Harnessing Topic Modelling to Hate-speech on Twitter: A Comparison of LDA and BERTopic Techniques

Abstract- **"We must confront bigotry by working to tackle the hate that spreads like wildfire across the internet" [1]. The challenge of addressing the soaring content of hate speech on digital tools has risen with the popularity of easily shareable social platforms. For example, on Twitter (now known as Platform X), before Elon Musk bought the company, slurs against Black Americans appeared 1,282 times a day. After his purchase, this number jumped to 3,876 times a day [2]. Therefore, in the spirit of combating this issue, we aim to utilize topic modeling techniques, especially Latent Dirichlet Allocation and BERTopic models, to help identify common topics most associated with hate speech sentiments. This will provide insights into the causes of the increase in hate speech, in hopes of protecting vulnerable communities from hate speech attacks.**

Keywords— Hate Speech detection, Twitter topic modelling, LDA, BERTopic.

[Please refer to coding files: *AE2_BERT.ipynb* and *AE2_LDA.ipynb* for detailed of coding]

I. Introduction

The rise of hate speech on social media platforms, such as Twitter (now known as Platform X), is increasingly concerning due to the platforms' ease of content sharing. Balancing the right to free speech with the imperative of safety presents formidable challenges. In this context, the detection of hate speech is crucial for identifying severe cases and effectively combating online abuse. This approach aligns with legislative efforts like the Online Safety Bill in the UK, which mandates that digital and social media platforms curb the proliferation of illegal content, including hate speech, particularly to protect children from harmful online content. [3]

In response to these challenges, this assignment employs topic modelling techniques, specifically through the use of Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers-Topic (BERT-topic) models, to analyse a dataset of tweets manually classified into categories of *hate speech, offensive speech, or neither*. By applying these models, we aim to uncover the topics most vulnerable to hate speech attacks and to distinguish content accurately as hate speech. Topic modelling is a method for the unsupervised classification of documents; It is an excellent technique for discovering hidden themes and organizing, summarizing, and searching documents. Utilising topic modelling methods may help us uncover hidden topics in hate speech tweets, for example, what specifically provokes hate speech comments and perhaps identify vulnerable groups subject to hate speech attacks.

In this analysis, we will deploy two distinct models on our dataset to evaluate their efficacy in topic modelling and their potential to enhance hate speech detection. First, we employ the Latent Dirichlet Allocation (LDA), a probabilistic statistical model that identifies abstract topics within a document collection. LDA operates by assigning words to topics based on their distribution probabilities, effectively grouping them into thematic categories. For instance, words that frequently coincide are aggregated under a specific topic—labelled as 'topic 1'—with the most probable words defining the core of this topic. This method allows us to systematically explore and represent the underlying themes within the dataset. [4]

Second, we utilize BERTopic, which combines transformer architectures with class-based TF-IDF (cTF-IDF) to enhance topic detection and interpretation. Transformers, integral to BERTopic, use self-attention mechanisms to process sequences of data, identifying intricate patterns and long-range dependencies between words. This capability enables the model to discern subtle connections and thematic nuances across the dataset, potentially revealing deep-seated correlations within the tweets. The cTF-IDF component further refines this process by weighting terms based on their class relevance, improving the clarity and distinctiveness of the resultant topic clusters. [5][6]

By applying both LDA and BERTopic, we aim to cross-validate findings and determine which model more effectively discerns coherent topics linked to hate speech. This comparative analysis will provide insights into the strengths and limitations of each approach, guiding future strategies for monitoring and mitigating hate speech on social media platforms.

II. Literature review

In their paper, Shastry and Prakash undertake a detailed comparative study of three prominent topic modelling techniques—Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix Factorization (NMF)—to analyse and extract hidden thematic elements from web log data. Recognising the vast availability of web data as a rich source for understanding user behaviour, they employed these models on the dataset, which sourced from the Nginx server access logs from an online shopping store. These logs provided critical user-related information, such as IP addresses, access dates, and URLs visited.

The primary aim of their research is to enhance the understanding of user behaviours, preferences, and interactions on the web, with potential applications including the development of advanced web recommendation systems and the optimization of user experiences through tailored content. To this end, the authors initiated their methodology with rigorous data pre-processing, employing techniques such as Bag of Words and TF-IDF vectorization to prepare the data

for LDA, which identifies latent topics by analysing word distributions.

For the LSA method, the data was converted into a term-document matrix, subsequently reduced in dimensionality through Singular Value Decomposition (SVD). In contrast, NMF also utilizes a term-document matrix but decomposes it into separate matrices that explicitly represent topics and their corresponding document associations, a process which generally allows for easier interpretation of the derived topics.

To evaluate the effectiveness of these models, the authors applied coherence scores and perplexity metrics. Their findings indicate that NMF yielded the highest coherence scores, suggesting superior topic relevance and distinction. LDA, while not achieving as high coherence scores as NMF, was noted for its stability and coherence across different datasets, making it a robust choice for real-world applications. LSA, however, was found to lag in performance when compared to the other two techniques. Shastry and Prakash also highlight significant challenges in web data analysis, such as scalability and the sheer volume of data, which demand robust models capable of efficiently processing large datasets and yielding high-quality, interpretable results. [7]

In the study by Xu et al., the authors adopt a cluster-based BERTopic modeling approach to analyze Twitter discourses on mask usage during the Covid-19 pandemic. Their research objective is to delineate how conversations about masks shift over time and how these shifts manifest among diverse user groups with different identities and interests. Notably, they focus on the political dimension, examining variances in perception and communication about mask-wearing across political lines and public health discussions.

The study utilises a comprehensive dataset from the COVID-19 Twitter corpus curated by Georgia State University's Panacea Lab. Prior to topic modelling, the researchers grouped Twitter users into clusters based on the information available on their bio pages. They used k-means clustering algorithms to organise users into coherent groups reflecting various demographics and interests—such as conservatives, progressives, and public health professionals. They then implemented BERTopic modeling, a technique that capitalizes on BERT's (Bidirectional Encoder Representations from Transformers) capability for grasping the nuanced, contextual meanings of words, particularly adept for analyzing the concise and dynamic nature of tweets. The authors also engaged LDA and Semantic Network-based Classification (Textnets) for a comparative perspective.

Their findings indicate that while LDA faces challenges with the brevity and disorderly nature of tweets, BERTopic, leveraging BERT's powerful embeddings, excels at interpreting the semantic nuances of language in such short texts. Additionally, Textnets offers a complementary approach, utilizing community detection algorithms to map out a network of related terms and visualize the interconnections between topics and terms. Despite this methodological ingenuity, the BERTopic approach is deemed superior due to its sophisticated use of contextual embeddings derived from BERT. Finally, the paper discusses broader issues beyond the technical realm, such as the implications of political polarization and the pervasive effects of misinformation and disinformation on social media platforms. These discussions highlight the societal challenges that interweave with the technical aspects of topic modeling in the context of public health crises. [8]

In the study by Sapul et al., the researchers employ clustering and topic modeling techniques to interpret trending topics on Twitter. Their investigation pivots on the comparative analysis of two clustering methods—k-means and CLOPE—and the Latent Dirichlet Allocation (LDA) topic modeling approach. Their analysis revolves around a corpus of Twitter data pertaining to the APEC 2016 summit. This dataset, comprising 10,668 tweets encompassing replies, retweets, and original content, was harvested during the summit's duration.

After implementing preprocessing protocols and executing feature selection and extraction procedures, the researchers applied k-means, CLOPE, and LDA to the data. The comparative analysis of these methods yielded notable insights: k-means clustering requires more contextually rich feature sets to function optimally.

CLOPE clustering, in contrast, exhibited an innate aptitude for generating coherent clusters even with minimal feature sets, such as standalone hashtags. Moreover, it surpassed other algorithms in terms of the number of meaningful topics recognized when analyzing more complex feature sets. This finding suggests that CLOPE's design, particularly adept at managing categorical data, allows for robust performance even with non-transactional datasets like Twitter data.

Lastly, LDA showed proficiency in recognizing topics that align with the predetermined thematic spectrum of the APEC summit, particularly when deploying a fusion of keywords and hashtags. Nonetheless, the consistency of LDA's performance fluctuated across varying feature sets. The model's core strength lies in extracting themes from textual data through the distributional patterns of words, but its efficacy is modulated by the specificity and relevance of the chosen features [9].

Lastly, in the work of Kstrati et al., the authors explore a different topic and aim to understand public engagement on Twitter during periods of rising energy prices by using sentiment analysis and topic modelling with transformers. The purpose is to track the evolution of public sentiment towards energy prices and to identify key topics of discussion related to this issue. The dataset comprises 366,031 tweets, collected from January 1, 2021, to June 18, 2022, using the Twitter API and selected based on keywords and hashtags related to energy prices.

After performing pre-processing steps, such as removing duplicates, non-English texts, URLs, hashtags, mentions, and any non-ASCII characters, they applied Transformer-Based Sentiment Annotation. The tweets were annotated for sentiment (positive, neutral, negative). They compared BERT with four other lexicon-based sentiment analysis methods, finding that BERT outperformed all others. Additionally, the authors employed LDA as a benchmark against BERTopic, with the findings suggesting that BERTopic was particularly effective in identifying clear and coherent topics, compared to LDA. [10]

III. Problem setting

In this assignment, we aim to leverage the topic modeling technique, specifically the BERTopic model, with LDA as a benchmark model, to identify hate speech patterns on the Twitter platform. We sourced the dataset from Kaggle, named 'Hate Speech and Offensive Tweets by Davidson et al.' This dataset contains approximately 24,000 examples and is classified into three categories: **0 = hate speech, 1 = offensive language, and 2 = neither**. Notably, Davidson et al. reported that only 5% of the tweets fall into the hate speech category. The authors explained that even tweets containing multiple racial or homophobic slurs were sometimes not identified as hate speech, due to the complexity of context and intent behind the words. Therefore, we will use the topic modeling technique to see if it reveals new perspectives on which topics constitute hate speech and which words are specifically used to represent it.[11]

IV. Model application and Discussion

1.0 Text pre-processing

Prior to the implantation of the LDA model, we performed Exploratory Data Analysis (EDA). We created a text preprocessing function **(`simple_normalised_text`)** to facilitate simple text cleaning for the generation of a word cloud. This included:

- **Checking for missing values**
- **Converting to lowercase**
- **Removing numbers**
- **Removing punctuation and special characters**

We also created another text preprocessing function **(`normalised_text`)** in preparation for the LDA model, which includes:

- **Checking for missing values**
- **Converting to lowercase**
- **Removing numbers**
- **Removing punctuation and special characters**
- **Removing Twitter handles and URLs**
- **Tokenization**
- **Removing stop words**
- **Lemmatization**

The additional steps reduce the noise in the data and make patterns clearer, especially for LDA. As seen in the literature review [8], LDA generally struggles with noisy texts; therefore, these steps could enhance LDA's performance.

For BERTopic, the text preprocessing is slightly different and includes:

- **Checking for missing values**
- **Converting to lowercase**
- **Removing numbers**
- **Removing punctuation and special characters**
- **Removing Twitter handles and URLs**
- **Tokenization**

We did not include further cleaning, such as removing stop words and lemmatization, because BERTopic can benefit from the extra noise to help it identify contextual relations.

2.0 Model Design

2.1. Latent Dirichlet Allocation (LDA) design

After applying the text pre-processing function to our dataset, we used the **Gensim** library to build a dictionary of all unique tokens in the dataset **(HST_token)**. Using this dictionary, we constructed a corpus where each tweet is transformed into a bag of words (BoW) format and represented by a list of tuples. Each tuple contains a token ID and its corresponding frequency. We then trained a Latent Dirichlet Allocation (LDA) model using **'LdaMulticore'** from **Gensim**. The model is configured to identify three latent topics from the corpus. Following this, we used the **pyLDAvis** package to generate visualisations to aid our analysis and to determine whether the model can generate coherent outputs. Additionally, we calculated the perplexity score to evaluate the model performance.

2.2 Bidirectional Encoder Representations from Transformers-Topic (BERT-topic)

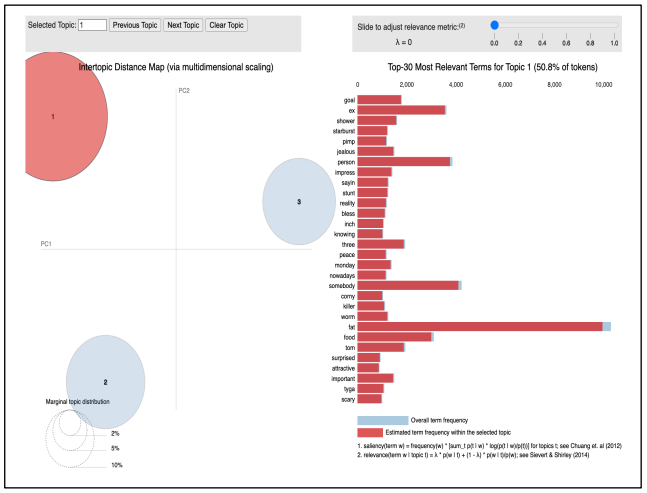
After applying the text pre-processing function, we initialise the model to generate coherent and informative topics from large text corpora. The transformer model we use is designated as **'paraphrase-MiniLM-L3-v2'**, and the **'min_topic_size'** is set to 7, indicating seven documents will not be considered during topic modeling. This helps focus the model on significant and robust topics. We use **model.get_topic_info** to retrieve information about the topics, allowing for human interpretation to assess whether the extracted topics make sense to human readers. Additionally, we use **model.visualize_barchart** to create a bar chart visualisation for better interpretation of the results.

V. Results discussion

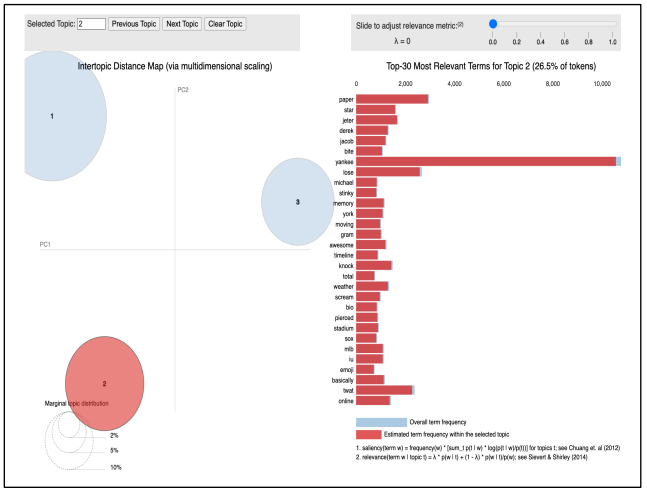
1.0 LDA results

Referring to figures 1-3, the Inter-topic Distance Map is presented on the left side, showcasing a two-dimensional, multidimensional scaling (MDS) representation of topic distances. In this map, each circle represents a different topic; the size of the circle reflects the marginal topic distribution—essentially, the prevalence of the topic in the dataset. The distance between any two circles suggests how similar or different they are; On the right side, a bar chart demonstrates the Top-30 Most Relevant Terms for each topic. By clicking on each topic, you can view the results for the respective topics. The terms are arranged in descending order of relevance, which combines term frequency within the selected topic and the term's distinctiveness across topics. This relevance metric can be adjusted by a slider (λ), which balances the weight between term frequency within the topic and exclusivity of the term to the topic.

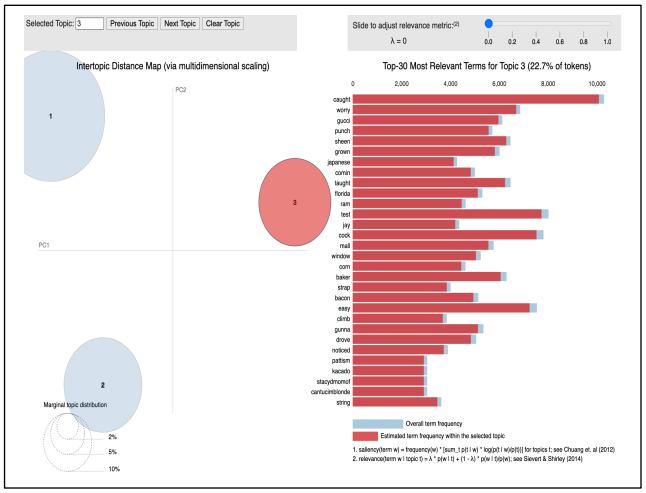
In Graph 1, the Top-30 Most Relevant Terms for Topic 1 comprise 50.8% of the total tokens. My examination does not reveal a clear pattern among these terms, but by analysing them individually, terms such as "goal," "ox," "shower," "starburst," "pimp," "jealous," and "person" emerge. These terms suggest the thematic content associated with Topic 1. The presence of seemingly unrelated terms such as "goal," "ox," and "starburst" indicates that the topic may be quite broad or that the model has grouped together documents with varied contexts. More suggestive terms like "pimp," "jealous," "impress," and "killer" may relate directly to aggressive or derogatory language potentially found in hate speech or offensive content. Overall, these results do not allow for more precise conclusions without a more in-depth analysis of specific content.



Graph 2 displays the Top-30 Most Relevant Terms for Topic 2, accounting for 26.5% of the total tokens. Specific names or nouns like "jeter," "derek," "jacob," "yankee," "michael" may indicate a discussion centered around public figures or entities. In the context of hate speech or offensive content, these could be targets or subjects of tweets. Terms that could be considered derogatory or relate to negative behaviors, such as "bite," "lose," "stinky," and "twat," might be used in a manner that is offensive or insulting. Topic 2 seems to be characterized by a mixture of personal names and potentially offensive terms interspersed with neutral words. This could indicate that the topic involves discussions about or directed at specific individuals, possibly with negative or hateful connotations, mixed with general commentary or descriptions of events.



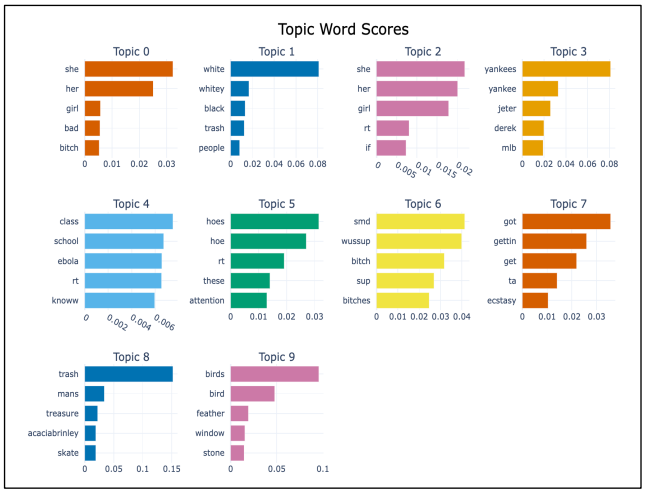
Graph 3 visualizes Topic 3 from the topic modelling analysis, which reveals a diverse set of terms accounting for 22.7% of the tokens in the dataset. The presence of brand names or cultural references (e.g., "gucci") could imply that the tweets incorporate contemporary or lifestyle topics. The use of potentially violent language (e.g., "punch") and judgmental terms points to the offensive nature of the tweets within this topic. The results suggest that offensive language or hate speech in this topic could be interwoven with everyday topics or personal identifiers, which is a common feature in our datasets. This underlines the complexity of automatically detecting and understanding offensive content on social media, highlighting the importance of context in interpretation.



2.0 BERT-Topic results

Refer to graph 4, where we have a grid of bar charts, each representing Topic Word Scores for different topics extracted from a collection of tweets. The word scores reflect the importance or frequency of each term within the respective topic.

In topics 0, 1, 5, 6, and 8, we can observe derogatory wording related to specific themes. For example, topic 0 includes terms such as "she," "her," "girl," "bad," and "bitch," suggesting a focus on female-related content, but with words that carry derogatory meanings. Topic 1, refers to graph 5 contains words like "white," "whitey," "black," "trash," "people,", 'racist', 'colored', 'whitecholo', which could indicate racial discussions or slurs, potentially pointing to racially charged content. Overall, we can identify potential targets of hate speech on social media platforms, including women, race, and anti-police sentiment (topic 8, with "acab," often standing for "All Cops Are Bastards").




```
[('white', 0.08098387906984113),
 ('whitey', 0.016838638394043766),
 ('black', 0.01334036799933145),
 ('trash', 0.012514308058452897),
 ('people', 0.008279996724963547),
 ('racist', 0.007332367955077791),
 ('colored', 0.005517705251782793),
 ('full', 0.005069708610815048),
 ('whitecholo', 0.0048876623577434443),
 ('man', 0.004761581010940485 5)]
```

3.0 Model comparison

Overall, using manual interpretation of the results, we can conclude that BERT-Topic was able to generate higher quality, more coherent topic modeling outputs compared to LDA. This is evident as LDA does not manage to clearly generate a coherent topic model that can precisely identify groupings targeted by hate speech comments. However, according to the perplexity score from the LDA results, at **-7.921707260264309**, a negative log perplexity implies that the model's perplexity is quite low, and the more negative the log perplexity, the better the model's predictions are. Nonetheless, manual examination of the topic modeling outputs suggests it does not perform well. Our dataset may generally lack distinct topics and have inherent overlap in topics, as also discussed in Davidson et al., [11]. The LDA model is not well-suited to handle this overlap effectively.

VI. Conclusion

In this assignment aims to employ topic modelling techniques to identify potential topics that may be targeted by hate speech comments. We hope to use Topic Modeling in the future to detect hate speech behaviours and assist in eliminating hate speech on social media platforms. We found that BERTopic delivers promising performance, identifies coherent topics, and suggests that certain demographics or topics could be subjected to hate speech comments. We recommend that future studies could utilise clustering techniques and BERTopic algorithms to provide better insights into combating hate speech on social media platforms.

VII. Reference

[1] United Nations, "What is hate speech?," *United Nations*, 2023. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[2] S. Frenkel and K. Conger, "Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find," *The New York Times*, Dec. 02, 2022. Available: https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html

[3] "Online Safety Bill: supporting documents," *GOV.UK*. https://www.gov.uk/government/publications/online-safety-bill-supporting-documents

[4] R. Kulshrestha, "Latent Dirichlet Allocation(LDA)," *Medium*, Sep. 28, 2020. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2#:~:text=It%20is%20one%20of%20the

[5] R. Kulshrestha, "Transformers," *Medium*, Nov. 22, 2020. https://towardsdatascience.com/transformers-89034557de14

[6] M. P. Grootendorst, "Home," *maartengr.github.io*. https://maartengr.github.io/BERTopic/index.html#quick-start (accessed Apr. 13, 2024).

[7] P. Shastry and C. O. Prakash, "Comparative analysis of LDA, LSA and NMF topic modelling for web data," *AIP Conference Proceedings*, Jan. 2023, doi: https://doi.org/10.1063/5.0178761.

[8] W. W. Xu *et al.*, "Unmasking the Twitter Discourses on Masks During the COVID-19 Pandemic: User Cluster–Based BERT Topic Modeling Approach," *JMIR Infodemiology*, vol. 2, no. 2, p. e41198, Dec. 2022, doi: https://doi.org/10.2196/41198.

[9] Ma. S. C. Sapul, T. H. Aung, and R. Jiamthapthaksin, "Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms," *IEEE Xplore*, Jul. 01, 2017. https://ieeexplore.ieee.org/document/8025911

[10] Z. Kastrati, A. S. Imran, S. M. Daudpota, M. A. Memon, and M. Kastrati, "Soaring Energy Prices: Understanding Public Engagement on Twitter Using Sentiment Analysis and Topic Modeling With Transformers," *IEEE Access*, vol. 11, pp. 26541–26553, 2023, doi: https://doi.org/10.1109/access.2023.3257283.

[11] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *arXiv:1703.04009 [cs]*, Mar. 2017, Available: https://arxiv.org/abs/1703.04009