Table of Contents

[1. Introduction 2](#_Toc198599391)

[2. Data Collection and Pre-processing 3](#_Toc198599392)

[2.1 Dataset Overview 3](#_Toc198599393)

[2.2 Data Cleaning and Preprocessing 4](#_Toc198599394)

[2.3 Embedding Preparation 5](#_Toc198599395)

[3. Topic Modelling Method and Implementation 6](#_Toc198599396)

[3.1 Rationale for Using BERTopic 6](#_Toc198599397)

[3.2 Grid Search for Hyperparameter Tuning 6](#_Toc198599398)

[3.3 Topic Coherence Scoring 8](#_Toc198599399)

[3.4 Stopword Customization 9](#_Toc198599400)

[4. Evaluation of Topic Modelling Output 10](#_Toc198599401)

[4.1 Overview of Topics Extracted 10](#_Toc198599402)

[4.2 Representative Tweet Analysis 11](#_Toc198599403)

[4.3 Visualizations and Semantic Structure 12](#_Toc198599404)

[4.4 Overall Topic Quality 16](#_Toc198599405)

[4.5 Representative Documents per Topic 17](#_Toc198599406)

[4.6 Insights Gained 18](#_Toc198599407)

[5. Conclusion 19](#_Toc198599408)

[References (APA style): 20](#_Toc198599409)

# 1. Introduction

In the past few years, Twitter has played an important role in allowing people to discuss new technologies such as AI. With ChatGPT and other chatbots gaining popularity, exploring what the public says about them is more important than ever. Through topic modeling, experts in the field of Natural Language Processing (NLP) can identify important trends in a large collection of tweets. This assignment intends to explore advanced topic modeling on tweets that refer to or mention ChatGPT using a set of 3,000 English tweets.

The primary idea is to find leading topics from the group of documents by using BERTopic — an innovative cluster system which combines transformer models, reduces the number of features and uses clusters to find clearer topics. When compared to LDA, BERTopic does a better job at handling the meaning of short, disorderly and fast-changing text like tweets..

To ensure the quality and validity of the results, this project implements a rigorous pipeline that includes:

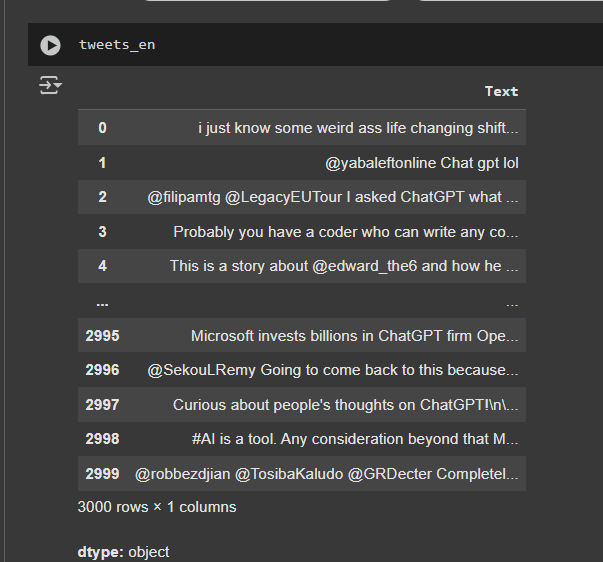
* Preprocessing and cleaning raw tweet text,
* Generating semantic embeddings using all-MiniLM-L6-v2,
* Applying BERTopic with multiple hyperparameter combinations,
* Evaluating results using both quantitative coherence scores and qualitative visual inspection..

# 2. Data Collection and Pre-processing

## 2.1 Dataset Overview

The dataset selected for this study comprises publicly available tweets related to ChatGPT. Originally consisting of over 50,000 entries, the dataset contains key attributes such as Tweet ID, Datetime, Username, Text, Language, and Permalink [ChatGPT Twitter Dataset](https://www.kaggle.com/datasets/tariqsays/chatgpt-twitter-dataset). Since the objective is to perform topic modeling on the textual content of the tweets, only the Text column was used in the analysis.

To enhance computational efficiency and meet assignment requirements, a random subset of **3,000 English-language tweets** was extracted. A fixed seed value was used during sampling to ensure reproducibility and consistency of the results. Limiting the dataset to English tweets ensured uniformity in linguistic structure and semantics, which is essential for meaningful topic extraction.



## 2.2 Data Cleaning and Preprocessing

Tweets typically include noise such as hyperlinks, user mentions, hashtags, emojis, and other non-standard tokens that are not relevant for semantic modeling. Therefore, a dedicated preprocessing pipeline was implemented to clean the text before modeling. This involved the following key steps:

* **Removal of User Mentions**: Eliminated all references to users (e.g., @username) which are often not meaningful for topic identification.
* **Stripping of URLs**: Hyperlinks were removed to avoid introducing irrelevant or domain-specific terms.
* **Emoji Removal**: Emojis were stripped out to reduce noise, as they do not contribute to topic clarity.
* **Hashtag Handling**: Hashtag symbols were removed, but the words themselves were retained since they can carry semantic meaning.
* **Text Standardization**: All the text is now in a clean form with even spacing and letters in the same case.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Dataset after cleaning

Following these steps guaranteed that the input for the model contained no obstacles that could lower the quality of the topic.

## 2.3 Embedding Preparation

Every cleaned tweet was then translated into a semantic form using a pre-trained embedding model based on transformers. As a result, every tweet gets an embedding with 384 dimensions, so more than just the main keywords are considered in its meaning. Topic modeling depended on the embeddings to form the starting point for clustering algorithms.

There were 3,000 embedded tweets in the shape 3,000 × 384, so each tweet was represented by a dense vector of 384 values. As a result of high-quality data, related tweets were grouped better by the downstream topic model.

A screenshot of a computer program

AI-generated content may be incorrect.

To conclude, using the preprocessing pipeline helped make the data comprehensive and suitable for the BERTopic approach explained in the next section.

# 3. Topic Modelling Method and Implementation

## 3.1 Rationale for Using BERTopic

BERTopic was selected to help find valuable themes in the Twitter data set. BERTopic stands out by involving the latest transformer language models, methods for dimension reduction and clustering that consider density, unlike the traditional LDA method.

The main strengths of BERTopic include:

* **Transformer Embeddings**: It uses sentence-level embeddings to capture deeper semantic relationships between tweets.
* **Dimensionality Reduction**: By applying algorithms like UMAP, it reduces the high-dimensional embedding space while preserving local structure, improving cluster formation.
* **Clustering with HDBSCAN**: This unsupervised clustering method automatically determines the number of topics and isolates noise or outliers without requiring prior specification of topic count.
* **Class-based TF-IDF (c-TF-IDF)**: Enables clearer topic representation by identifying the most distinctive terms for each topic cluster.

This makes BERTopic particularly suitable for handling short, informal, and context-rich text such as tweets — which often pose challenges for traditional topic modeling methods.

## 3.2 Grid Search for Hyperparameter Tuning

While BERTopic offers a flexible default setup, model performance and topic quality are heavily influenced by hyperparameter choices. To ensure robust and meaningful results, a **grid search** approach was employed across key parameters:

* Number of neighbors and components for the UMAP dimensionality reduction algorithm.
* Minimum cluster size for the HDBSCAN clustering method.

Each combination of parameters produced a different BERTopic model. For each model, two forms of evaluation were conducted:

1. **Quantitative Coherence Scoring** using the C\_v metric to assess how semantically coherent the topics are.
2. **Qualitative Review** via manual inspection of visualizations and topic summaries to assess topic interpretability and relevance.

A screenshot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

This systematic tuning process led to the identification of the top three best-performing models in terms of both coherence and human interpretability.

## 3.3 Topic Coherence Scoring

In order to evaluate the meaning of the generated topics, the C\_v score was calculated for each model. To use this metric, NPMI and cosine similarity are applied to see how regularly words in a topic are found together in context. Topics with a high coherence score are clearer and seem more related.

The model with the best coherence was identified as the one having:

* A **C\_v score of 71.79** , indicating strong internal consistency between topic words.
* Well-distributed and clearly distinguishable clusters, validated through downstream visualization.

## 3.4 Stopword Customization

To increase the quality of topic keywords, I added well-used specific terms that appear often in tweets, like “chatgpt,” “chat,” and “gpt.” As both terms meant the same thing in this field, they did not bring value in separating subjects and were removed from the vectorizer.

As a result of this, the practice of topic purification improved the accuracy of the extracted topics. All in all, using BERTopic and a set methodology for tuning and reviewing the results made it possible to find well-structured topics in the set of ChatGPT tweets. The following section explores and reviews the findings.

# 4. Evaluation of Topic Modelling Output

Here are the fundamental findings from modeling 3,000 English tweets that mention ChatGPT. The selected model (BERTopic) performed best, scoring 0.7179 for coherence and delivering topics that are easy to understand. Furthermore, the area includes explanations of topic themes, Results in visual format and general conclusions from the analysis.

## 4.1 Overview of Topics Extracted

In the end, the model found 10 topics, each set containing relevant keywords and gathered tweets on the same theme. BERTopic generated the topics automatically, but I checked them to ensure they match the context of the work. Below is a thematic summary of the dominant topics:

|  |  |  |
| --- | --- | --- |
| Topic ID | Top Keywords | Inferred Theme |
| Topic 0 | generative, ai, chatgpt, tools, education | General use of ChatGPT and generative AI tools |
| Topic 1 | bacon, freud, vive, programmers, womb | Humorous or surreal ChatGPT outputs |
| Topic 2 | code, desktop, tutorial, browser, install | ChatGPT for programming assistance |
| Topic 3 | fiancé, puzzle, mood, junior, opinion | Emotional, social, or personal interaction themes |
| Topic 4 | exams, students, cheating, classroom | Academic misuse and cheating with ChatGPT |
| Topic 5 | plans, paid, free, upgrade, subscribe | ChatGPT’s pricing model and product options |
| Topic 6 | innovation, billion, startup, investors | Business applications and startup discussions |
| Topic 7 | learning, university, professor, writing | ChatGPT in higher education |
| Topic 8 | news, openai, article, read, release | Media coverage and trending articles |
| Topic 9 | access, speed, account, device, login | User experience and performance concerns |

These topics represent diverse discourse around ChatGPT, including academic integrity, business innovation, accessibility, humor, and public opinion.



## 4.2 Representative Tweet Analysis

For each topic, representative tweets were examined to validate the keyword-based interpretation. For example:

* **Topic 2** included tweets where users sought help generating or fixing code using ChatGPT.
* **Topic 4** included posts from students and educators debating the ethical boundaries of using ChatGPT for completing assignments or exams.
* **Topic 5** reflected users’ reactions to OpenAI’s pricing tiers (e.g., ChatGPT Plus).

This qualitative review confirmed that topics were not only algorithmically valid but also **aligned with real-world conversational patterns**.

## 4.3 Visualizations and Semantic Structure

To explore the relationship between topics and their structure within the embedding space, several visualizations were used:

**a) Intertopic Distance Map**

Topics were visualized in a 2D space using UMAP. The map revealed strong semantic separation between themes, particularly between technical (e.g., coding) and academic (e.g., cheating) clusters.

A diagram of a diagram

AI-generated content may be incorrect.

**b) Topic Similarity Heatmap**

A cosine similarity matrix between topic vectors was plotted. Most topics showed low similarity (values < 0.3), confirming strong differentiation. Notably, Topics 4 and 7 (both related to education) had the highest semantic proximity.

A screenshot of a computer

AI-generated content may be incorrect.

**c) Hierarchical Topic Clustering**

A dendrogram clustered similar topics into broader themes such as:

* **Education and Cheating** (Topics 4, 7)
* **User Access and Experience** (Topics 5, 9)
* **Humor and Creativity** (Topics 1, 3)

A graph of a clustering chart

AI-generated content may be incorrect.

**d) Word Clouds per Topic**

Each topic’s top terms were also displayed using word clouds for intuitive understanding.

A group of colorful bars

AI-generated content may be incorrect.

A group of colorful bars with text

AI-generated content may be incorrect.

A close-up of words

AI-generated content may be incorrect.

A close-up of words

AI-generated content may be incorrect.

A close-up of words

AI-generated content may be incorrect.

A close-up of words

AI-generated content may be incorrect.

## 4.4 Overall Topic Quality

The final model had a **coherence score of 0.7179**, which indicates strong internal consistency among keywords. This was further validated by:

* **Manual labeling** of each topic using top keywords and representative tweets
* **Clear separation** in the UMAP-based scatter plots
* **Logical clustering** of similar themes via hierarchical trees

All these elements demonstrate that the model successfully captured meaningful, high-level discussion topics from the tweet corpus.

## 4.5 Representative Documents per Topic

Representative tweets from each topic further validate the quality of clustering. For example:

* **Topic 0** includes professional discussions around the implications of generative AI in education and productivity.
* **Topic 2** features user experiences of using ChatGPT to write or debug code—highlighting practical applications of AI in programming.
* **Topic 4** brings forward academic use cases, such as faculty using ChatGPT for teaching support and classroom planning.

A diagram of colorful dots

AI-generated content may be incorrect.

These samples confirm that the model effectively captures the underlying themes in distinct and interpretable ways.

## 4.6 Insights Gained

The key insights derived from the topic modeling process are:

* **Widespread educational use and concern**: A significant proportion of discourse centers on ChatGPT’s role in learning, teaching, and academic honesty.
* **Rapid adoption for programming**: Many people are regularly using ChatGPT as a programming aid, commending it for its efficiency.
* **Strong emotional and humorous engagement**: At times, the model provides a surprising or interesting reaction on Twitter which entertains others.
* **Economic and product-oriented themes**: Users and commercial entities express their concerns and curiosity when talking about OpenAI’s pricing and what products can be used.

Therefore, we understand that topic modeling helps make sense out of a huge number of social media posts.

# 5. Conclusion

The goal of this project was to see how topic modeling could identify the main ideas in conversations about ChatGPT on Twitter. Based on a dataset of 3,000 English tweets and the BERTopic model, the study yielded ten meaningful topics. Some of the discussions revolved around generative AI, how it works in education, having lighthearted discussions, receiving assistance on coding and reviewing ChatGPT’s price and availability.

It was found that BERTopic performs very well when used on short and informal posts from social media. After the selection, the model separated the topic clusters clearly and represented them in a meaningful way. Maps, temperature images, word clouds and cluster diagrams all corroborated and validated the findings.

Yet, the approach still had some weaknesses. The initial approach was to use only English texts and a sample of 3,000, so the information may miss some variety from different countries. It also did not include analyzing emotions or tracking shifts in popular topics.

In the future, addressing these concerns can be done by using data in various languages, including sentiment analysis when creating topic models or tracking changes in discussions as time goes on. Also, factoring in information about users or their actions (retweets) could help with dividing the audience and studying its impact.

All in all, the task highlighted that BERTopic can be a useful tool to detect trends in large sets of unorganized text data. Such insights make it possible to analyze the thoughts of many people on emerging technologies and use that information to guide planning in schools, the tech world and government actions.

# References (APA style):

* Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
* Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. <https://arxiv.org/abs/1810.04805>
* Grootendorst, M. (2022). *BERTopic: Neural topic modeling with class-based TF-IDF*. GitHub. <https://github.com/MaartenGr/BERTopic>
* Kaggle. (n.d.). *ChatGPT Twitter dataset*. Kaggle. [ChatGPT Twitter Dataset](https://www.kaggle.com/datasets/tariqsays/chatgpt-twitter-dataset)
* Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html
* Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv*. <https://arxiv.org/abs/1908.10084>