**National Research University Higher School of Economics**

**Faculty of Computer Science**

**HSE and University of London Double Degree Programme in Data Science and Business Analytics**

**BACHELOR'S THESIS**

**(Research Project)**

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

**Prepared by the student of Group 191, Year 4 (year of study),**

**Dzhkha Anika**

**Thesis Supervisor:**

**Candidate of Science (PhD), Assistant Professor at ICEF,**

**Dimova Elena**

**Moscow**

**2023**

**Table of contents**

[1 Abstract 4](#_Toc136320977)

[2 Introduction 5](#_Toc136320978)

[2.1 Background of the study 5](#_Toc136320979)

[2.2 Problem Statement 6](#_Toc136320980)

[2.3 Relevance 6](#_Toc136320981)

[2.4 Research questions 6](#_Toc136320982)

[2.5 Objectives and significance 7](#_Toc136320983)

[2.1.1 Objectives 7](#_Toc136320984)

[2.1.2 Significance 7](#_Toc136320985)

[3 Literature review and previous studies 9](#_Toc136320986)

[3.1 Understanding consumer behaviour and marketing strategies 9](#_Toc136320987)

[3.2 Predicting consumer behaviour using machine learning 10](#_Toc136320988)

[4 Methodology 13](#_Toc136320989)

[4.1 Data preparation and pre-processing 13](#_Toc136320990)

[4.2 Modelling 18](#_Toc136320991)

[4.3 Topic models overview 18](#_Toc136320992)

[4.4 Latent Dirichlet Allocation model 20](#_Toc136320993)

[4.5 BERTopic model 24](#_Toc136320994)

[4.6 Model effectiveness evaluation 26](#_Toc136320995)

[5 Results 28](#_Toc136320996)

[6 Discussion and analysis 30](#_Toc136320997)

[7 Business applications 34](#_Toc136320998)

[7.1 Insights from the identified topics 34](#_Toc136320999)

[7.2 Usability of models for business 37](#_Toc136321000)

[8 Limitations and ethical considerations 41](#_Toc136321001)

[9 Future work 43](#_Toc136321002)

[10 Conclusion 45](#_Toc136321003)

[11 References 46](#_Toc136321004)

[12 Appendices 47](#_Toc136321005)

# Abstract

This research aims to explore the application of machine learning models, specifically BERTopic and Latent Dirichlet Allocation (LDA), for the analysis and prediction of socio-behavioural trends in the context of business operations. The goal is to investigate the potential of these models in identifying and analysing consumer behaviour patterns, enabling businesses to enhance marketing strategies and decision-making processes. The findings highlight the effectiveness of BERTopic and LDA in uncovering hidden trends, understanding customer preferences and sentiments, and optimizing marketing campaigns. The implications suggest that adopting these machine learning models can lead to increased sales, improved customer satisfaction, enhanced operational efficiency, and cost savings for businesses. The research also addresses the limitations and ethical considerations associated with using machine learning in consumer behaviour analysis. By examining the application of machine learning models, including BERTopic and LDA, in the analysis and prediction of socio-behavioural trends, this research contributes to the growing body of knowledge in the field.

Key words: Machine learning models, Socio-behavioural trend, Consumer behaviour, Business applications, Predictive analysis, Trend identification, Latent Dirichlet Allocation, BERT Topic Modelling

# Introduction

## Background of the study

This research focuses on utilizing machine learning models to analyse and forecast socio-behavioural patterns for commercial applications. Socio-behavioural trends refer to the dominant attitudes and behavioural patterns demonstrated by a given society, which are shaped by various factors, including cultural transformations, economic fluctuations, and technological progress. Acknowledging the significance of comprehending these patterns, enterprises are progressively resorting to machine learning algorithms to acquire discernments and attain a competitive edge in the marketplace.

Despite the notable progress made in the field of machine learning, there remain various obstacles and areas of limited comprehension. To successfully apply machine learning models to socio-behavioural trends, it is necessary to address various factors that can impact the outcomes, including but not limited to shifting social norms and economic conditions. Furthermore, there is a need to delve deeper into the interpretation and practical application of the outcomes derived from these models.

The objective of this study is to examine the application of machine learning models, namely BERTopic and LDA (Latent Dirichlet Allocation), in the analysis and forecasting of socio-behavioural patterns for commercial ends. This research aims to augment comprehension regarding the effective application of these models for the purpose of evaluating and predicting trends. The ultimate goal is to furnish businesses with actionable insights that enable them to make informed decisions. The objective of this study is to enable businesses to improve their market performance by utilizing machine learning models in conjunction with socio-behavioural trends.

## Problem Statement

The study aims to address is how to effectively apply machine learning models to the analysis and prediction of complex and variable socio-behavioural trends for business, with a focus on identifying and predicting trends in consumer behaviour, social norms, and communication and interaction patterns.

## Relevance

Numerous studies have investigated the potential advantages of different machine learning methods in the realm of consumer behaviour analysis. However, there exists a requirement for a more all-encompassing and pragmatic methodology that can be implemented in various sectors and situations. Moreover, a more profound comprehension is required regarding how enterprises can utilize socio-behavioural patterns to guide their tactics and functions, and to maintain their competitiveness in a swiftly evolving marketplace. The study's emphasis on the amalgamation of managerial knowledge and machine learning techniques underscores the significance of incorporating domain expertise and data-driven methodologies.

The relevance of this research lies in its capacity to provide enterprises with the necessary resources and knowledge to adjust, create, and react efficiently to socio-behavioural patterns, resulting in a competitive advantage and long-term prosperity in the industry.

## Research questions

We can identify the following research questions for the current study:

1. What are the most effective machine learning algorithms for analysing socio-behavioural trends from existing data?
2. How can the accuracy and reliability of machine learning models be improved?
3. Can we accurately predict socio-behavioural trends using machine learning?
4. How can businesses effectively use socio-behavioural trend analysis to inform their marketing strategies, product development, and overall business operations?
5. What limitations and ethical considerations should be taken into account when using a model for businesses?

## Objectives and significance

### Objectives

1. To explore the use of machine learning models for the analysis and prediction of socio-behavioural trends.
2. To identify the socio-behavioural trends and evaluate the potential implications for businesses.
3. To develop a methodology for applying machine learning models to socio-behavioural trend analysis and prediction that is reliable, accurate, and efficient.
4. To assess the effectiveness of using socio-behavioural trend analysis for informing business strategies and operations.

### Significance

This study has the potential to advance the field of socio-behavioural trend analysis by introducing novel methods and instruments for analysing and predicting these trends. By analysing key socio-behavioural patterns and their relevance to business operations, this research enables companies to obtain valuable insights and make informed decisions. This study's findings can assist companies in formulating strategies that align with emergent trends, allowing them to respond to shifting consumer preferences and market dynamics. In turn, this can improve their competitiveness, market position, and overall success.

In addition, this study acknowledges the ethical considerations and potential hazards associated with the use of machine learning models for analysing socio-behavioural trends. By casting light on these aspects, businesses and policymakers will be able to develop guidelines and best practices to ensure the responsible and ethical application of these models. This knowledge can assist organizations in navigating the challenges of data privacy, transparency, and impartiality, thereby fostering confidence and accountability in the use of machine learning algorithms for trend analysis.

# Literature review and previous studies

## Understanding consumer behaviour and marketing strategies

Within the domain of business, consumer behaviour and marketing strategies are intertwined. This literature review examines several books that cast light on the comprehension of consumer behaviour and the implementation of effective marketing strategies to reach customers. The chosen works provide insights into the evolving landscape of consumer behaviour, emergent trends, and the role of advanced technologies such as machine learning and data analytics in analysing and predicting consumer behaviour.

The author of "Marketing in the Moment" (Tasner, 2011) examines the shift in consumer behaviour toward an "always-on" mentality and its implications for businesses. Tasner emphasizes the significance of adapting marketing strategies to reach consumers in real-time via Web 3.0 technologies, such as machine learning and data analytics. The book focuses on the use of machine learning algorithms to analyse large datasets of consumer behaviour, identify patterns, and adjust marketing strategies accordingly.

Charles Duhigg's "The Power of Habit" (2012) examines the fundamental influence of habits on human behaviour, both personally and professionally. Duhigg stresses the significance of identifying and shaping critical behaviours in individuals and organizations. These insights enable the development of targeted marketing campaigns and employee training programs that are consistent with consumer behaviours, thereby enhancing customer engagement, employee productivity, and organizational effectiveness.

"Contagious" (2013) by Jonah Berger explores the scientific rationale for why certain products, ideas, and behaviours become popular. This book focuses predominantly on the scientific explanations for why certain products, ideas, and behaviours become popular, as well as strategies for businesses to leverage social influence and word-of-mouth marketing. The book emphasizes the role of social influence, emotion, and narrative in generating viral content and identifies the "STEPPS" framework (social currency, triggers, emotion, public, practical value, and stories) to increase the likelihood of virality and generate consumer interest.

The 2011 book "Thinking, Fast and Slow" by Daniel Kahneman examines the two systems of thought and their influence on decision-making in personal and professional contexts. Kahneman contends that prejudices and heuristics frequently affect decision-making, resulting in suboptimal outcomes. The author uses numerous cognitive psychology experiments and findings to illustrate the two systems of thought and their impact on decision-making.

Solomon and Armstrong's "Consumer Behavior" (2019) provides an in-depth examination of the psychological factors that have a significant impact on consumer behaviour. This book emphasizes the significance of understanding these factors to improve marketing effectiveness. It explores the intricate interplay between perception, cognition, motivation, and emotion, which collectively influence the decision-making processes of consumers. The authors present actionable strategies for utilizing this knowledge to create effective marketing campaigns. In addition, the book acknowledges the influence of emerging trends such as social media and mobile technologies, providing insight into how these trends are influencing the future of marketing.

## Predicting consumer behaviour using machine learning

The second part of the literature review provides an overview of machine learning methods for analysing and predicting consumer behaviour. The studies explore the potential of machine learning methods and big data analytics in understanding consumer behaviour, predicting preferences, identifying significant predictors, and enhancing marketing strategies. The findings highlight the importance of feature selection, dataset balancing, theory-guided analysis, and domain knowledge in achieving accurate predictions and improved business performance.

Agarwal et al. (2010) conducted a study to investigate the predictive potential of web search volume for consumer behaviour in various domains. By employing boosting trees, the authors developed baseline and combined models using publicly available and search data. They found that the predictive power of search data varies across domains, influenced by factors such as population size, search proximity to outcomes, and query relevance. The study concluded that search data can enhance the performance of baseline models, but its utility depends on speed, convenience, and adaptability.

Moura et al. (2021) aimed to improve the performance of a wine quality prediction model by utilizing machine learning techniques. They compared the performance of artificial neural network (ANN), support vector machine (SVM), and naïve Bayes (NB) algorithms using red and white wine datasets. The ANN algorithm achieved the highest accuracy, outperforming the other classifiers. The study emphasized the significance of feature selection and dataset balancing in enhancing model performance.

Yen et al. (2021) compared the performance of six supervised classification techniques to predict customer churn in the banking industry. Using demographic and personal attributes, including gender, geography, and product usage, the study employed algorithms such as Naïve Bayes, k-nn, SVM, decision trees, random forest, and ANN. The ANN and random forest algorithms demonstrated superior performance, with gender, geography, active membership, and product usage identified as significant predictors of churn.

Bradlow, Gangwar, and others (2017) examined the role of big data and predictive analytics in retailing, focusing on customer, product, time, location, and channel dimensions. The study emphasized the importance of theory and domain knowledge in analysing big data. Ethical and privacy issues related to big data usage were also discussed. The authors provided an example of a pricing field experiment that utilized A/B testing and predictive econometric modelling, resulting in enhanced retailer profitability. They concluded that the use of theory, domain knowledge, and statistical tools is crucial in leveraging big data for retailing.

The literature reviewed in this study provides valuable insights into the potential of combining managerial knowledge and machine learning to achieve intelligent data compression and its subsequent utilization in business operations. Nonetheless, there is a dearth of empirical studies on socio-behavioural patterns. This research gap indicates the need for further investigation in understanding how machine learning techniques can be effectively utilized to analyse and interpret socio-behavioural data, thereby enhancing decision-making processes in various business contexts. Future research efforts should aim to bridge this gap and explore the potential of machine learning in capturing and leveraging socio-behavioural patterns to inform business strategies and improve organizational performance.

# Methodology

## Data preparation and pre-processing

In our research, we are using the following dataset from Kaggle: The News Category Dataset. The dataset contains around 210k news headlines from 2012 to 2022 and is one of the largest news datasets available, serving as a benchmark for various computational linguistic tasks. Each record in the dataset includes information such as the article's category, headline, authors, link, short description, and publication date.

The dataset is extensive, containing a large number of news headlines spanning a significant period of time. This allows us to analyse trends and patterns over an extended timeframe and provides a comprehensive representation of news articles. Moreover, the dataset covers various categories, including politics, sports, entertainment, technology, and more. This diversity enables us to explore a wide range of topics and examine how different categories are represented within the dataset. The additional metadata as dates enhances the richness of the research and is also important to show better results.

The dataset includes a total of 42 news categories, with the top 10 categories being: POLITICS, WELLNESS, ENTERTAINMENT, TRAVEL, STYLE & BEAUTY, PARENTING, HEALTHY LIVING, FOOD & DRINK, BUSINESS, COMEDY.

Sample of the data:



Table 4.1 Dataset sample

Let’s look at the top 10 categories of news articles by the overall number of their appearance in the dataset:

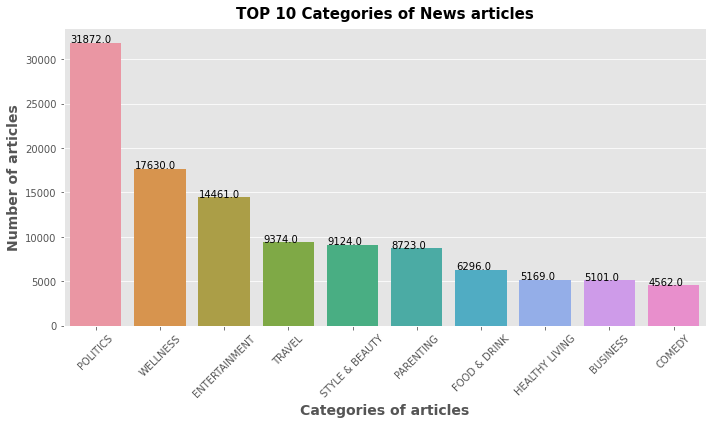


Fig. 4.1 Top 10 categories of news in the dataset

Key findings of the dataset exploration:

1. Dataset has total 41 distinct categories of news articles.
2. ‘Politics’ is the most common category of news in our dataset. ’Wellness’ and ’Entertainment’ are taking 2nd and 3d places.
3. We have total of 180740 different short descriptions of the news.
4. Maximum length of short descriptions is 1472 while median length is 121.
5. Dates of news starts from 2012-01-28 and goes up to 2022-09-23. The total amount of news from 2019 to 2022 is significantly lower than in other years.

Here we can take a look at top 20 categories:

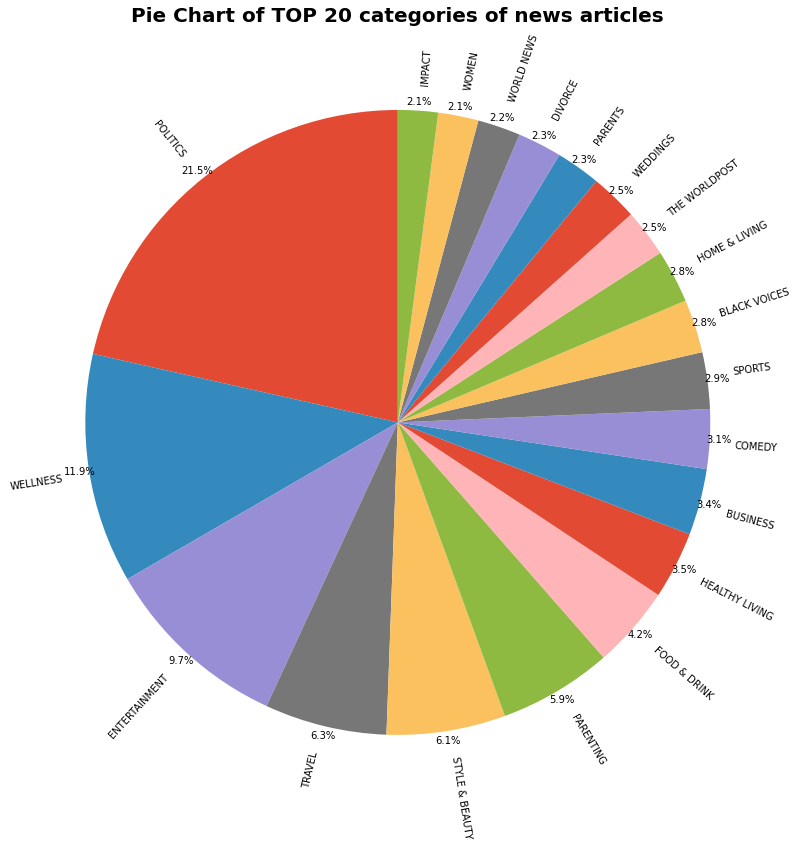


Fig. 4.2 Top 20 categories of news.

It is noticeable that categories ‘Weddings’, ‘Parents’, and ‘Divorce’ are having roughly the same amount of appearance, as well as ‘Women’ and ‘Impact’ topics in the dataset.

The overview of most frequent words is also helpful for our research as it coincides with society’s mindset. The following ‘wordclouds’ can show us the importance of some words in each category. It is created by arranging the words in a cloud-like shape where the size of each word/phrase corresponds to its frequency or importance. Typically, the more frequent a word appears in the text, the larger and bolder it appears in the word cloud.



Fig. 4.3 Wordclouds of popular categories

These word clouds offer a qualitative perspective alongside quantitative analysis. They serve as a quick visual summary of the main themes, making it easier to grasp the key categories and see the catchy words. What is interesting is that the appearance of ‘time’ and similar words in almost all topics signal that time plays a crucial role in many aspects of everyday life.

We can see the visualisation of the mean amount of news for the 10 most popular categories:

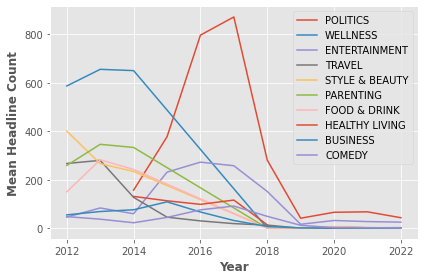


Fig. 4.4 Mean headlines count of top 10 categories over the years

The graph shows that the number of headlines in each category varied over the years, with POLITICS having the fast grow starting from 2014 and the highest mean amount of news since 2016 and WELLNESS having the highest number of headlines in 2012-2015 with the continuous fall after 2014. Significant events such as the U.S. presidential election, policy changes, political controversies, and international relations can generate extensive media coverage and drive up the number of political headlines. The shift in the highest number of headlines from wellness to politics can be influenced by media organisations' editorial decisions and public interest. Additionally, the average number of headlines in each category decreased by 2019 compared to the previous years.

## Modelling

To define socio-behavioural trends in the news data, we employ natural language processing (NLP) techniques to extract and analyse the short descriptions of news articles. Through topic modelling, we aim to identify the primary topics discussed in the short descriptions and examine how these topics evolve over time.

The topics and associated words obtained through topic modelling serve as a foundation for showcasing the business applications of our model. By analysing the extracted topics and their corresponding words, we can identify the key themes and interests prevalent in the news dataset. This understanding enables us to derive actionable insights for various business sectors.

## Topic models overview

Topic modelling is a powerful text analysis technique that allows us to discover latent themes or topics within a collection of documents without prior knowledge or manual annotation. It is particularly useful when dealing with large volumes of unstructured text data, such as news articles, social media posts, research papers, or customer reviews. The goal of topic modelling is to automatically identify and extract the main topics or thematic patterns present in the text corpus. It provides a way to uncover the underlying structure and content distribution within the documents, enabling us to gain insights into the key themes and subjects discussed.

The process of topic modelling involves several steps. First, the text data is pre-processed, which typically includes tasks such as tokenization, removing stop words, stemming or lemmatization, and handling of special characters or numbers. Next, a topic modelling algorithm is applied to the pre-processed text data to identify the topics. Once the topics are identified, each topic is represented by a distribution of words or keywords that are most strongly associated with that topic. These keyword distributions help in interpreting and understanding the meaning and content of each topic.

There are various topic modelling algorithms available, each with its own strengths and characteristics. Some popular algorithms include Latent Dirichlet Allocation (LDA), BERTopic, Non-Negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA). These algorithms employ different mathematical and statistical techniques to uncover topics based on word co-occurrence patterns, word distributions, or semantic similarities.

In this paper, the BERTopic algorithm was chosen as the primary method, while the Latent Dirichlet Allocation algorithm was utilised as the baseline method. The selection of BERTopic was driven by its aptitude to apprehend semantic significance and contextual details in textual data by means of pre-trained BERT embeddings. BERT, also known as Bidirectional Encoder Representations from Transformers, is a cutting-edge deep learning model that has garnered significant attention for its exceptional performance in NLP tasks. Through the utilisation of BERT embeddings, BERTopic endeavours to furnish precise and significant topic representations. The selection of BERTopic was also motivated by its efficacy in managing extensive and heterogeneous datasets.

The selection of BERTopic and LDA was based on a thorough analysis of literature on topic modelling algorithms and their implementation in various fields. According to (R. Egger and J. Yucorresponding, 2022) BERTopic demonstrates better results compared to LDA, making it a more promising choice for analysing social media data. The study acknowledges the effectiveness of BERTopic in capturing contextual information, which may be concealed by other models considered. Additionally, the authors suggest considering NMF as an alternative to LDA for traditional topic modelling in social science research. It is important to note that each model has its strengths and weaknesses, and the findings require careful qualitative interpretation.

The rationale behind opting for LDA as the baseline approach stems from its ubiquitous adoption as a traditional probabilistic methodology for topic modelling. The LDA algorithm assumes that documents are generated from a combination of topics and has exhibited favourable outcomes in diverse use cases. Through a comparative analysis of BERTopic and LDA, the study sought to evaluate the potential superiority of BERT-based methodologies over conventional topic modelling techniques.

This study elucidated the capabilities and constraints of diverse methodologies, encompassing BERT-derived models and LDA. Furthermore, it highlighted the necessity for sophisticated algorithms that can effectively capture semantic significance, contextual details, and proficiently manage vast and varied datasets.

By integrating BERTopic and LDA into the investigation, our study endeavours to augment the current corpus of knowledge by assessing the efficacy of BERTopic in contrast to a firmly established benchmark technique. The comparative analysis results will yield valuable insights into the efficacy of BERT-based models for topic modelling tasks and their potential superiority over conventional methods such as LDA.

## Latent Dirichlet Allocation model

Latent Dirichlet Allocation is a probabilistic generative model widely used for topic modelling; a task aimed at uncovering latent thematic structures in a collection of documents. LDA assumes that each document in the corpus is a mixture of multiple topics, and each topic is a probability distribution over a fixed vocabulary of words.

The underlying principle of LDA is to identify the latent topics by estimating the distribution of topics in each document and the distribution of words in each topic. It follows a Bayesian framework where the goal is to infer the posterior distribution of the latent variables (topic assignments for words and document-topic distributions) given the observed words in the documents.

The LDA algorithm works as follows:

1. Initialization: Randomly assign topics to each word in each document and initialise the topic-word and document-topic distributions.
2. Iterative Inference: Repeatedly update the topic assignments for each word and refine the topic-word and document-topic distributions based on the observed words and the current estimates.
3. Convergence: Iterate the inference process until a convergence criterion is met, such as a maximum number of iterations or a small change in the estimated distributions.
4. Output: The final output of the LDA algorithm is the estimated topic-word and document-topic distributions, which represent the discovered topics and their associated word probabilities and document-topic proportions.

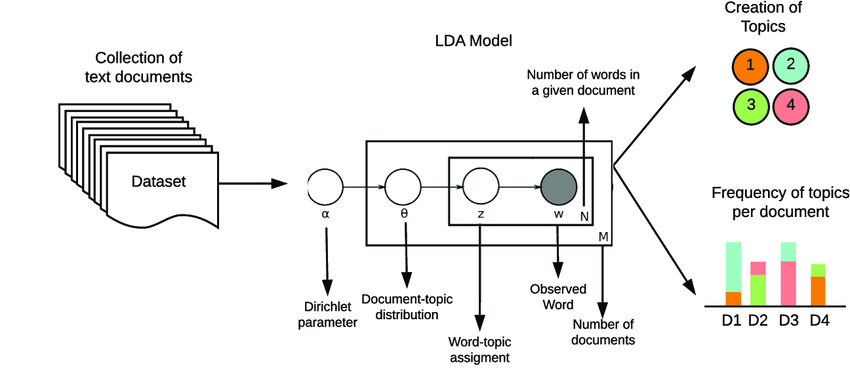


Fig. 4.5 LDA topic model

The LDA model has strengths that make it effective in topic modelling. One key strength is its ability to generate interpretable topics. By estimating word distributions in each topic, LDA reveals thematic structures, enabling meaningful analysis and understanding. LDA is also flexible, handling diverse types of textual data such as documents, social media posts, or reviews. It can handle multilingual datasets, making it applicable across different languages. Additionally, LDA is scalable, efficiently processing large-scale datasets through parallel computing techniques, ensuring computational efficiency and faster analysis.

However, it is important to acknowledge the drawbacks of the LDA model. One limitation is the assumption that each document is a mixture of a fixed number of topics. This assumption may not hold true in all cases, as real-world documents can often cover multiple topics simultaneously or exhibit varying degrees of topic dominance. Additionally, LDA requires researchers to predefine the number of topics, which can be a challenging task, especially when working with unfamiliar or ambiguous datasets. Another drawback of LDA is its reliance on bag-of-words representation, which disregards the word order and context within a document. This limitation can lead to the loss of important semantic information and potentially affect the quality of the topic modelling results.

The model used in the code is LdaMulticore from the Gensim library, which is an implementation of Latent Dirichlet Allocation (LDA) for topic modelling.

lda\_model = LdaMulticore(corpus=corpus,

id2word=id2word,

num\_topics=NUM\_TOPICS,

random\_state=42,

passes=10,

workers=4,

alpha = 0.01,

eta=0.99)

After tuning the hyperparameters, the model is configured as follows:

* corpus: The corpus of documents used for training the model.
* id2word: The mapping of word IDs to their corresponding words in the vocabulary.
* num\_topics: The desired number of topics to be extracted from the corpus.
* random\_state: A fixed random seed for reproducible results.
* passes: The number of passes through the corpus during training.
* workers: The number of parallel workers to use for training.
* alpha: The parameter controlling the sparsity of the per-document topic distributions. A lower value (0.01) results in more focused topics per document.
* eta: The parameter controlling the sparsity of the per-topic word distributions. A higher value (0.99) encourages more diverse topic assignments for each word.

## BERTopic model

BERTopic is a topic modelling algorithm that leverages BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art deep learning model, to generate topic representations. Unlike LDA, BERTopic is an embedding-based model that captures contextual information and semantic relationships between words, resulting in more accurate and meaningful topic representations.

The BERTopic algorithm works as follows:

1. Embedding Generation: BERTopic first converts the input text into numerical vectors using a pre-trained BERT model. This step captures the contextual information and word semantics.
2. Document Clustering: BERTopic applies an efficient clustering algorithm, such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), to group similar documents into clusters. Each cluster represents a topic.
3. Topic Representation: BERTopic calculates the centroid (mean vector) for each cluster, representing the topic. These centroid vectors capture the most representative words and their embeddings within the topic.
4. Topic Ranking: BERTopic ranks the topics based on their prevalence across the corpus. This ranking considers the number of documents assigned to each topic and their similarity to the centroid vectors.
5. Topic Interpretation: Finally, BERTopic provides keywords that are most relevant to each topic, allowing for better interpretation and understanding of the discovered topics.

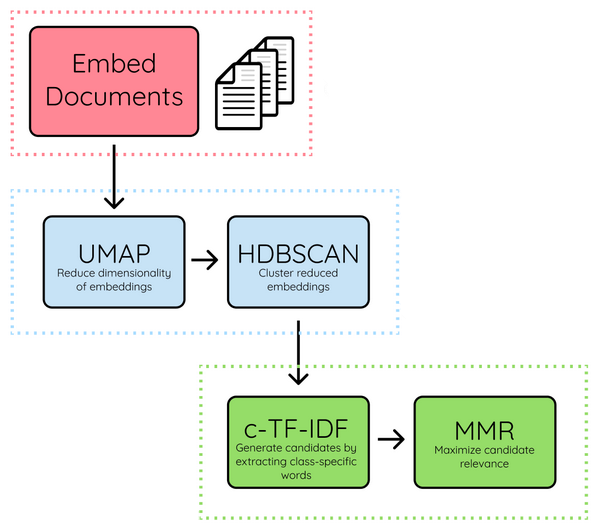


Fig. 4.6 BERTopic model structure

One major strength of BERTopic is its ability to encode contextual information, capturing the nuances and complexities of language. This enables the model to generate more accurate and contextually rich topics compared to traditional topic modelling algorithms like LDA. BERTopic also has the advantage of not requiring the pre-specification of the number of topics, as it automatically discovers topics based on the data. Moreover, BERTopic tends to produce topics with higher coherence scores, indicating that the words within each topic are more semantically related and form more coherent clusters. This contributes to the interpretability of the topics and enhances the understanding of the underlying thematic structures within the corpus.

However, it is important to note that BERTopic has its limitations. The main drawback is its computational complexity and resource requirements due to the use of BERT, which is a large and computationally expensive model. Fine-tuning BERT for topic modelling is time-consuming and requires substantial computational resources. Additionally, BERTopic may not perform optimally on small datasets with limited training examples, as it heavily relies on the representations learned from pre-training on large-scale corpora.

## Model effectiveness evaluation

The coherence score is calculated to assess the interpretability of the generated topics. It measures the degree of semantic coherence between the words within a topic. The calculation involves estimating the probability of word co-occurrence within the context of a topic and comparing it to a background distribution. The coherence score is derived from the average of these probabilities across all word pairs within a topic.

(1)

where:

M is the number of unique word pairs within the topic, represents the coherence measure between word w\_i and word w\_j, is the average coherence measure of word wi with all other words in the topic

The coherence score is a valuable measure for assessing the interpretability of topic models. It helps quantify the semantic relationships between words within a topic, indicating how well the topics capture distinct themes or concepts. Higher coherence scores indicate that the words within a topic are more semantically related and provide a clearer representation of a specific theme. In contrast, lower coherence scores suggest that the words within a topic are less connected and may not convey a coherent concept.

In the context of LDA, the coherence score reflects the extent to which the words within a topic are related and provide a clear representation of a particular theme. A coherence score of *0.415* in the baseline LDA model indicates a moderate level of topic interpretability, suggesting that the topics generated have some degree of coherence but could be further improved. After tuning the hyperparameters, the coherence score increased to *0.44*, indicating a slight improvement in the interpretability of the topics. While this increase is positive, it still suggests room for further enhancement.

BERTopic demonstrates higher coherence scores up to ***0.64*** in our final model. This score is noticeably higher than the one we got in LDA. It indicates a stronger level of topic interpretability and coherence compared to LDA. BERTopic utilises contextual embeddings to capture the semantic relationships between words, enabling it to generate more coherent topics. The higher coherence score suggests that the topics produced by BERTopic are more representative of distinct themes and exhibit stronger semantic cohesion.

# Results

Output of the LDA model:

Topic 0: show, night, late, week, new, host, music, last, news, first

Topic 1: life, make, people, way, one, need, change, well, time, take

Topic 2: year, new, study, old, cancer, month, find, researcher, percent, report

Topic 3: trump, say, president, donald, house, campaign, republican, obama, white, former

Topic 4: man, film, star, woman, movie, game, video, play, name, one

Topic 5: get, time, like, one, day, know, thing, say, make, kid

Topic 6: look, make, food, fashion, wear, dress, like, see, photo, good

Topic 7: dollar, company, million, business, pay, industry, money, market, cost, price

Topic 8: check, city, travel, world, new, want, style, sure, facebook, hote

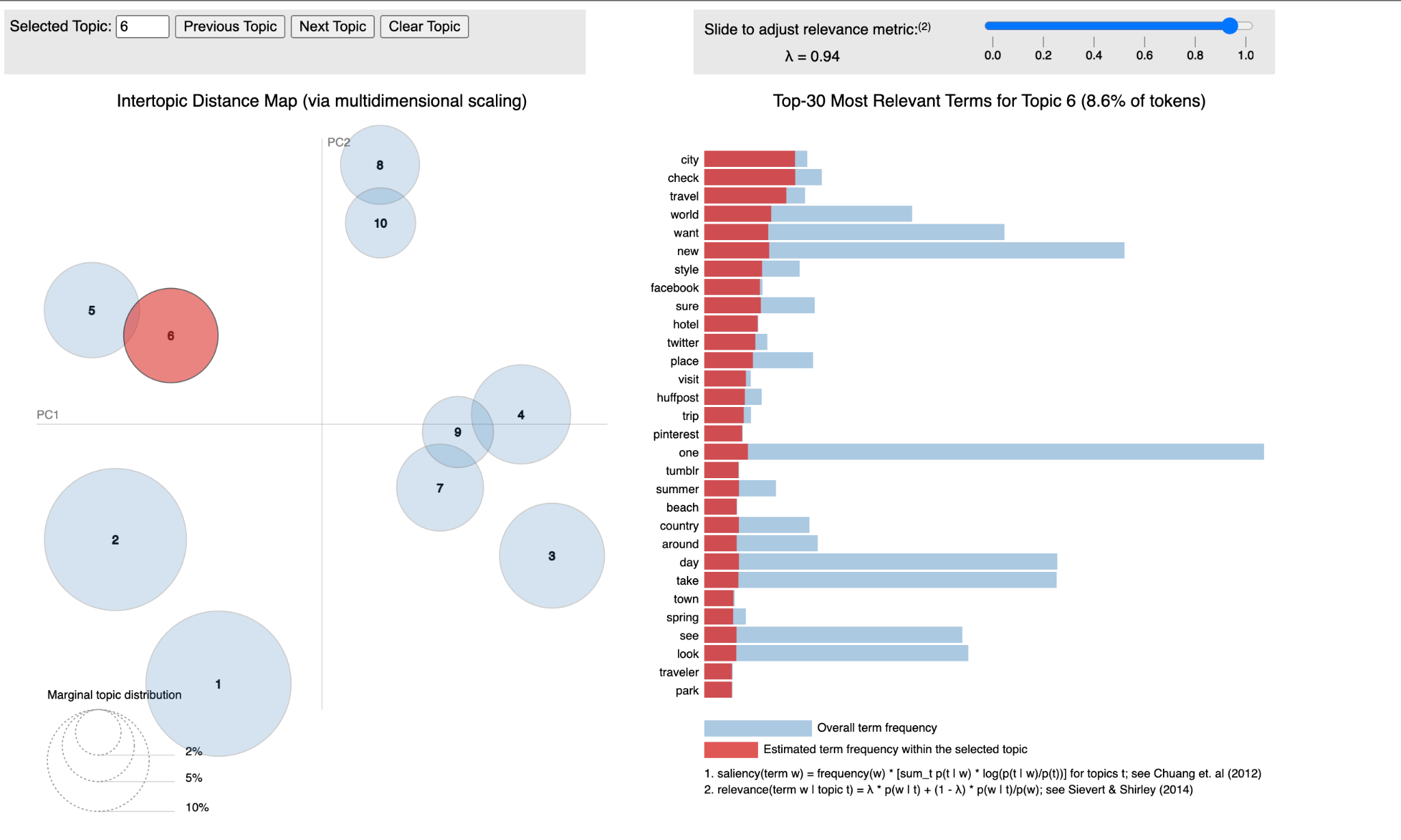
Topic 9: state, say, law, health, public, would, american, united, government, careThe visualisation of the LDA model results: 

Fig. 5.1 Posterior word distribution of LDA topic 6

Output of the BERTopic model looks like:

Topic: -1 Words: say, make, year, life, trump, day, president, world, live, family

Topic: 0 Words: say, make, year, day, president, life, thing, work, sleep, help

Topic: 1 Words: obamacare, repeal, medicaid, shutdown, aca, republicans, resign, medicare, senate, reform

Topic: 2 Words: coronavirus, hiv, zika, outbreak, covid, vaccination, ebola, pandemic, infect, meningitis

Topic: 3 Words: obesity, gluten, nutrition, vegan, diet, diabetes, obese, healthy, celiac, gmo

Topic: 4 Words: disability, homelessness, homeless, poverty, struggle, advocate, elderly, accessibility, shelter, single

Topic: 5 Words: menopause, hormone, estrogen, woman, menopausal, hysterectomy, postmenopausal, testosterone, ovarian, exercise

Topic: 6 Words: introvert, narcissism, narcissist, extrovert, narcissistic, introversion, impostor, family, extroversion, shyness

Topic: 7 Words: mariasfarmcountrykitchen, maria, maya, novel, author, adventure, milkweed, pagans, nonfiction, brindisi

Topic: 8 Words: ready, jealous, shiny, confirm, able, probably, line, close, set, let

At the visualisations of this method we going to look further.

# Discussion and analysis

The comparison between the results of BERTopic and LDA models reveals interesting insights into the topic modelling performance of both approaches.

BERTopic appears to generate more specific and diverse topics compared to LDA. The BERTopic topics cover a wide range of subjects, including politics, health, social issues, personal experiences, and more. Each topic consists of a distinct set of words that are contextually relevant to the given theme. This granularity allows for a more detailed understanding of the underlying topics in the dataset. It's worth noting that the BERTopic model assigns a label of "-1" to some topics, indicating that those topics may not have a clear and identifiable theme. This highlights the challenge of automatically assigning meaningful labels to all generated topics.

On the other hand, the topics generated by LDA tend to be broader and less specific. The LDA topics revolve around general themes such as entertainment, life, studies, politics, and business. While these topics provide a high-level overview of the content, they may lack the fine-grained details and context captured by BERTopic.

Let’s look at the one of the visualisations of BERTopic generated topics:

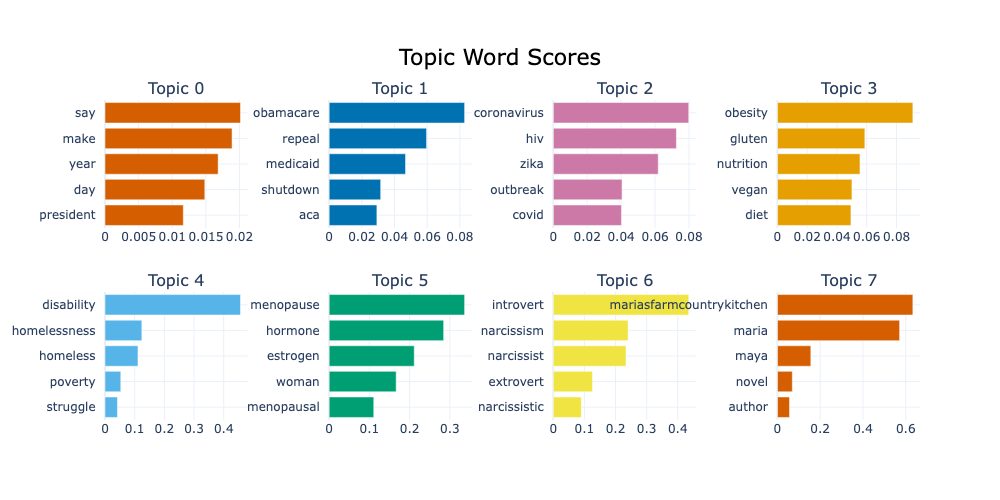


Fig. 6.1 BERTopic word scores

Word topic scores in BERTopic represent the relevance or importance of a specific word within a particular topic. These scores indicate the degree to which a word contributes to the overall definition or theme of a topic. The word topic scores are calculated based on the frequency and distribution of words across the entire corpus and within individual topics. A higher score for a word within a topic suggests that the word is more strongly associated with that topic. Conversely, a lower score indicates that the word has less relevance to the topic.

One more important result of the BERTopic model can be displayed in the following graph that provides an insightful representation of how certain topics evolve and change over time:

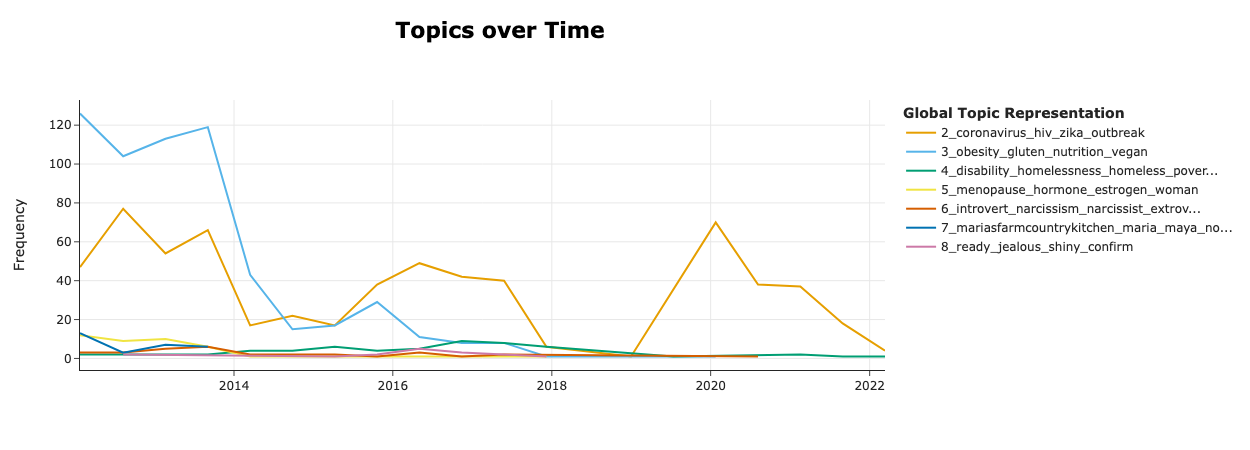


Fig. 6.2 BERTopic topics 2-8 over Time

By specifying a set of topics, in this case 2-8, the graph displays the temporal distribution of these topics, showcasing their prevalence and popularity during different time periods. For example, topic 2 that is related to global health issues consistently maintains a high level of prominence throughout the observed time period (except for 2018-2019, drop right before the COVID-19), it suggests that the corresponding theme or concept remains consistently relevant over time. On the other hand, topic 3 related to health & dietary choices experiences a surge in popularity in 2012-2024, but experiences a decline during all other time, which may indicate a significant event or trend that captured public attention during that period.

Moreover, if we look through the topic 2 evolution more closely as an example, we can see that it developed from words: hiv, infection, aids in 2012 to outbreak, ebola, microbiome in 2014.

|  |  |
| --- | --- |
|  |  |

Fig. 6.3 Evolution of topic 2 from 2012 to 2014

By 2018 in global issues, we define the most popular words like ‘cyber’, ‘wannacry’, ‘cyberattack’, and starting from 2019 words related to COVID-19 and pandemic.

|  |  |
| --- | --- |
|  |  |

Fig. 6.4 Evolution of topic 2 from 2018 to 2020

Using this visualisation result we can distinguish most popular words related to each topic historically or find ones that are most recent.

Let’s select the top 6 topics with the most meaningful results and define their headlines manually, which are:

* Topic: 1 Politics (Words: obamacare, repeal, medicaid, shutdown, aca, republicans, resign, medicare, senate, reform)
* Topic: 2 Global health issues (Words: coronavirus, hiv, zika, outbreak, covid, vaccination, ebola, pandemic, infect, meningitis)
* Topic: 3 Health and Dietary Choices (Words: obesity, gluten, nutrition, vegan, diet, diabetes, obese, healthy, celiac, gmo)
* Topic: 4 Socioeconomic Challenges and Advocacy (Words: disability, homelessness, homeless, poverty, struggle, advocate, elderly, accessibility, shelter, single)
* Topic: 5 Menopause and Women's Health (Words: menopause, hormone, estrogen, woman, menopausal, hysterectomy, postmenopausal, testosterone, ovarian, exercise)
* Topic: 6 Introversion, Extroversion, and Personality Traits (Words: introvert, narcissism, narcissist, extrovert, narcissistic, introversion, impostor, family, extroversion, shyness)

The six topics and their components will be used further to identify socio-behavioural trends because they reflect shifts in overall social behaviour, values, attitudes, and preferences within their respective domains. These trends indicate broader societal changes and highlight the evolving dynamics of human behaviour and decision-making.

# Business applications

## Insights from the identified topics

This section of the paper will examine the outcomes derived from the BERTopic model and investigate plausible business implementations for the recognised patterns and for the utilisation of the model. Through the analysis of patterns and tendencies, it is possible to acquire significant knowledge regarding the evolving inclinations and actions of customers. Such information can be utilized to make well-informed corporate judgments. Our proposal entails offering recommendations that can assist organizations in maintaining a competitive edge, predicting changes in consumer demand, and adjusting their tactics to keep up with the swiftly changing business environment of today.

The news media is an essential aspect in shaping social behaviour by acting as a primary conduit for information dissemination and influencing public opinion. The interplay between news and social behaviour is complex and involves diverse mechanisms that can influence individuals and the broader society.

Firstly, the identified topics from the BERTopic model can provide valuable insights into socio-behavioural trends and have several applications in various business domains. Let’s address some potential applications:

1. By analysing topics related to politics, such as Topic 1 ("Politics") in the example, businesses can gain insight into public sentiment and opinions regarding political issues. This data can be utilized to comprehend consumer behaviour, anticipate market trends, and make informed business decisions. By investigating political concerns such as healthcare reform, a news media organization can modify its coverage and advertising to appeal to audiences interested in political discussions and policy debates.
2. Topics such as Topic 2 ("Global health issues") concentrate on infectious diseases and vaccinations. Monitoring these topics can assist businesses in the healthcare industry in keeping abreast of emerging health concerns, adapting their strategies, and developing products and services to meet public health requirements. In order to resolve pressing public health concerns and contribute to disease prevention and control, a pharmaceutical company can allocate resources to the development of vaccines and therapies, as well as open lectures for global health issues such as HIV, Zika, and Ebola.
3. Targeting Health-Conscious Consumers: Topic 3 ("Health and Dietary Choices") emphasizes health-related important terms such as nutrition, diet, and diabetes. These insights can be utilized by businesses in the food and wellness industries to develop and market products that appeal to health-conscious consumers, promote healthy lifestyles, and meet specific dietary requirements. A brand of healthy snacks may launch a line of gluten-free, vegan, and nutritionally balanced products in response to the rising demand for healthier dietary options.
4. Topics such as Topic 4 ("Socioeconomic Challenges and Advocacy") cast light on social problems such as homelessness and poverty. Businesses can use these insights to align their corporate social responsibility initiatives, contribute to social change, and support relevant causes, thereby enhancing their brand image and reputation. The goal of a social advocacy organization is to improve the socioeconomic conditions of local communities by initiating campaigns and partnerships to address homelessness and poverty.
5. Topic 5 ("Menopause and Women's Health") is centred on women's health-related keywords. This information can be utilized by businesses in the healthcare and wellness industries to develop products, services, and educational materials that cater to women's health requirements, particularly those related to menopause and hormone-related issues. A women's health clinic can provide specialized services, information, and products to address menopause-related symptoms and hormone imbalances, catering to the particular requirements of menopausal women.
6. Understanding Characteristics of Personality and Consumer Preferences: Introversion, extroversion, and personality traits are discussed in Topic 6 ("Introversion, Extroversion, and Personality Traits"). This information can be useful for businesses in areas such as marketing and customer service, allowing them to modify their communication strategies and customer experiences to the various personality types of their customers. A marketing agency is able to create personalized marketing campaigns that resonate with both introverted and extroverted individuals, ensuring that communication styles and channels align with varying personality types.

As it can be seen the identification of socio-behavioural trends through the analysis of topics can greatly benefit businesses in understanding consumer behaviour and driving various business processes. By leveraging these particular insights, businesses can tailor their strategies, develop relevant products and services, engage with customers effectively, and contribute to social causes, ultimately leading to improved customer satisfaction, brand reputation, and business success.

## Usability of models for business

In the realm of identifying consumer behaviour, the usability of models for business has become increasingly important. A new tool that has emerged is the utilization of the BERTopic model, which offers a novel approach to identifying trends and building processes. In order to gain a comprehensive understanding of the implications of this novel approach, it is imperative to juxtapose it with conventional techniques that are frequently utilized for discerning consumer conduct.

Traditional techniques for ascertaining consumer behaviour generally entail the gathering of data via surveys, interviews, focus groups, and market research. The aforementioned techniques utilize organized surveys and predetermined classifications to acquire knowledge regarding consumer inclinations, buying habits, and cognitive processes involved in making choices. Although traditional methods have been extensively utilized and have yielded significant findings, they frequently encounter constraints such as partiality, probable response prejudices, and a protracted process of data gathering.

In contrast, the usage of the BERTopic model offers a more data-driven and automated approach to identifying consumer behaviour. The BERTopic model is based on natural language processing and machine learning techniques, specifically leveraging the Bidirectional Encoder Representations from Transformers (BERT) algorithm. This model can analyse large volumes of unstructured text data, such as social media posts, online reviews, and customer feedback, to extract topics and trends without the need for predefined categories.

One of the key advantages of the BERTopic model is its ability to identify latent topics and uncover emerging trends in an unsupervised manner. It can capture nuanced patterns and sentiments present in consumer-generated content, providing a more comprehensive and real-time understanding of consumer behaviour. Additionally, the BERTopic model can handle large datasets efficiently and can adapt to evolving trends and language usage over time.

Compared to traditional methods, the BERTopic model offers several benefits. It enables businesses to gain insights into consumer behaviour in a more agile and scalable manner, as it can process vast amounts of unstructured data quickly. This facilitates the identification of emerging trends and allows businesses to respond promptly and adapt their strategies accordingly. Moreover, the automated nature of the BERTopic model reduces human bias and potential errors that can arise in manual categorization processes.

* The BERTopic model can be applied in market research activities to extract insights from diverse textual sources, including customer feedback, social media conversations, and online reviews. By analysing these data, the model can identify prevalent topics, evaluate consumer sentiment, and uncover emerging trends. For example, a company in the hospitality industry could employ the BERTopic model to analyse online reviews and identify key topics such as room cleanliness, customer service, and amenities. This analysis enables businesses to gain a deeper understanding of customer preferences, satisfaction levels, and areas for improvement.
* Utilizing the BERTopic model, businesses can leverage unstructured text data to drive product development and enhancement efforts. By analysing customer feedback, product reviews, and user-generated content, the model can extract valuable insights regarding customer preferences, feature requests, and pain points. For instance, an e-commerce company could employ the BERTopic model to identify emerging topics related to customer desires for enhanced product functionality or improved user experience. This analysis enables businesses to prioritize and tailor product development initiatives to meet customer expectations.
* The BERTopic model can be instrumental in monitoring brand perception and managing online reputation. By analysing social media conversations, news articles, and customer reviews, the model can identify sentiment trends and detect potential reputation risks or opportunities. For example, a telecommunications company could employ the BERTopic model to monitor social media discussions and identify topics related to customer complaints or service issues. This analysis enables businesses to proactively address concerns, manage reputation crises, and enhance brand perception.
* Employing the BERTopic model in customer support processes can enable businesses to provide more personalized and efficient customer service. By analysing customer inquiries, support tickets, and chat logs, the model can identify common issues, extract relevant information, and suggest appropriate responses or solutions. For instance, a software company could use the BERTopic model to analyse customer support tickets and identify recurring topics or patterns. This analysis allows businesses to optimize support workflows, automate responses to common issues, and provide timely resolutions to customer inquiries.

The utilization of the BERTopic model in business processes has significant implications for various aspects of business performance, including increased sales, improved customer satisfaction, enhanced operational efficiency, and possible cost savings. By analysing unstructured text data, such as customer feedback and social media conversations, the BERTopic model allows businesses to acquire a deeper understanding of customer preferences, requirements, and emotions. This knowledge enables businesses to align their products, services, and marketing strategies with customer expectations, resulting in increased customer satisfaction and loyalty. This targeted approach enhances customer engagement, improves conversion rates, and drives sales. Furthermore, the model's ability to analyse large volumes of text data efficiently contributes to operational efficiency by automating the process of trend identification and customer behaviour analysis.

Let's consider a retail company with multiple physical stores. Traditionally, market research and consumer behaviour analysis required hiring external consultants or conducting expensive surveys to gather customer insights. However, by leveraging the BERTopic model, the company can analyse customer feedback from various sources, such as online reviews, social media posts, and customer service interactions, in a cost-effective manner.

# Limitations and ethical considerations

Several limitations of the research conducted on the application of the BERTopic model to consumer behaviour analysis must be acknowledged. These limitations impact the research's generalizability, data quality, interpretability, and contextual comprehension.

The limitations of the study are notable in regard to the sample size and generalizability. In the event that the dataset employed is restricted or exhibits partiality towards a specific group or platform, the conclusions drawn may not accurately reflect the entirety of the consumer demographic. In order to augment the credibility and generalizability of the results, it is imperative to ascertain a heterogeneous and inclusive cohort. In addition, constraints are imposed by the quality and dependability of the data used to train and evaluate the BERTopic model. Biases, inaccuracies, or disturbances in the data may affect the veracity and validity of the research's conclusions. The most recent data should be considered for more precise results. It is necessary to employ rigorous data pre-processing and validation techniques to ensure data integrity.

Another limitation of the BERTopic model is its interpretability. The BERTopic model, like many other machine learning algorithms, is a "black box" model, making it difficult to interpret the underlying decision-making process. While the model may generate accurate predictions, it can be challenging to comprehend the rationale behind those predictions. This restriction may restrict the model's insights and actionable recommendations.

The application of the BERTopic model to consumer behaviour analysis is limited by its limited contextual comprehension. The model largely relies on textual data and may not account for nonverbal cues, personal experiences, and cultural nuances that play a significant role in consumer behaviour. As a result, the model might not be able to provide an exhaustive comprehension of consumer behaviour.

Additionally, privacy and data protection are important ethical considerations. The utilization of the BERTopic model involves processing and analysing large amounts of consumer data, necessitating measures to ensure the secure handling of personal and sensitive information. Adhering to relevant privacy regulations and guidelines is essential to protect individuals' privacy and data rights.

In terms of ethical considerations, obtaining informed consent from individuals whose data is being analysed is of utmost importance. Clear communication about the purpose, scope, and potential implications of data analysis should be provided, allowing individuals to make informed decisions about their participation.

Lastly, unintended consequences should be considered. The utilization of the BERTopic model may have unintended effects, such as amplifying existing biases or impacting vulnerable populations. Businesses must remain vigilant and implement safeguards to prevent and mitigate any adverse consequences.

# Future work

Future research in the study of utilizing the BERTopic model for consumer behaviour analysis presents several promising avenues for further exploration and investigation. These future directions encompass understanding the life cycle of trends, quantifying the financial impact of trends on businesses, assessing the speed of capitalizing on identified trends, and incorporating theoretical frameworks such as the Earth Observation Hype Cycle.

One area of future work involves examining the life cycle of trends and its implications for business profitability. Understanding how trends evolve over time and whether they follow predictable patterns can provide valuable insights into their monetary value for businesses. Researchers can delve into the temporal dynamics of trends identified through the BERTopic model, analysing their emergence, growth, maturity, and decline. By discerning the life cycle of trends, businesses can make informed decisions on the timing and duration of their investments, optimizing their potential for generating revenue.

Another important aspect to investigate is the speed at which businesses can leverage identified trends and convert them into profitable opportunities. The BERTopic model's ability to detect and analyse trends in near real-time presents an exciting prospect for businesses to act swiftly and capitalize on emerging consumer behaviours. Future research can focus on understanding the timeframe from trend identification to effective implementation of corresponding marketing strategies. This analysis can shed light on the efficiency and agility of businesses in adapting to changing consumer preferences and taking advantage of market opportunities.

Incorporating theoretical frameworks, such as the Earth Observation Hype Cycle, can provide a valuable perspective on the adoption and commercialization of trends identified through the BERTopic model. This theoretical framework, commonly used in the field of technology and innovation management, can help assess the maturity and potential risks associated with trends. Future research can explore how the Earth Observation Hype Cycle, or similar frameworks, can be adapted to analyse consumer behaviour trends identified by the BERTopic model. This integration can offer insights into the readiness of businesses to adopt these trends, the associated investment risks, and the timing of market saturation.

Additionally, future research can explore limitations improves related to the utilization of the BERTopic model. These may include investigating the model's scalability to accommodate larger datasets, examining the robustness of its performance across different industries or markets, and evaluating the model's effectiveness in multi-channel consumer behaviour analysis.

# Conclusion

This study has investigated the application of machine learning models, such as BERTopic and LDA, for the analysis and forecasting of socio-behavioural trends in the business context. The findings emphasize the potential for these models to identify and forecast trends in consumer behaviour, social norms, and communication patterns, allowing businesses to make informed decisions and develop effective strategies. The significance of this study resides in its contribution to the development of new methods and instruments for comprehending and capitalizing on socio-behavioural trends, thereby augmenting business success. The insights and implications from this research can guide businesses and policymakers in harnessing the power of machine learning while upholding ethical standards and promoting sustainable growth.

# References

[1] Tasner, M. Marketing in the Moment: The Practical Guide to Using Web 3.0 Marketing to Reach Your Customers First. 2011.

[2] Duhigg, C. The Power of Habit: Why We Do What We Do in Life and Business. 2012.

[3] Berger, J. Contagious: How to Build Word of Mouth in the Digital Age. 2013.

[4] Kahneman, D. Thinking, Fast and Slow. 2011.

[5] Solomon, M.R., Armstrong, G.B. Consumer Behavior. 2019.

[6] Agarwal et al. "Predicting consumer behaviour with Web search data: a dynamic approach using boosting trees." 2010.

[7] Moura et al. "Using machine learning to predict consumer preferences: an application in the wine industry." 2021.

[8] Yen et al. "Predictive modelling of customer churn in banking industry using machine learning algorithms." 2021.

[9] Bradlow, E.T., Gangwar, M., others. "The Role of Big Data and Predictive Analytics in Retailing." 2017.

[10] Guo, L., Shi, F., Tu, J. "Textual analysis and machine learning: Crack unstructured data in finance and accounting

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

[12] Aravind. Earth Observation Hype Cycle: 2023 Edition. [Электронный ресурс]. URL: <https://newsletter.terrawatchspace.com/p/earth-observation-hype-cycle-2023>

# Appendices

GitHub link: <https://github.com/AnikaJha/Socio-Behavioral_trends>