SOGIE Bill Discourse Analysis Using Latent Dirichlet Allocation, Nonnegative Matrix Factorization, BERTopic, and Latent Semantic Analysis

Ernest Joseph Curativo
Department of Computer, Information Sciences, and Mathematics
University of San Carlos
Cebu City, Philippines
ernestcurativo@gmail.com

Neil Christian Sagun
Department of Computer, Information Sciences, and Mathematics
University of San Carlos
Cebu City, Philippines
neilsagun7@gmail.com

Angie Ceniza-Canillo
Department of Computer, Information Sciences, and Mathematics
University of San Carlos
Cebu City, Philippines
amceniza@usc.edu.ph

***Abstract*—The Sexual Orientation, Gender Identity, and Expression (SOGIE) Bill is a proposed law seeking to address discrimination based on an individual's SOGIE. The SOGIE Bill has sparked widespread debates on social media, leading to diverse public sentiments, stances, and opinions expressed through textual data. Since there are virtually no searchable papers about applying machine learning (ML) or natural language processing (NLP) techniques in SOGIE Bill-related documents in the current literature, the researchers intend to address this research gap by fulfilling their objective: using sentiment analysis and topic modeling techniques to analyze public discourse around the SOGIE Bill. Sentiment analysis identifies the overall sentiment of a text (negative, positive, or neutral), while topic modeling uncovers the themes discussed in the text. To achieve the main objective, the researchers took the following steps: (a) utilized the unsupervised ML model RoBERTa to assign sentiment scores to the entire corpus, dividing it into 4 subcorpora according to the overall sentiments of the texts (negative, positive, neutral, and other); (b) created and trained Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), BERTopic, and Latent Semantic Analysis (LSA) topic models for each subcorpus; (c) evaluated the performance of the topic models using CV coherence, UMass coherence, and perplexity scores; (d) determined the best-performing topic models based on their scores and extracted their generated topics (each represented by the top 30 words); (e) deduced the general topics of the best-performing models based on the top words; and (f) validated the general topics deduced by the researchers with a lawyer to ensure their correctness and the absence of bias. Results showed that LDA consistently excelled in perplexity across the four subcorpora, NMF and LSA equally excelled in CV coherence, and BERTopic excelled in UMass coherence but with fewer and more repetitive topics. The outputs of this study, which are the general topics representing the sentiments and stances of Filipinos about the SOGIE Bill, will help lawmakers revise the bill and foster an understanding of the different perspectives on the bill.**

Keywords—Sentiment Analysis, Topic Modeling, Natural Language Processing (NLP), Machine Learning (ML), Sexual Orientation and Gender Identity/Expression

I. Introduction

The proposed SOGIE Bill in the Philippines has been pending for over two decades in both the House of Representatives and the Senate. In those decades, the bill has caused numerous nationwide debates and received backlash, praise, and revisions. The bill was created as a response to the discrimination experienced by the LGBTQ+ (Lesbian, Gay, Bisexual, Transgender, and Queer) community in the country. To fight against SOGIE-based discrimination, the bill claims to protect all kinds of people regardless of their SOGIE, including cisgender and heterosexual (straight) people. However, many Filipinos doubt this claim and believe that the bill only gives special privileges to the LGBTQ+ community while giving unfair treatment to cisgender and straight people in the process [1]. While there are some Filipinos who approve of the bill, some continue to have strong convictions against it, claiming that it is a threat to freedom of speech, religion, family values, conventional marriages, young people, etc. [2][3][4]. Lawmakers must understand these different public perspectives to revise and adjust the bill accordingly so that it truly benefits all of the Filipino people. These polarizing opinions are often expressed online in social media posts. However, due to their sheer number, it is difficult to manually read all of them and analyze people's sentiments and stances about the bill. To address this problem, the researchers aim to use NLP and ML techniques to automate the identification of people's sentiments and stances about the bill. This allows the researchers to quickly perform SOGIE Bill public discourse analysis, which is the objective of this study. This study presents the novelty of addressing the research gaps from the lack of ML and NLP papers about the SOGIE Bill by combining sentiment analysis (RoBERTa) with topic modeling (LDA, NMF, LSA, and BERTopic) techniques and applying them to SOGIE Bill-related texts.

Sentiment analysis allows the researchers to identify the overall tone of a text (positive, negative, or neutral) and group the texts according to their sentiment. Topic modeling allows the researchers to identify the topics, including the stances, expressed in the texts belonging to a sentiment group. The purpose of augmenting sentiment analysis with topic modeling is to determine what stances are expressed in every sentiment, which the researchers can use to conclude the correlation between a person's sentiment and stance about the SOGIE Bill. Furthermore, combining sentiment analysis with topic modeling not only gives the researchers the surface-level sentiments of the texts but also their actual content and stances, which are valuable insights for public discourse analysis. Therefore, the output of this research is a summary of general topics and stances discussed in SOGIE Bill-related texts per sentiment. The researchers hope that it can help with the lawmakers' problem of properly revising the bill.

Due to the large volume of text data, the researchers acknowledge the following limitations: a) potential bias in the texts, inaccuracies in the automation of Filipino-to-English translations, and irrelevant documents (tests not related to the SOGIE Bill). The researchers do not account for them due to the impracticality of manually validating a corpus containing over 45,000 texts. The researchers also disregard sarcasm in the texts as detecting it requires complex linguistic and contextual analysis. Translation errors may impact the study's accuracy, while irrelevant data are likely classified as outliers by the topic models or discarded during manual evaluation by a lawyer.

Lastly, the test scenarios of this study include the comparison of the LDA, NMF, LSA, and BERTopic models in the CV, UMass, and Perplexity metrics, validation of the researchers' deduced general topics by a lawyer, and correlation between Filipinos' sentiments and stances about the SOGIE Bill.

II. RELATED LITERATURE

A. RoBERTa

RoBERTa (Robustly Optimized BERT Approach) is an extension of Bidirectional Encoder Representation from Transformers (BERT), both falling under the Transformers family that was developed to address the issue of long-range dependencies problem in sequence-to-sequence modeling [5]. The model is trained on vast amounts of data, allowing it to capture contextual information and understand the nuanced relationships between words, even slang in a sentence, whether formal or informal [6]. The model is most crucially adjusted for sentiment analysis using the TweetEval benchmark [7], which is consistent with the researchers' goal of doing a sentiment analysis. RoBERTa excels in performance across various NLP tasks, thanks to its training on a larger corpus and improved training process, resulting in better fine-tuning and interpretability [8]. However, it has limitations, including a reduced ability for zero-shot learning, high computational costs for fine-tuning, and susceptibility to biases or misinformation from its training data.

B. Latent Dirichlet Allocation (LDA)

LDA works by assuming that each document is composed of a blend of topics, represented as probability distributions over words, and aims to reveal these hidden topics and their distribution across the corpus using Bayesian inference, thereby facilitating the extraction of meaningful themes from the text data [9]. LDA is advantageous for its simplicity and efficiency, particularly with longer texts where different topics are covered in separate sections [10]. However, it has notable limitations: it requires the number of topics to be predefined, disregards sentence structure and semantics, and struggles with contextual concepts like irony. Additionally, it treats related topics as entirely separate, missing correlations that might be important in understanding the text.

C. Nonnegative Matrix Factorization (NMF)

Nonnegative Matrix Factorization is an unsupervised learning algorithm used to "factorize a nonnegative matrix, X, into the product of two lower rank matrixes, A and B, such that AB approximates an optimal solution of X" [11]. Simply put, it decomposes a given nonnegative matrix into two nonnegative matrices whose product approximates the original matrix. NMF effectively reveals latent structures in high-dimensional, sparse, or noisy data and handles missing values and outliers well. However, it faces challenges in determining the optimal number of clusters and can be sensitive to initial conditions, potentially leading to suboptimal solutions [12].

D. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) takes a matrix of documents and terms and decomposes it into two separate matrices (document-topic matrix and topic-term matrix) using Singular Value Decomposition (SVD) [13]. In high-dimensional text data, LSA is useful for retrieving documents, data extraction, and clustering. LSA is effective at reducing noise and capturing semantic structures by representing documents in a lower-dimensional space, improving retrieval by reflecting term correlations. Its disadvantage, however, is that it requires substantial storage due to large data sizes and can be inefficient for processing long documents because of the computational demands and loss of sparsity [14].

E. BERTopic

BERTopic is a state-of-the-art Python library designed to streamline the topic modeling method by utilizing diverse embedding techniques and c-TF-IDF to create dense clusters, enabling easy interpretation of topics while retaining important words within the topic descriptions [15]. The model offers some advantages, including a consistently strong performance across a variety of language models, flexibility as it separates the process of embedding documents from representing topics, and the ability to model the dynamic and evolutionary aspects of topics using class-based versions of TF-IDF [16]. Its disadvantage is that it often produces too many outliers and topics, requiring labor-intensive inspection [17].

F. Previous Research

The current literature on NLP and ML techniques applied to texts related to the SOGIE Bill is limited, with the only searchable work providing charts that depict the growth of anti-SOGIE networks, thematic clusters of anti-SOGIE narratives, and viral disinformation [18]. However, the methods used to generate these charts are not specified. While there is a notable research gap in the application of sentiment analysis or topic modeling specifically to the SOGIE Bill issue, similar techniques have been successfully employed in other contexts. For instance, the RoBERTa model combined with ABSA achieved 94.7% accuracy in sentiment analysis of the Russia-Ukraine crisis [19]. LDA has been used to analyze comments on Tiket.com, identifying promo discounts as a popular topic and reservation issues as a source of negative feedback [20]. The NMF model outperformed other methods in analyzing Urdu social media texts [21], and BERTopic demonstrated a 34.2% improvement over LDA and Top2Vec in topic clustering on Weibo and Twitter [22]. Additionally, LSA, when enhanced with entropy, provided better classification accuracy than TF-IDF, proving effective for large text corpora [23]. These findings underscore the effectiveness of RoBERTa, LDA, NMF, BERTopic, and LSA techniques in text analysis and topic generation. Building on these successful approaches, this research aims to apply these methods to the unexplored area of the SOGIE Bill, addressing the existing research gap and contributing new insights to the fields of NLP and ML.

III. METHODS

A. *Sources of Data*

The sources of data came from six social media platforms, namely, Reddit, TikTok, YouTube, Instagram, X (formerly Twitter), and Facebook. The duration or time span of the data is 2019-2024. Moreover, the researchers only scraped the comments, posts (in texts), and replies as they are text data, not the pictures, videos, or any kind of media. The documents were scraped with the following keywords and hashtags: "sogie", "sogie bill", "sogiebill", "#sogie", "#sogiebill", "sogie 2019", "sogie 2020", "sogie 2021", "sogie 2022", "sogie 2023", and "sogie 2024".

B. *Gathering of Data*

To collect the data from Reddit, the researchers utilized the PRAW (Python Reddit API Wrapper) library to interact with the Reddit API. For the data scraped from Facebook, TikTok, YouTube, Instagram, and X, the researchers utilized the Facebook Comments Scraper, TikTok Comments Scraper, Youtube Comments Scraper, Instagram Posts/Comments Scraper, and Twitter Scraper, respectively, which are all available on the Apify Store. The compiled data consisted of 32,693 texts from Facebook, 5,703 from YouTube, 3,043 from TikTok, 2,685 from X (formerly Twitter), 2,589 from Reddit, and 1,302 from Instagram. After removing duplicates, the total count of texts across all platforms is 45,458.

C. *Preprocessing*

During the sentiment analysis phase, The researchers removed noise from the text data by removing emojis, mentions, hashtags, URLs, special and newline characters, extra spaces, and empty texts. This preprocessing was necessary to avoid potential lexical errors. In the topic modeling phase, the researchers performed the following preprocessing steps:

- **Lemmatization:** Reducing words to their most basic or canonical form (lemma) is known as lemmatization. For instance, "mice" becomes "mouse", "running" becomes "run", etc.

- **Stop Words Removal:** Removing common words that occur frequently in sentences but often carry little semantic meaning, such as "and", "the", "is", etc.

- **Lowercasing:** Lowercasing involves converting all words to lowercase, ensuring consistency in the representation of words and preventing the duplication of tokens due to case differences (e.g. "SOGIE" and "sogie").

- **Tokenization:** Tokenization is the process of dividing a text document into discrete words, or tokens, to provide the basic analytical units for the models. The statement "I love myself" is tokenized, breaking down into the words "I", "love", and "myself".

- **N-grams Creation:** N-grams are sequences of N contiguous tokens (words or characters) from a given text. Common examples of them in this paper are bi-grams like *sexual_orientation*, *gender_identity*, etc.

- **Term Frequency-Inverse Document Frequency (TF-IDF) Removal:** TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. The purpose of employing this preprocessing step is to remove words that are too common in each document to the point that they hold invaluable meaning.

D. *Analysis and Design*

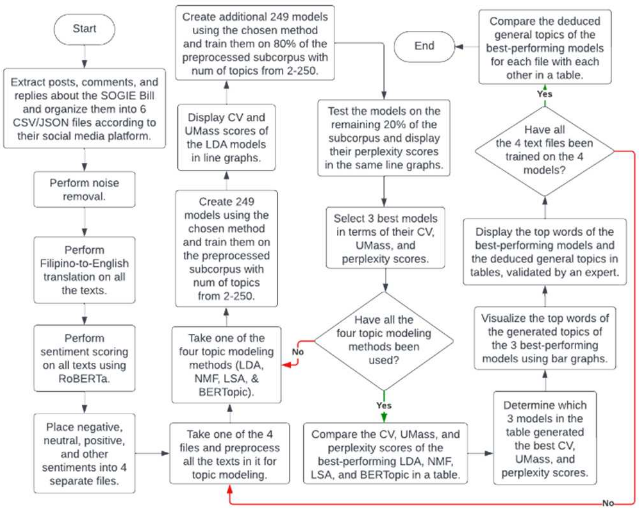
Start
Extract posts, comments, and replies about the SOGIE Bill and organize them into 6 CSV/JSON files according to their social media platform.
Perform noise removal.
Perform Filipino-to-English translation on all the texts.
Perform sentiment scoring on all texts using RoBERTa.
Place negative, neutral, positive, and other sentiments into 4 separate files.
Take one of the 4 files and preprocess all the texts in it for topic modeling.
Take one of the four topic modeling methods (LDA, NMF, LSA, & BERTopic).
Create 249 models using the chosen method and train them on the preprocessed subcorpus with num of topics from 2-250.
Display CV and UMass scores of the LDA models in line graphs.
Create additional 249 models using the chosen method and train them on 80% of the preprocessed subcorpus with num of topics from 2-250.
Test the models on the remaining 20% of the subcorpus and display their perplexity scores in the same line graphs.
Select 3 best models in terms of their CV, UMass, and perplexity scores.
Have all the four topic modeling methods been used?
No
Yes
Compare the CV, UMass, and perplexity scores of the best-performing LDA, NMF, LSA, and BERTopic in a table.
Determine which 3 models in the table generated the best CV, UMass, and perplexity scores.
Visualize the top words of the generated topics of the 3 best-performing models using bar graphs.
Display the top words of the best-performing models and the deduced general topics in tables, validated by an expert.
Have all the 4 text files been trained on the 4 models?
Yes
No
Compare the deduced general topics of the best-performing models for each file with each other in a table.
End

Fig. 1. Process Flow

Figure 1 outlines the researchers' process, starting with collecting textual data from the six social media platforms. They first removed noise from the data (discussed further in Subsection C) and used the Googletrans library to translate Filipino words into English, leaving English words unchanged. They then used a modified RoBERTa model to assign sentiment scores (positive, neutral, negative, and *other*) to each text. Texts were classified as positive, negative, or neutral if their respective sentiment score exceeded 0.5. Texts not meeting this threshold or containing 2000+ characters were categorized as *other*. This resulted in four subcorpora: negative (17,257 texts), positive (6,103 texts), neutral (19,307 texts), and *other* (2,791 texts).

For every subcorpus (s) from the four subcorpora (negative, positive, neutral, *other*), the researchers took the following steps:

1) Performed the preprocessing steps detailed in subsection C.

2) For every topic modeling technique (t) from the four techniques (LDA, NMF, LSA, and BERTopic), the researchers performed the following steps:

a) Created 249 models using the technique, with the number of topics (k) ranging from 2 to 250 topics; trained them on 100% of the subcorpus to evaluate CV and UMass coherence scores. Training the models on 100% of the subcorpus does not cause overfitting because the CV and UMass coherence scores are designed to measure the quality of the topics generated by the model based on the data it was trained on, rather than on held-out data (which is the case for perplexity).

b) Created another 249 models using the same technique (with the same k range), trained them on 80% of the subcorpus, and tested them on the remaining 20% to calculate perplexity scores.

c) Selected the three best-performing models in terms of CV, UMass, and perplexity scores for the current topic modeling technique (*t*).

3) Selected the overall three best-performing models (among all the techniques) in terms of CV, UMass, and perplexity scores for the current subcorpus (*s*).

The researchers then extracted the topics (sets of top 30 words) of the best-performing models per subcorpus (*s*) and deduced general topics from them, validated by a lawyer. Finally, the researchers concluded by analyzing the models' performance, comparing the generated topics, and examining the correlation between sentiment and stance on the SOGIE Bill.

E. Evaluation

- **CV Coherence Score:** The coherence score measures the interpretability of topics by evaluating the degree of semantic similarity between words within topics, ensuring that the words in each topic are meaningful and related [24]. The higher the CV score, the better the model is performing [25]. This equation is derived from [26].

$$c_v = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N} s\cos\left(\overrightarrow{W_{n,k}},\overrightarrow{W_k^*}\right)}{N\ X\ K} \qquad (1)$$

where K is the number of topics, N is the number of top words per topic, s is the sliding window size, $\overrightarrow{W_{n,k}}$ is a vector to represent topic word *w* at index *n* in topic *k*, where $\overrightarrow{W_{n,k}}$ = *N*, and $\overrightarrow{W_k^*}$ is a topic vector.

- **UMass Coherence Score:** UMass coherence is based on the aggregation of word probabilities derived from document frequencies of the original documents used for learning the topics. This metric considers the ordering among the top words within a topic and utilizes a smoothed conditional probability approach to assess the coherence between word pairs, with a summation procedure that emphasizes the relationship among adjacent words in the topic [27]. The closer the UMass score is to 0, the better [25]. The equation referenced here is from [27].

$$c_{UMass} = \frac{2}{N\,.\,(N-1)} \sum_{i=1}^{N}\sum_{j=1,j\neq 1}^{N} \log\frac{P(w_i,w_j)+\in}{P(w_j)} \qquad (2)$$

where N is the number of top words per topic, w is a word in the corpus, *i* and *j* are indices, and ∈ is a small constant used to avoid a log of 0.

- **Perplexity Score:** This assesses how well the model performs on new unseen data [24]. Training and testing sets of the dataset are separated. After the model is trained on the training set, it is tested on the testing set by asking it to forecast the probability distribution of words in the testing set or in documents that have not yet been viewed. This is done using the model's acquired word distributions and topic structure from training. The lower the perplexity score, the better the model is performing [28]. This equation is adapted from [29].

$$per(D_{test}) = \exp\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d} \qquad (3)$$

where D_{test} is the testing document, *M* is the number of documents, w_d is a word in document *d*, and N_d is the number of words in document *d*.

- **Human Expert Validation:** Lastly, a lawyer validates the results of the study, which are the general topics deduced by the researchers from the top keywords generated by the best-performing models, to ensure the interpretation of the researchers is valid, correct, and not biased.

IV. Results

For each subcorpus, the researchers first created and trained 249 models each of LDA, NMF, BERTopic, and LSA, with the number of topics (*k*) ranging from 2 to 250, and evaluated these models using the CV (1) and UMass (2) formulas. To assess perplexity, they then created and trained an additional 249 models for each technique with the same *k* values, evaluating these using the perplexity formula (3).

Afterward, the researchers selected the three best models from each technique (LDA, NMF, BERTopic, and LSA) based on their CV (1), UMass (2), and perplexity (3) scores. From these, they identified the three overall best models for each subcorpus, considering their performance across all three metrics. The table below shows the overall best models per subcorpus.

TABLE I. BEST-PERFORMING LDA, NMF, LSA, AND BERTOPIC MODELS IN NEGATIVE SUBCORPUS

Subcorpus	**CV (1)**	**UMass (2)**	**Perplexity (3)**
Negative	0.74253 (NMF *k* = 6)	-1.74957 (BERTopic *k* = 2)	0.84701 (LDA *k* = 3)
Positive	0.78326 (LSA *k* = 2)	-2.56898 (BERTopic *k* = 3)	0.80840 (LDA *k* = 2)
Neutral	0.83425 (NMF *k* = 3)	-2.08333 (BERTopic *k* = 2)	0.82830 (LDA *k* = 4)
Other	0.63465 (LSA *k* = 2)	-1.59872 (LSA *k* = 2)	0.89494 (LDA *k* = 5)

Table I shows the top-performing topic models for each sentiment subcorpus: NMF (6 topics) excelled in CV coherence for negative sentiment; BERTopic (2 topics) led in UMass coherence; and LDA (3 topics) had the lowest perplexity. For positive sentiment, LSA (2 topics) was best in CV coherence, BERTopic (3 topics) in UMass coherence, and LDA (2 topics) in perplexity. In the neutral subcorpus, NMF (3 topics) excelled in CV coherence, BERTopic (2 topics) in UMass coherence, and LDA (4 topics) in perplexity. For *other* sentiment, LSA (2 topics) was top in both CV and UMass coherence, and LDA (5 topics) had the lowest perplexity.

The outputs of these best-performing models were the topics they discovered in each subcorpus. Each topic was represented by a set of top 30 words, which means the discovered topics were not the actual topics per se but just representations of them. The researchers were the ones deducing topics from each set of top 30 words, which were then validated by a lawyer for accuracy.

V. Discussion

This section outlines the top 30 words generated by the best models in each subcorpus (as shown in Table I) and presents the researchers' general topics deduced from their top 30 words, which have been validated by a lawyer.

TABLE II. TOP WORDS AND GENERAL TOPICS IN NEGATIVE SUBCORPUS

NMF topic 0 **Top Words:** gay, man, make, know, sin, person, people, many, one, go, world, lesbian, problem, bible, homosexual, straight, bear, give, love, accept, become, hate, choose, life, thing, create, great, see, reason, look **Deduced General Topic:** Perceptions of the Bible on homosexuality; choosing to be your desired identity and its consequences
NMF topic 1 **Top Words:** woman, man, right, want, sex, real, child, tran, get, thing, gender, people, use, transwoman, bathroom, lgbtq, make, respect, feel, transgender, see, enter, give, abuse, male, bear, happen, wear, call, fellow **Deduced General Topic:** Issues revolving transgender women using certain bathrooms; man and woman are the only genders
NMF topic 2 **Top Words:** people, say, gender, think, get, lgbtq, use, know, accept, many, church, discrimination, religion, straight, person, mean, believe, religious, bad, feel, life, bible, wrong, lgbt, thing, discriminate, see, issue, call, need **Deduced General Topic:** Negative attitudes of the church and religion towards the LGBTQ+ community
NMF topic 3 **Top Words:** bill, sogie, right, people, law, discrimination, pass, country, lgbtq, philippine, gender, community, make, need, many, discriminate, give, protect, read, straight, equality, let, religious, include, big, problem, anti, human, base, filipino **Deduced General Topic:** The passage of the SOGIE Bill in the country and the protection and equality it claims to give
NMF topic 4 **Top Words:** right, say, wrong, law, man, human, sex, use, homosexuality, bible, make, act, bad, thing, country, word, sin, give, let, homosexual, happen, religion, stop, many, argument, truth, point, change, get, equal **Deduced General Topic:** Religious condemnation of homosexuality
NMF topic 5 **Top Words:** want, law, know, go, think, sex, gender, respect, get, one, discrimination, pass, make, many, country, world, see, come, equality, call, thing, marriage, hate, need, issue, stupid, change, first, let, accept **Deduced General Topic:** LGBTQ+ discrimination; SOGIE Bill's aim for equality and effects on marriage
BERTopic topic 0 **Top Words:** bill, people, sogie, sogie bill, man, say, woman, right, law, want, know, gay, philippine, make, go, get, problem, stupid, farmer, many, think, sex, big, use, country, one, ugly, pass, discrimination, bad **Deduced General Topic:** Create laws to support farmers instead of the SOGIE Bill
LDA topic 0 **Top Words:** problem, stupid, philippine, many, make, country, farmer, say, ugly, time, issue, people, first, focus, vote, useless, poor, law, go, get, waste, senator, song, government, money, know, stop, bad, good, pay_attention **Deduced General Topic:** The Philippines has more other problems to solve (e.g. problems faced by farmers) instead of focusing and wasting time on the SOGIE Bill
LDA topic 1 **Top Words:** people, right, gender, say, discrimination, think, use, want, know, gay, human, get, law, discriminate, need, lgbtq, person, religion, religious, call, belief, go, community, make, respect, understand, point, mean, straight, feel **Deduced General Topic:** Discussion on LGBTQ+ rights and discrimination; understanding and respecting points and feelings revolving around the discussion; respecting the law and religion
LDA topic 2 **Top Words:** bill, sogie, woman, man, people, want, gay, law, know, say, sex, sin, right, make, pass, go, let, child, bible, wrong, thing, get, bad, many, world, big, marriage, one, happen, create **Deduced General Topic:** The LGBTQ+ community and children; perceptions about the queer community according to the Bible; God only created man and woman

Table II presents the top 30 words per topic from the best-performing models in the negative subcorpus. The common theme revolves around the polarized discourse on the SOGIE Bill, with most leaning towards negative views. NMF topics 2, 4, and LDA topic 2 highlight the strong association between the LGBTQ+ community and negative terms like "sin", "bad", and "church", reflecting the negative perception of Christianity in the Philippines. NMF topic 0 discusses the word "choose" and various identities, linking it to debates on whether being LGBTQ+ is a choice, with some arguing it leads to divine punishment. NMF topics 3 and 5 cover discussions on discrimination, equality, and the bill, with concerns about special treatment rather than true equality. The word "marriage" in NMF topic 5 brings up debates on same-sex marriage and traditional views on marriage between a man and a woman, while LDA topic 2 connects "man", "woman", and "create" to arguments that God created only man and woman. The word "law", appearing in multiple topics, is linked to discussions on prioritizing other laws (like those supporting farmers) over the SOGIE Bill, as shown in BERTopic and LDA topic 0. LDA topic 1 emphasizes understanding and respecting diverse viewpoints, including religious perspectives, while addressing discrimination against the LGBTQ+ community. NMF topic 1 focuses on the controversy surrounding transgender women using female bathrooms, with words like "male", "feel", and "abuse" reflecting concerns about gender identity and perceived misuse of the bill. Overall, while most sentiments in the negative subcorpus are against the bill, the topics often overlap, and some topics were challenging to deduce due to vague or general word sets.

TABLE III. TOP WORDS AND GENERAL TOPICS IN POSITIVE SUBCORPUS

LSA topic 0 **Top Words:** love, know, right, beautiful, good, thank, see, bill, people, make, sogie, want, get, let, else, world, say, community, life, go, hope, believe, respect, live, pride, fight, give, happy, lgbtq, understand **Deduced General Topic:** Gratitude and love towards the SOGIE Bill
LSA topic 1 **Top Words:** know, love, beautiful, else, see, bill, sogie, get, thank, good, desperately, overwhelmed_smile, way_flip_hair, ground_hard_tell, look, light, people, community, world, pride, make, baby, harry, let, understand, believe, say, lgbtq, fight, hope **Deduced General Topic:** Positive expressions and general optimism for the LGBTQ+ community and the fight for their rights; spam lyrics from "You Don't Know You're Beautiful" by One Direction
BERTopic topic 0 **Top Words:** love, good, bill, right, thank, hope, sogie, people, let, go, sogie bill, say, know, community, make, life, well, woman, get, want, equality, respect, fight, time, see, give, man, bless, gender, happy **Deduced General Topic:** Praises and gratitude towards Senator Nancy Binay for speaking out her opinion against the SOGIE Bill
BERTopic topic 1 **Top Words:** contact, opportunity, link, peso, thank, deposit, duplon, earn, get, mrs, mrs lieze, lieze, lieze duplon, great, connect, profit, financially, click link, click, help, good news, great opportunity, peso get, hour thank, financially truly, philliphine, send promise, philliphine deposit, contact thank, later connect **Deduced General Topic:** Spam message thanking Mrs. Lieze Duplon for helping them financially
LDA topic 0

Top Words: happy, bless, hope, true, point, straight, feel, okay, correct, people, watch, use, funny, join, take, brother, see, favor, bad, good, laugh, queen, kind, video, boy, become, write, luck, end, wish **Deduced General Topic:** Positive expressions for someone expressing their opinion in a video
LDA topic 1 **Top Words:** make, good, let, truth, bro, sense, bless, know, glory, real, read, perfect, look, word, time, respi, keep, long, catholic, actually, change, pray, share, fine, special, want, think, rest_peace, talk, continue **Deduced General Topic:** Thanking someone for speaking out their "truth" against the SOGIE Bill
LDA topic 2 **Top Words:** go, first, good, clean, win, girl, back, country, philippine, image, many, help, together, aman, problem, important, time, ahead, create, ma_am, nice, fun, drag, rap, church, use, worry, welcome, powerful, let **Deduced General Topic:** Humans are created with the image of God; praises for drag queens for expressing their support for the SOGIE Bill; praises for a rapper for making a rap about preaching and repenting to God
LDA topic 3 **Top Words:** love, right, people, life, respect, holy, fight, live, accept, give, world, know, long, want, happy, let, human, lgbtq, say, praise, sin, come, one, believe, learn, much, gender, part, brother, judge **Deduced General Topic:** Acceptance and respect for LGBTQ+ individuals
LDA topic 4 **Top Words:** good, say, thank, well, man, great, binay, get, woman, right, finally, nice, song, madam, like, senator, news, gay, brain, job, comment, point, link, want, farmer, mrs_lieze, god_send_promise_share, opportunity_phillliphine_deposit_peso, philippine_singapore_contact, thank_later_connect **Deduced General Topic:** Praising Senator Binay for her opinion and another spam message thanking Lieze Duplon
LDA topic 5 **Top Words:** thank, bill, sogie, bless, much, risa, fight, good, right, vote, salute, people, hontivero, senator, hope, country, explain, continue, pastor, stand, song, view, philippine, big, educate, pay, mute, next, group, help **Deduced General Topic:** Gratitude towards Senator Risa Hontiveros for her support for the SOGIE Bill; thanking pastors for expressing their opinions against the bill
LDA topic 6 **Top Words:** bill, support, sogie, agree, community, pride, equality, lgbtq, month, flag, law, pass, time, right, let, make, hope, well, senator, proud, celebrate, discrimination, first, friend, protect, recognize, member, color, smart, last **Deduced General Topic:** Pride month celebration and support; Senator Binay and her "color" argument against the SOGIE Bill
LDA topic 7 **Top Words:** pray, let, go, know, keep, see, work, come, answer, beautiful, get, proud, bless, take, home, guy, leave, marry, make, hope, brave, care, need, well, else, opportunity, mind, talent, treatment, awesome **Deduced General Topic:** Providing support and positive reinforcement to the LGBTQ+ community; brave people expressing their opinions on the SOGIE Bill

Table III summarizes the top 30 words per topic from the best-performing models in the positive subcorpus. The topics generally reflect gratitude towards the SOGIE Bill, supporters, and critics. LSA topics 0 and 1 feature positive words like "beautiful", "good", and "thank", though "beautiful" was used sarcastically in some contexts. BERTopic topic 0 and LDA topic 4 show that words like "good" and "thank" were used to praise those who criticized the bill. LDA topics 0, 1, and 5 reveal praise and blessings for those attacking the bill, with LDA topic 5 also showing support for Senator Risa Hontiveros' advocacy for the bill, despite some misspellings. LDA topic 2 includes distinct topics: "image" relates to creations in God's image, "drag" to drag queens and SOGIE advocacy, and "rap" to a religious rap attacking the bill. LDA topic 6 celebrates LGBTQ+ pride and includes references (particularly, "discrimination", "protect", and "color") to Senator Nancy Binay's past comments about the bill. The senator exclaimed that she gets discriminated against for her skin color, yet she does not feel the need to file a bill to protect people like her. LDA topic 7 praises bravery of those expressing opinions about the bill. LSA topic 1 contains spam lyrics from One Direction and references to Harry Styles, while BERTopic topic 1 and LDA topic 4 detect spam thanking individuals for financial help, with consistent misspellings and suspicious links. Overall, the positive subcorpus topics, though mostly positive, reflect both pro- and anti-SOGIE sentiments, and the researchers found it challenging to separate some topics due to vagueness, with multiple topics sometimes emerging from single models.

TABLE IV. TOP WORDS AND GENERAL TOPICS IN NEUTRAL SUBCORPUS

NMF topic 0 **Top words:** people, woman, say, man, know, right, want, respect, let, make, love, go, word, gay, think, accept, give, way, one, life, wrong, person, truth, live, bible, understand, believe, sin, many, change **Deduced General Topic:** The truth of what it means to be LGBTQ+ according to the Bible; LGBTQ+ acceptance and respect
NMF topic 1 **Top words:** bill, sogie, right, woman, equality, discrimination, pass, lgbtq, man, equal, protect, human, straight, let, give, include, need, philippine, community, people, transwoman, know, individual, identity_expression, read, gay, first, member, fight, senator **Deduced General Topic:** Protection against discrimination; gender identity and expression, including straight and trans people
NMF topic 2 **Top words:** law, gender, male, female, sex, bill, need, discrimination, country, make, marriage, go, philippine, pass, identity, get, equality, want, read, mean, church, religion, think, community, time, human, issue, change, biological, study **Deduced General Topic:** Effects of the passage of the bill on marriage; discussions on Biology, transgender people, and changing identities
BERTopic topic 0 **Top Words:** bill, sogie, right, sogie bill, say, law, people, man, know, go, woman, gender, make, want, let, pass, read, point, discrimination, think, first, come, gay, get, need, philippine, respect, bible, one, church **Deduced General Topic:** Read the provisions of the SOGIE Bill first
LDA topic 0 **Top Words:** marriage, sex, family, mother, sister, church, think, post, religion, brain, look, go, get, tell, separate, ready, hit, catholic, father, bring, culture, couple, say, pregnant, song, priority, year, fix, jessy, union **Deduced General Topic:** Separation between church and state; Filipino and Western culture; prioritizing the SOGIE Bill among others; marriage and family dynamics
LDA topic 1 **Top Words:** bill, sogie, right, law, people, need, pass, discrimination, philippine, equality, respect, lgbtq, protect, want, fight, community, country, read, religious, let, say, senator, think, gender, pray, make, human, know, straight, equal **Deduced General Topic:** Fighting for equality and respect
LDA topic 2 **Top Words:** woman, man, know, people, say, want, gender, gay, go, right, male, make, respect, bible, change, female, person, give, let, accept, love, word, world, create, truth, sex, wrong, one, life, think **Deduced General Topic:** Changing the truth; changing gender from woman to man and vice versa; respecting the LGBTQ+ community or the Bible
LDA topic 3 **Top Words:** say, point, take, first, get, come, make, farmer, talk, issue, time, go, next, government, hope, use, year, day, big, let, sign, last, video, philippine, watch, country, respi, remember, correct, win

Deduced General Topic: The government focusing on the SOGIE Bill issue instead of talking about farmers

Table IV summarizes the top words per topic from the best-performing models in the neutral subcorpus. NMF topic 0 features terms like “bible”, “word”, and “truth”, reflecting anti-SOGIE Bill views based on biblical teachings. NMF topic 1 contains “include”, “protect”, and “discrimination” and the bigram “identity_expression”, most likely referring to the documents talking about how the SOGIE Bill protects everyone from discrimination regardless of their gender identity and expression, including straight (cisgender) people. NMF topic 2 covers “bill”, “marriage”, and “pass”, focusing on the bill’s impact on marriage and transgender identity. BERTopic topic 0 suggests people should read the bill’s provisions first to avoid misconceptions. LDA topic 0 addresses church-state separation, LGBTQ+ spaces, and family dynamics, with “jessy” being a misinterpreted name. LDA topic 1 advocates for LGBTQ+ equality and protection through the bill, while LDA topic 2 explores changes to gender identity and conflicting truths with the Bible. LDA topic 3 emphasizes prioritizing farmers’ needs over the SOGIE Bill. The subcorpus reveals polarized stances on the bill with less topic overlap but similar vagueness in the top words.

TABLE V. TOP WORDS AND GENERAL TOPICS IN *OTHER* SUBCORPUS

LSA topic 0 **Top words:** bill, woman, sogie, people, say, man, right, law, gender, lgbtq, sex, good, wrong, act, know, child, person, think, make, use, discrimination, love, want, sexual, sin, let, gay, school, go, accept **Deduced General Topic:** Discrimination and treatment towards the LGBTQ+ community in schools
LSA topic 1 **Top words:** act, wrong, good, bill, woman, sogie, sexual, sex, morally, think, say, man, lgbtq, make, procreation, point, contrary, engage, gender, people, argument, order_procreation, couple, respect, school, action, child, way, end, hold **Deduced General Topic:** Sexual acts ordered or aimed towards procreation are morally good; otherwise, they are not
LDA topic 0 **Top words:** law, take, right, good, vote, work, philippine, time, know, sorry, show, hope, want, first, stop, senator, family, use, jail, lose, gay, job, go, get, election, government, wake, say, call, become **Deduced General Topic:** Senators pandering to the LGBTQ+ community for votes; jail punishment under the SOGIE Bill
LDA topic 1 **Top words:** say, people, know, sin, want, make, go, love, gay, life, bible, man, sex, think, let, one, truth, believe, read, marriage, judge, understand, respect, word, come, wrong, argument, person, follow, live **Deduced General Topic:** What it means to be a homosexual according to the truth in the Bible; only God can judge people
LDA topic 2 **Top words:** good, wrong, say, act, sex, right, people, sexual, think, make, know, morally, point, way, see, respect, never, judge, procreation, come, time, bear, bad, go, love, engage, many, give, disagree, action **Deduced General Topic:** Engaging in sex can be morally right or wrong depending on the intention (e.g. procreation)
LDA topic 3 **Top words:** woman, right, man, people, gender, bill, law, want, sogie, lgbtq, let, accept, child, know, love, respect, gay, person, discrimination, think, country, school, human, thing, different, world, use, need, philippine, go **Deduced General Topic:** Discrimination and treatment towards the LGBTQ+ community in schools
LDA topic 4 **Top words:** bill, sogie, people, say, good, many, get, law, right, farmer, big, know, make, well, need, give, use, discrimination, thank, issue, man, let, hope, go, member, problem, senator, read, lgbtq, discriminate **Deduced General Topic:** Create laws for the farmers instead of the LGBTQ+ community

Table V outlines the top words per topic from models applied to the *other* subcorpus. LSA topic 0 and LDA topic 3 address “school”, “child”, and “discrimination”, reflecting issues faced by LGBTQ+ individuals in educational settings with strict dress codes. LSA topic 1 and LDA topic 2 discuss “procreation”, “wrong”, “good”, “morally”, “sexual”, and “act” and the bigram “order_procreation”, highlighting debates on whether sexual acts should aim for or ordered toward procreation and their moral implications. LDA topic 0 involves “vote”, “senator”, and “government”, with discussions on politicians pandering to the LGBTQ+ community to get their votes and people facing jail punishments under the bill, with claims of inequality especially toward the poor in its application. LDA topics 1 and 4 revisit themes from other subcorpora, including biblical judgments on LGBTQ+ existence and the neglect of farmers’ needs. The topic explores how the Bible’s perceived truth conflicts with the LGBTQ+ community’s existence and includes “judge”, reflecting debates on whether only God can judge LGBTQ+ individuals, while some religious people claim they are merely stating biblical facts. This subcorpus includes documents that could not be classified by the sentiment analysis model, resulting in diverse sentiments and stances, and introduces unique topics like school discrimination and procreation debates.

The lawyer has validated the researchers’ interpretations of the top words and marked them as accurate. The lawyer only has problems with the misspellings in the top words and one top word that does not belong in the topic, which reflects the model’s performance and is beyond the researchers’ control.

VI. CONCLUSIONS

In this study, the researchers proposed a method combining sentiment analysis using a modified RoBERTa model and topic modeling using LDA, NMF, LSA, and BERTopic models to analyze public discourse on the SOGIE Bill.

The researchers observed that positive sentiments did not necessarily imply support for the bill, as some positive texts expressed gratitude towards those opposing the bill, which was reflected in some of the topics in the positive subcorpus. Negative sentiments were often associated with opposition to the bill, with few exceptions. The neutral and *other* subcorpus also predominantly contained negative stances despite having neutral or unclassified sentiments. This suggested a general conservatism among Filipinos, with a notable minority supporting the bill.

As for the performance of the topic models, the researchers found that the LDA models consistently achieved the best perplexity scores across all four subcorpora. The NMF and LSA models performed equally well in terms of CV scores, with NMF excelling in the negative and neutral subcorpora and LSA in the positive and *other* subcorpora. While BERTopic often scored highest in UMass, its generated topics were less insightful, often repetitive, and sometimes incoherent. In the *other* subcorpus, LSA outperformed

BERTopic in UMass, showing its strength in identifying latent topics in long documents since the subcorpus mainly contains texts that are too long that the RoBERTa model could not assign sentiment scores to them. Additionally, LSA models were the quickest to train and evaluate among all the methods.

For future work, the researchers recommend refining the sentiment analysis model to be able to handle long documents and exploring additional topic modeling techniques (e.g. Top2Vec) to enhance the interpretability and diversity of the generated topics. Moreover, further research could involve expanding the corpus to include more social media platforms or conducting a longitudinal analysis to capture shifts in public sentiment and discourse over time.

Acknowledgment

We would like to offer our heartfelt gratitude to the following individuals for their contributions in making this research possible: To our thesis adviser, Ms. Angie M. Ceniza-Canillo, for her unwavering guidance and encouragement throughout the duration of our research which have been instrumental in the completion of this study; to Mr. Ryan Pelongco for his professional assistance in validating our research, as his expertise and meticulous attention to detail have significantly contributed to the accuracy and reliability of our findings; and to our families and friends for their endless support and understanding, which have been a source of strength and inspiration.

References

[1] Abad, M. (2022, November 12). *LGBTQ+ not asking for 'special rights' with SOGIE bill – expert*. Rappler. Retrieved August 19, 2024, from https://www.rappler.com/philippines/lgbtq-not-asking-special-rights-sogie-bill-expert/

[2] Cepeda, M. (2019, August 28). *Eddie Villanueva claims SOGIE bill 'threatens' freedoms of non-LGBTQ+*. Rappler. Retrieved August 19, 2024, from https://www.rappler.com/philippines/238780-eddie-villanueva-claims-sogie-bill-threatens-freedoms-non-lgbtq/

[3] Olaer, V. (2020). *Why Say "NO" to SOGIE Bill*. The Disciplers. Retrieved August 19, 2024, from https://thedisciplers.com/why-say-no-to-sogie-bill/

[4] Philippine Daily Inquirer. (2023, April 22). *Well-meaning but harmful | Inquirer Opinion*. Inquirer Opinion. Retrieved August 19, 2024, from https://opinion.inquirer.net/162567/well-meaning-but-harmful

[5] Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, *10*, 21517-21525.

[6] Prasanthi, K. N., Madhavi, R. E., Sabarinadh, D. N. S., & Sravani, B. (2023, April). A Novel Approach for Sentiment Analysis on social media using BERT & ROBERTA Transformer-Based Models. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)* (pp. 1-6). IEEE.

[7] Hugging Face. (2024, January 4). *cardiffnlp/twitter-roberta-base-sentiment · Hugging Face*. Hugging Face. Retrieved February 29, 2024, from https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[8] Kumar, M. (n.d.). *GPT-3, BERT, and RoBERTa | AI Model Analysis & Comparison*. CronJ. Retrieved August 19, 2024, from https://www.cronj.com/blog/gpt-3-bert-and-roberta-ai-model-analysis-comparison/

[9] Lark Editorial Team. (2023, December 25). *Latent Dirichlet Allocation Lda*. Lark. Retrieved February 29, 2024, from https://www.larksuite.com/en_us/topics/ai-glossary/latent-dirichlet-allocation-lda

[10] Data Basecamp. (2023, June 14). *What is LDA (Latent Dirichlet Allocation)?* Data Basecamp. Retrieved August 19, 2024, from https://databasecamp.de/en/statistics/lda-en

[11] Eunus, S. I. (2022, March 19). *What is Non-Negative Matrix Factorization (NMF)? | by Salman Ibne Eunus | CodeX*. Medium. Retrieved April 6, 2024, from https://medium.com/codex/what-is-non-negative-matrix-factorization-nmf-32663fb4d65

[12] Numerical Analysis. (2023, April 17). *NMF for Clustering: Pros, Cons, and How-To*. LinkedIn. Retrieved August 19, 2024, from https://www.linkedin.com/advice/0/what-advantages-disadvantages-using-nmf-clustering

[13] *Goyal, C. (2021b, June 26). Topic Modelling using LSA | Guide to Master NLP (Part 16)*. Analytics Vidhya. Retrieved April 6, 2024, from https://www.analyticsvidhya.com/blog/2021/06/part-16-step-by-step-guide-to-master-nlp-topic-modelling-using-lsa/

[14] Gandhi, Y. (2022, February 10). *What is Latent Semantic Analysis? Advantages and Disadvantages*. Analytics Steps. Retrieved August 19, 2024, from https://www.analyticssteps.com/blogs/what-latent-semantic-analysis-nlp-advantages-and-disadvantages\

[15] Strien, D. v., & Grootendorst, M. (2023, May 31). *Introducing BERTopic Integration with the Hugging Face Hub*. Hugging Face. Retrieved April 6, 2024, from https://huggingface.co/blog/bertopic

[16] Gaire, B. (2023, January 5). *Summary of BERTopic*. Medium. Retrieved April 6, 2024, from https://medium.com/@bgaire2053/summary-of-bertopic-3e3a5b07d4e1

[17] Vishwanath, Y. (2023, March 14). *BerTopic Modelling -Advanced Topic Modelling: | by Yogeshwar Vishwanath | Digital Engineering @ Centific*. Medium. Retrieved August 19, 2024, from https://medium.com/digital-engineering-centific/bertopic-modelling-advanced-topic-modelling-73af7697b7f3

[18] Hapal, D. K. (2023, February 12). *Disinformation on SOGIE Bill Spreads As Filipino Queers Face Real-World Discrimination*. Pulitzer Center. Retrieved August 19, 2024, from https://pulitzercenter.org/stories/disinformation-sogie-bill-spreads-filipino-queers-face-real-world-discrimination

[19] Sirisha, U., & Bolem, S. C. (2022). Aspect based sentiment & emotion analysis with ROBERTa, LSTM. *International Journal of Advanced Computer Science and Applications*, *13*(11).

[20] Puspita, B. H., Muhajir, M., & Aliady, H. (2020, October). Topic Modeling Using Latent Dirichlet Allocation (LDA) and Sentiment Analysis for Marketing Planning Tiket. com. In *The 2nd International Seminar on Science and Technology (ISSTEC 2019)* (pp. 16-22). Atlantis Press.

[21] Latif, S., Shafait, F., & Latif, R. (2021). Analyzing LDA and NMF topic models for Urdu tweets via automatic labeling. *IEEE Access*, *9*, 127531-127547.

[22] Gan, L., Yang, T., Huang, Y., Yang, B., Luo, Y. Y., Richard, L. W. C., & Guo, D. (2023, October). Experimental Comparison of Three Topic Modeling Methods with LDA, Top2Vec and BERTopic. In *International Symposium on Artificial Intelligence and Robotics* (pp. 376-391). Singapore: Springer Nature Singapore.

[23] Neogi, P. P. G., Das, A. K., Goswami, S., & Mustafi, J. (2020). Topic modeling for text classification. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (pp. 395-407). Springer Singapore.

[24] Vaj, T. (2023, July 31). *How to evaluate a novel topic modeling method. | by Tiya Vaj | Medium*. Tiya Vaj. Retrieved March 13, 2024, from https://vtiya.medium.com/how-to-evaluate-novel-topic-modeling-method-104ad968442

[25] Andronikou, K. (2022, October 21). *Topic Modeling with BERTopic*. Cmotions. Retrieved April 10, 2024, from https://cmotions.nl/en/topic-modeling-with-bertopic/

[26] Rijcken, E. (2023, January 16). *C_v topic coherence explained.* Medium. https://towardsdatascience.com/-topic-coherence-explained

[27] Röder , M., Both, A., & Hinneburg, A. (2015, February 2). *Exploring the Space of Topic Coherence Measures*. ACM Conferences. https://dl.acm.org/doi/10.1145/2684822.2685324

[28] Bismi, I. (2023, May 15). *Topic Modelling, LDA, NLP, HIDDEN TOPICS, COHERENCE, PERPLEXITY*. Medium. Retrieved April 12, 2024, from https://medium.com/@iqra.bismi/topic-modelling-using-lda-fe81a2a806e0

[29] Hörster, E., Lienhart, R., & Slaney, M. (2007, July 1). *Image retrieval on large-scale image databases*. ACM Conferences. https://dl.acm.org/doi/10.1145/1282280.1282283