**ICT 202 Assignment 2 Final Report**

*Murdoch University*

EDITED BY

**Keith Neo Kai Si**

**35107628**

*Harnessing BERT for Enhanced Topic Modelling in Natural Language Processing*

Table of Contents

[**Introduction** 3](#_Toc172765369)

[**Data Collection and Pre-processing** 3](#_Toc172765370)

[**Topic Modelling** 6](#_Toc172765371)

[**Unsupervised Topic Modelling Evaluation** 7](#_Toc172765372)

[**Topic Modelling Results** 8](#_Toc172765373)

[**Conclusion** 9](#_Toc172765374)

[**Reference** 10](#_Toc172765375)

## **Introduction**

Social media platforms such as Twitter play a vital role in expressing opinions, disseminating information, and fostering connections among users. Analysing text data from these platforms through topic modelling techniques is essential for gaining insights into user sentiments and preferences. This project centres on topic modelling using Natural Language Processing (NLP) techniques, particularly leveraging Bidirectional Encoder Representations from Transformers (BERT). The objective is to uncover prevalent topics from Twitter data, revealing underlying themes and trends.

**Assignment Overview:**

Data Source:

* The dataset “All Trump’s Twitter insults (2015-2021)” from [Kaggle](https://www.kaggle.com/datasets/ayushggarg/all-trumps-twitter-insults-20152021).

Topic Modelling Analysis:

* **Data Exploration and Visualisation**
* **Data Preprocessing**
* **Feature Extraction**
* **Model Training**
* **Model Visualisation**
* **Model Evaluation**
* **Interpretation of the Results**

Justification and Explanation:

* Explanation of chosen techniques and steps, with justifications.
* Documentation of any deviations from the suggested steps.

Interpretation of Results:

* Interpretation of the identified topics, including visualisations.
* Concluding the analysis, providing insights into the data’s significance.

## **Data Collection and Pre-processing**

**Data Source**

The dataset titled “All Trump’s Twitter Insults (2015-2021)” from [Kaggle](https://www.kaggle.com/datasets/ayushggarg/all-trumps-twitter-insults-20152021) was chosen for its comprehensive collection of tweets, which provides a solid foundation for analysing public discourse and identifying key themes. This dataset is particularly valuable due to its rich content and significance in understanding political communication and social media behaviour.

**Data Pre-processing**

1. Importing Libraries and Dataset

* Key libraries such as pandas and NLTK were imported to facilitate data manipulation and text processing.
* The dataset was loaded into a DataFrame, allowing for efficient handling and analysis of the data.

1. Initial Data Cleaning

* Unnecessary columns were removed to concentrate on the relevant text data.
* The **data** column was converted to **datetime** format, enabling accurate temporal analysis of the tweets.
* All tweets were consolidated into a single string for initial exploration and preliminary text analysis.

1. Text Cleaning and Preparation

* **Lowercase Conversion:** All text was converted to lowercase to ensure uniformity and reduce redundancy.
* **Punctuation Removal:** Punctuation marks were eliminated to remove irrelevant characters that do not enhance the text’s meaning.
* **Numerical Data Removal:** Numbers were removed to focus on the linguistic aspects of the content.
* **Tokenization:** The text was divided into individual tokens (words), facilitating a detailed word-level analysis.
* **Stop Words Removal:** Commonly used words that lack significant meaning (e.g. and, the) were removed to highlight key content.
* **Lemmatization:** Words were reduced to their base or root form, consolidating different word variations and improving data consistency.

1. Application of Cleaning Function

* The cleaning function was applied to each tweet, standardising the text data for analysis

1. Saving Cleaned Data

* The cleaned tweets were saved to a new CSV file, ensuring data integrity and readiness for subsequent analysis stages.

This comprehensive preprocessing ensured that the text data was clean, consistent, and properly formatted for effective BERT topic modelling. Each step was carefully designed to enhance data quality, thereby improving the accuracy and reliability of the analysis.

## **Topic Modelling**

I utilised the BERT (Bidirectional Encoder Representation from Transformers) model for topic modelling. BERT’s transformer-based architecture effectively captures the context of words by analysing their surrounding terms, making it well-suited for grasping the nuances of language and identifying topics.

**Selected Method: BERTopic**

BERTopic was selected for its capability to leverage BERT embeddings in conjunction with clustering algorithms to uncover coherent and contextually relevant topics. The model was adapted BERTopic documentation. (Grootendorst, n.d.) and an article on topic modelling in Python (David, 2021).

**Implementation Steps:**

1. **Importing and Cleaning Data:**

* The dataset “All Trump’s Twitter Insults (2015-2021)” underwent cleaning and preprocessing and is imported in this step to train the BERTopic model.

1. **Training the BERT Model:**

* The BERTopic model was trained on the cleaned dataset. If a pre-existing model is unavailable, it will be trained from scratch; otherwise, the pre-trained model was loaded to optimise time efficiency. An essential parameter set during training was the **verbose=True** flag, which provided detailed logs throughout the training process.

1. **Libraries and Tools:**

* **BERTopic:** For topic modelling utillising BERT embeddings.
* **Pandas:** For data manipulation and management.
* **NLTK:** For text preprocessing tasks, including tokenisation and lemmatisation.
* **Scikit**-**Learn:** For additional machine learning functions, including feature extraction and evaluation.
* **Genism**: For calculating coherence scores.
* **Matplotlib and seaborn**: For data visualisation
* **Wordcloud**: For creating word clouds to visualise the most frequent words within topics.

## **Unsupervised Topic Modelling Evaluation**

Two primary metrics were utilised to assess the quality of the topics generated by the BERTopic model: Coherence Score and Silhouette Score.

**Coherence Score:**

The coherence Score quantifies the semantic similarity among the words within a topic. Higher coherence scores suggest that the issues are more meaningful and accessible to interpret.

**Implementation:**

I employed Gensim’s Coherence Model to compute the coherence score. This model was initialised with the topics identified by BERTopic, the preprocessed text data and a Gensim dictionary. The coherence score was then calculated using the **c\_v** metric.

**Silhouette Score:**

The silhouette score evaluates how similar a data point is to its cluster compared to others. This score ranges from -1 to 1, where higher values indicate better-defined clusters.

**Implementation:**

The cleaned dataset was shuffled and subsampled to include 1,000 tweets. Topic assignments and probabilities were extracted from the BERTopic model. The Silhouette score was calculated using the silhouette\_score function from scikit-learn based on the topic probabilities and assignments.

## **Topic Modelling Results**

**Results:**

The BERTopic model successfully identified a range of coherent topics from the dataset. Each topic is characterised by a collection of critical terms that encapsulate its main theme.

**Visualisation:**

Multiple visualisations were created to illustrate the outcome of the topic modelling effectively:

1. **Topic Information:**

An overview of the topics, including their size, was obtained using the **get\_topic\_info** method.

1. **Top Words in Each Topic:**

The **get\_topic** method was utilised to showcase the top words associated with each topic, offering insights into the primary themes.

1. **Topic Frequency:**

The frequency of each topic was visualised to highlight the most prominent topics within the dataset.

1. **Topic Visualisations:**

Interactive visualisations were generated using the **visualize\_topics, visualise\_barchart, visualise\_heatmap,** and **visualise\_hierarchy** methods, illustrating the relationships and hierarchical structure of the topics.

1. **Word Clouds:**

Word clouds were created for each topic to visually represent the most frequent words, facilitating the interpretation of the topics.

**Interpretation:**

The identified topics were analysed to uncover the underlying themes and patterns present in the dataset. The visualisations played a crucial role in identifying key topics and their relevance. Additionally, the coherence and silhouette scores provided quantitative metrics to evaluate the model’s performance, reflecting the quality of the generated topics.

## **Conclusion**

In this assignment, I investigated the application of BERT for topic modelling on the dataset. The key findings are as follows:

* **Effective Topic Identification:** BERT successfully identified coherent and contextually relevant topics, revealing patterns and common targets of insults.
* **Importance of Preprocessing:** Comprehensive preprocessing, which included text cleaning, tokenisation, stopword removal, and lemmatisation, was essential for accurate topic extraction.
* **Quantitative Evaluation:** Coherence and silhouette scores validated the model’s capability to produce well-defined topics.
* **Comprehensive Visualisations:** Various visualisations facilitated a deeper understanding of the main themes and the relationships among topics.

**Limitations:**

* **Computational Constraints:** The large dataset size necessitated using a subset for analysis, which may have limited the comprehensiveness of the findings.
* **Topic Granularity:** Some topics appeared broad or overlapping, indicating a need for further model refinement.

**Future Work:**

* **Model Refinement:** Exploring different parameters or preprocessing techniques could improve topic coherence and specificity.
* **Analysis of Larger Datasets:** Examining the entire dataset or utilising advanced computational resources could yield more comprehensive insights.

Overall, BERT proved to be an effective tool for modelling political social media data topics, providing valuable insights and establishing a foundation for future research and enhancements.

## **Reference**

David, D. (2021, August 24). *NLP Tutorial: Topic Modeling in Python with BerTopic*. Retrieved from Hackernoon: https://hackernoon.com/nlp-tutorial-topic-modeling-in-python-with-bertopic-372w35l9

Grootendorst, M. P. (n.d.). *BERTopic*. Retrieved from BERTopic: https://maartengr.github.io/BERTopic/