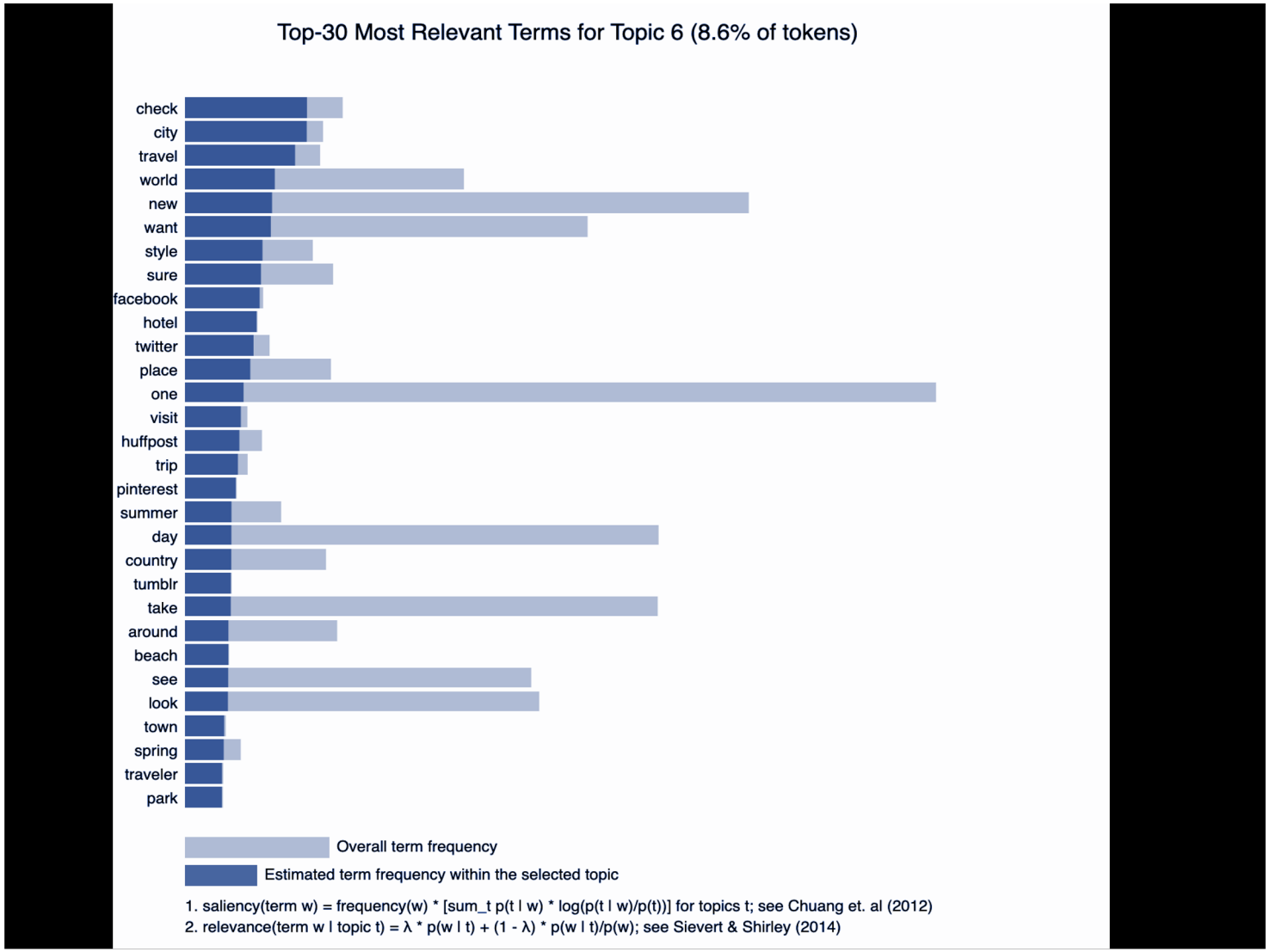
$$Coherence\ Score = \left(\frac{1}{M * (M-1)}\right) * \Sigma\left(sim(w_i, w_j) - sim(w_i)\right)$$

- M represents the total number of words in the topic.
- sim(wi, wj) is the similarity score between word wi and wj based on their co-occurrence in a given context.
- sim(wi) represents the average similarity score of word wi with all other words in the topic.

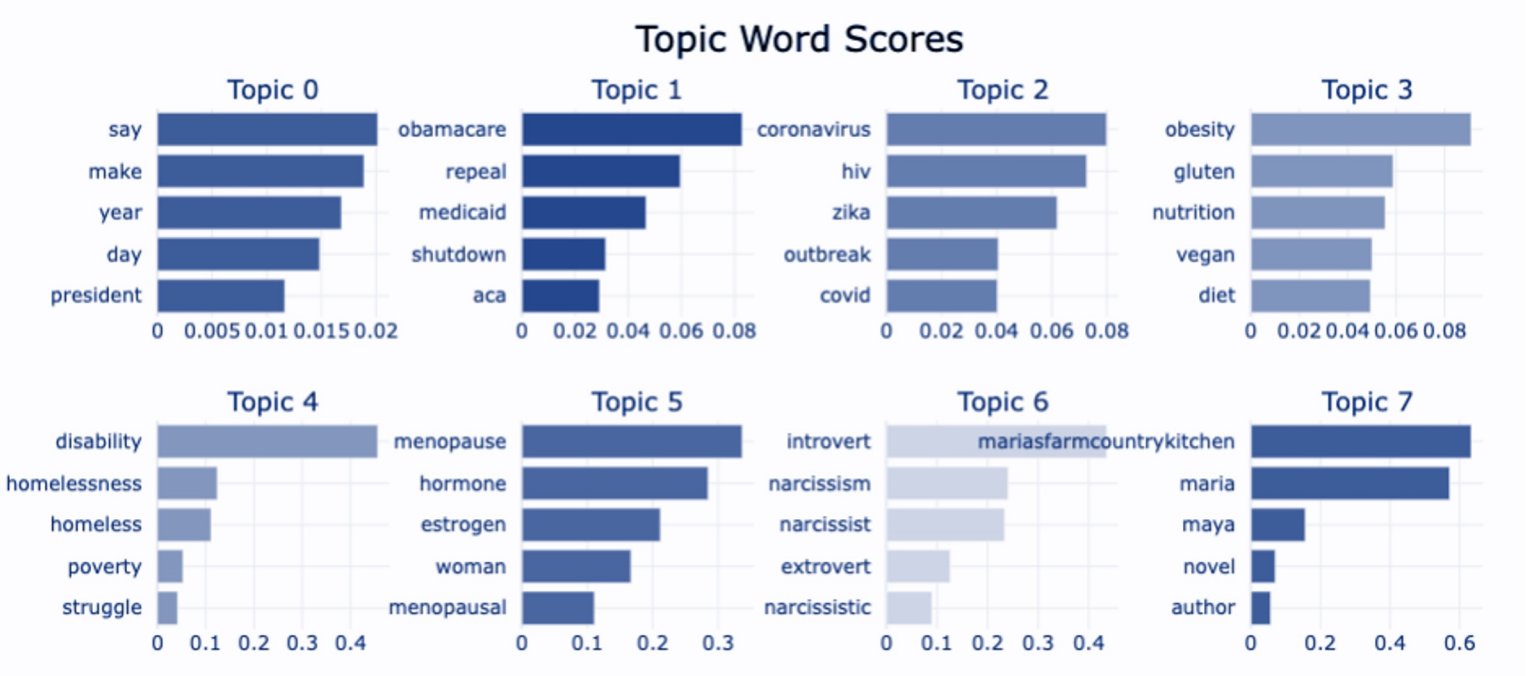
4

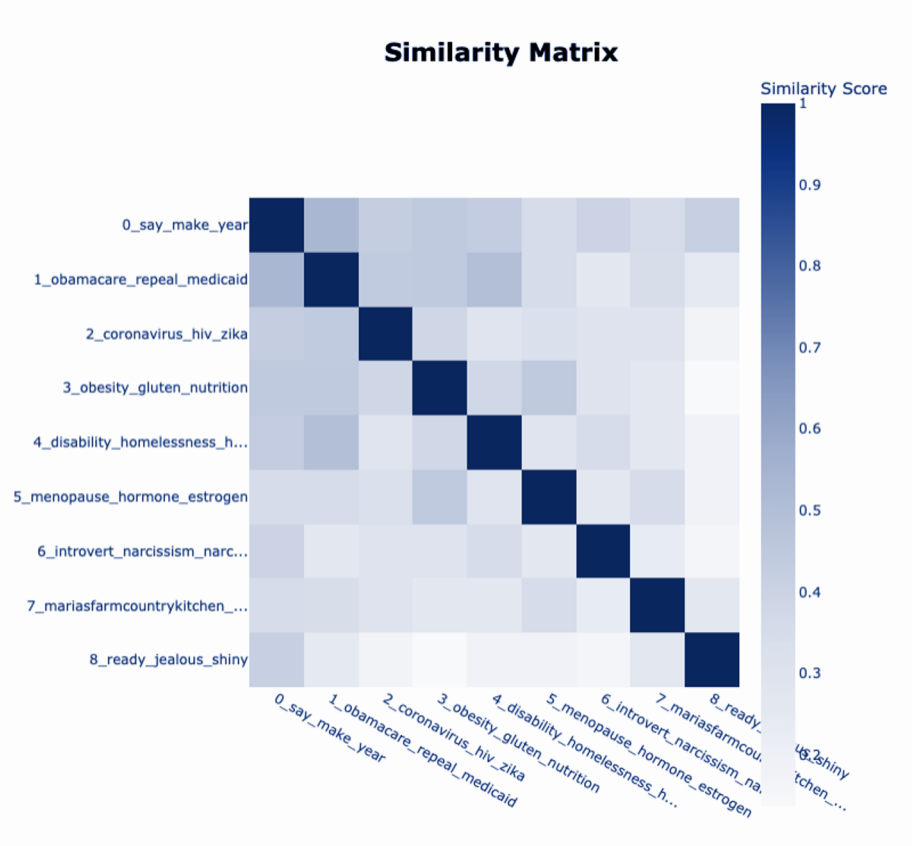
RESULTS & FINDINGS

Results - LDA

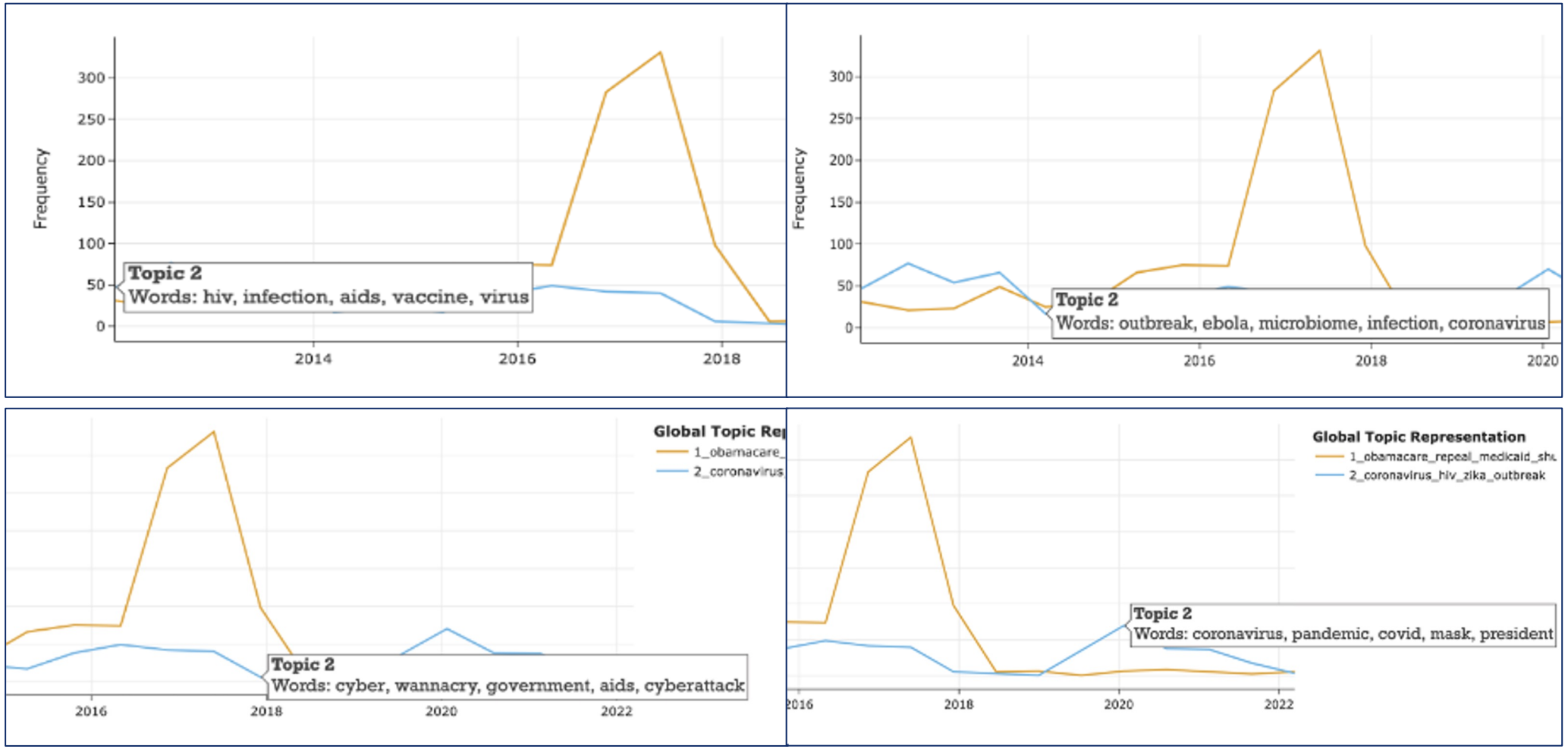

Top-30 Most Relevant Terms for Topic 6 (8.6% of tokens)
check
city
travel
world
new
want
style
sure
facebook
hotel
twitter
place
one
visit
huffpost
trip
pinterest
summer
day
country
tumblr
take
around
beach
see
look
town
spring
traveler
park
Overall term frequency
Estimated term frequency within the 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)

Results - BERTopic



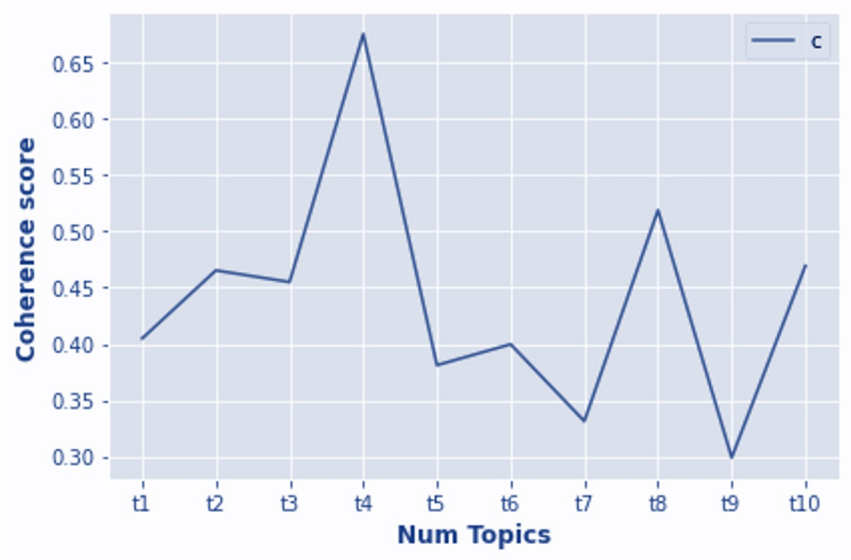


Results - BERTopic (continued): evolution of topics / dynamic modelling



Results - Performance: coherence

Coherence scores of LDA topics



Coherence Score of LDA model: ~0.44

BERTopic

Best Coherence score of BERTopic model

~0.70

***Coherence Score of the final model: ~0.64**

Results - Topics comparison

	LDA model	BERTopic
1	Topic 9: state, say, law, health, public, would, american, united, government, **care**	Topic 1: **obamacare**, repeal, medicaid, shutdown, aca, republicans, resign, **medicare**, senate, reform
2	Topic 2: year, new, study, old, **cancer**, month, find, researcher, percent, report	Topic: 2: coronavirus, hiv, zika, outbreak, covid, vaccination, ebola, pandemic, infect, meningitis
3	Topic 6: **look**, make, **food**, fashion, wear, dress, like, see, photo, good	Topic 3: **obesity**, gluten, nutrition, vegan, **diet**, diabetes, obese, healthy, celiac, gmo
4	Topic 7: dollar, company, million, business, **pay**, industry, money, market, **cost**, price	Topic 4: disability, homelessness, homeless, **poverty**, struggle, advocate, elderly, accessibility, shelter, single
5	Topic 4: man, film, star, woman, movie, game, video, play, name, one	Topic 5: menopause, hormone, estrogen, woman, menopausal, hysterectomy, postmenopausal, testosterone, ovarian, exercise

Conclusions: Usability of models for business

- BERTopic model offers a data-driven and automated approach to identifying consumer behavior.
- It can handle large datasets, identify latent topics, and uncover emerging trends.
- Benefits compared to traditional methods include agility, scalability, real-time understanding, and reduced bias.
- BERTopic model can be applied in market research, product development, marketing, reputation management, and customer support.

Main efficiency improve:

improved speed of trend identification
&
more precise understanding of consumer needs

Thanks for attention