Hot Topic Detection with Topic Modeling Methods

1st Wenyang Lyu
Auckland University of Technology
Auckland, New Zealand
sxb7657@autuni.ac.nz

2nd Henry Hu
Auckland University of Technology
Auckland, New Zealand
mvf6715@autuni.ac.nz

3rd Parma Nand
Auckland University of Technology
Auckland, New Zealand
parma.nand@aut.ac.nz

***Abstract*—In this project, we analyze a dataset consisting of 19,320 XML-formatted blog files from an anonymous blogging site. These blogs, annotated with anonymized metadata, cover a period from 2001 to 2004. Our objective is to identify the two most popular topics discussed within specific demographics: males, females, age less than 20 and over 20, students, and the general population. The aim is to provide insights that can drive the innovation of new products and services. We employ various pre-processing techniques to handle noise in the data, such as removing non-ASCII characters and unnecessary sections. For topic extraction, we utilize multiple methods including noun counting, TF-IDF, n-grams, Non-negative Matrix Factorization (NMF) with TF-IDF vectorizer, and Latent Dirichlet Allocation (LDA) with count vectorizer. The evaluation of these methods is conducted using coherence scores and human judgment through clause extraction. Additionally, visualization techniques such as termite plots and word clouds are utilized to represent the data. Our analysis aims to provide actionable insights into prevalent themes within different demographic groups.**

***Index Terms*—Topic Modeling, NMF, LDA**

CONTRIBUTIONS

This paper is written by Wenyang Lyu and Henry Hu. Most tasks were conducted collaboratively, with certain tasks being more heavily contributed to by one author.

Wenyang Lyu: Responsible for pipeline design, counting, TF-IDF and N-gram testing, and LDA and NMF topic number testing.

Henry Hu: Focused on LDA method design, extracted clause analysis, results analysis, and methods comparison.

I. INTRODUCTION

In the era of digital communication, blogs have become a significant source of personal expression and public discourse. The vast amount of text data generated through blogs provides a unique opportunity for text mining and natural language processing (NLP) to uncover underlying themes and trends. This paper focuses on analyzing a dataset of 19,320 XML-formatted blog files, collected from an anonymous blogging site, to identify the two most dominant topics discussed among different demographics.

The task involves processing and analyzing the blog dataset to determine the two most popular topics within different demographics as indicated by files name. Given the raw nature of the dataset, the initial step is to pre-process the data to handle noise, such as non-ASCII characters and unnecessary sections. The subsequent steps involve extracting metadata, segmenting the data into the specified demographics, and employing various strategies to extract and identify dominant topics.

Understanding the prevalent themes in blogs can provide valuable insights for businesses and researchers. For an innovation company, such insights can drive the development of new products and services tailored to the interests and needs of different demographic groups. The ability to analyze and visualize textual data effectively can also enhance decision-making processes and strategy formulation.

To achieve the objectives, we apply several text mining techniques. We pre-process the data to clean and normalize it then different strategies are employed for topic extraction for a better topic detection results, including:

- Counting all types of nouns
- Counting all subjects, direct objects, and prepositional objects
- TF-IDF (Term Frequency-Inverse Document Frequency)
- N-gram analysis
- Non-negative Matrix Factorization (NMF) with TF-IDF vectorizer
- Latent Dirichlet Allocation (LDA) with count vectorizer

The results produced by these different methods are further used to extract clauses and infer dominant topics. We compare the outcomes of the various methods and explain why we prefer using the Non-negative Matrix Factorization (NMF) approach.

In this paper, we provide a detailed description of our methodology, the pre-processing steps, the topic extraction techniques employed, the clauses extraction methods and topic interpretation methods, and the dominant topics of all demographics.

II. LITERATURE REVIEW

In this paper, we review various related works to understand the LDA and NMF methodologies for hot topic detection, the techniques to visualize the topics, and how to interpret the topics from extracted clauses (Table I). These related works have inspired our program design and provided insights for improving our data models. By reading these papers, we have not only gained an understanding of existing technologies and methodologies but also obtained valuable ideas on how to apply and optimize these techniques in our own project.

TABLE I
RELATED WORKS OF TOPIC MODELLING

Study	Methodologies	Key Findings and Innovations	Applications
Blei et al. [1]	LDA	Fundamental and Generative probabilistic approach	General topic detection
Rawat et al. [2]	LDA	Integration with PyLDAvis for visualization	Consumer behavior analysis
Mifrah [3]	LDA, NMF	LDA better for coherence, NMF for specificity	Research paper categorization
Yin and Yuan [4]	LDA with Word2Vec, BERT	Improved topic coherence with semantic embeddings	Blended learning research
Altarturi et al. [5]	HTML Topic Model (HTM)	Outperformed LDA in HTML web content analysis	Web content topic detection
Indra et al. [6]	Modified TF-IDF, LDA	Improved hot topic detection with temporal adjustments	Online news analysis
Röder et al. [7]	LDA, GloVe embeddings	Enhanced topic coherence and diversity	Social media analysis
Wallach and Lee [8]	NMF, LDA	Comparative study highlighting the strengths of each method	Health informatics
Chang et al. [9]	LDA, BERT	Leveraged contextual embeddings for improved topic clarity	Legal document analysis

A. Topic Modeling Methods

In the evolving field of text mining and natural language processing, Latent Dirichlet Allocation (LDA) has established itself as a foundational technique for topic modeling, enabling the discovery of latent topics within large text corpora. Blei et al. [1] introduced LDA as a generative probabilistic model, which assumes that documents are mixtures of topics and that topics are mixtures of words. This model has been widely adopted for various applications, including consumer behavior analysis, where Rawat et al. [2] successfully used LDA to analyze consumer tweets and identify key topics related to product quality, customer service, and pricing. The effectiveness of LDA in providing actionable insights for businesses underscores its utility in understanding customer preferences and tracking changes in consumer behavior over time.

Comparative studies between LDA and Non-negative Matrix Factorization (NMF) highlight the strengths and limitations of each approach. NMF, proposed by Lee and Seung [10], is known for producing distinct and interpretable topics by decomposing a document-term matrix into two lower-dimensional matrices representing topics and term distributions. In the context of research paper categorization, Mifrah [3] demonstrated that while LDA generally performs better in terms of topic coherence, NMF offers more specific topic categorization. This comparative analysis provides valuable insights into the applicability of different topic modeling techniques for various text mining tasks.

The integration of semantic information into LDA models represents a significant advancement in topic modeling. Recent innovations, such as the use of pre-trained word embeddings like Word2Vec and BERT, have enhanced the semantic representation of topics. Yin and Yuan [4] applied LDA with semantic embeddings to analyze blended learning research, identifying key trends and shifts in research focus over time. The integration of semantic embeddings with LDA improves topic coherence and relevance, offering a more nuanced understanding of complex textual data.

Specific applications of topic modeling in web content analysis have led to the development of models that consider the unique structure of HTML documents. Altarturi et al. [5] proposed the HTML Topic Model (HTM), which leverages HTML tags to better understand the structure of web pages and extract meaningful topics. HTM significantly outperforms traditional LDA in terms of topic coherence when applied to web content data, highlighting the importance of considering document structure in topic modeling. This approach demonstrates the potential for tailored models to enhance topic detection in specialized domains.

The combination of term frequency-based approaches with topic modeling has proven effective for detecting hot topics in dynamic text streams. Indra et al. [6] combined modified TF-IDF with LDA to detect hot topics in Indonesian online news, incorporating temporal and public attention adjustments. This methodology improved the accuracy of hot topic detection, demonstrating the importance of integrating contextual and temporal information in topic modeling. The ability to dynamically adjust term weights based on temporal trends and public engagement ensures that topic models remain relevant and accurate in rapidly evolving datasets.

Evaluating the accuracy and coherence of extracted topics is a crucial aspect of topic modeling, ensuring that the identified topics are meaningful and relevant. Various methods and metrics have been developed to assess the quality of topics generated by models like Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

B. Evaluation Methods

One of the most widely used methods for evaluating topic quality is through coherence metrics, which measure the semantic similarity between high-scoring words in a topic. Röder et al. [7] introduced a set of coherence metrics, including C_v, C_umass, and C_npmi, that evaluate the coherence of a set of words based on their co-occurrence in a reference corpus. These metrics have been shown to correlate well with human judgments of topic quality, making them a reliable tool for assessing topic models. The C_v metric, in particular, combines several coherence measures and has been widely adopted for its robustness.

Perplexity is another common metric used to evaluate topic models, particularly for LDA. It measures how well a probabilistic model predicts a sample of held-out data, with lower perplexity indicating better generalization. While perplexity is

useful for assessing model performance, it does not always correlate with human judgment of topic quality. Wallach et al. [8] explored the limitations of perplexity and suggested combining it with other metrics for a more comprehensive evaluation. Additionally, the log-likelihood of the data given the model parameters provides insight into the model's fit to the data, although it shares similar limitations with perplexity.

Beyond automated metrics, human judgment remains a gold standard for evaluating topic quality. Chang et al. [9] emphasized the importance of human evaluations by conducting studies where human subjects rated the coherence and interpretability of topics. They found that certain automated metrics, while useful, often miss nuances that human evaluators can detect. This has led to the development of hybrid evaluation frameworks that combine quantitative metrics with qualitative analysis.

A comprehensive evaluation methodology involves using a combination of coherence metrics, perplexity, and human judgment. The process typically includes:

- **Data Preparation and Model Training:** Train the topic model on a representative dataset.
- **Quantitative Evaluation:** Calculate coherence metrics (C_v, C_umass, C_npmi) and perplexity to assess the model's performance.
- **Human Evaluation:** Conduct surveys or studies where human subjects rate the coherence and interpretability of topics.
- **Comparative Analysis:** Compare the results from quantitative metrics and human evaluations to identify any discrepancies and improve the model.

This multi-faceted approach ensures a robust evaluation of topic models, combining the strengths of automated metrics with the nuanced insights of human judgment.

III. Methodology

At the beginning, we first processed 100 student files to quickly set up the entire pipeline of data preparation, preprocessing, topic modeling, clause extraction, and result evaluation. Most of these tasks were completed within 1 minute, which saves us a lot of time. Afterward, we analyzed topics of the entire 5120 student files with different methods, which took about 10-20 minutes per task. After analyzing all the topics and extracted clauses, we found the Non-negative Matrix Factorization (NMF) interpret the most reasonable topics. The final processing pipeline is as shown in Figure 1.

A. Dataset Preparation

The dataset, provided as a zipped file named Assignment2BlogData.zip containing 19320 files, is stored in the Google Drive folder '/content/drive/MyDrive/COMP814Data/'. As you can find from our code in 'Part 1: Dataset Preparation' (Appendix), we first mount Google Drive, then load and unzip all the files into the temporary file path. Then we define the function to process files with a filter to segment all the files based on demography: 'student', 'male', 'female', 'ageOlder', 'ageYoung', and 'everyone'. The 'ageOlder' demography is users who are older than 20 years old, while 'ageYoung' is users less or equal then 20. Then all the files are segmented into different variables named accordingly.

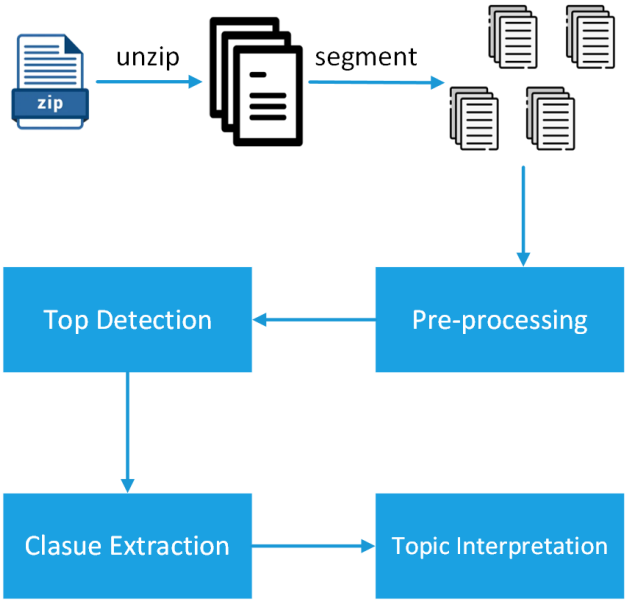

Fig. 1. Hot topic detection methodology steps

TABLE II
Data Distribution by Demography

Demography	Original Files	Processing Files
student	5120	5000
male	9660	9357
female	9660	9337
ageOlder	11080	10614
ageYoung	8240	8080
everyone	19320	18694

Some files are removed during processing because they contain more than 100,000 characters, which exceeds the model's processing ability. The original and trimmed number of files are shown in Table II.

We do not unzip and segment all files into our Google Drive for possible faster processing because we find that the total size of all folders would exceed 6GB, and copying files to different folders takes too much time.

B. Pre-processing

The pre-processing is described in 'Part 2: Pre-processing' of our code, which involves several steps to clean and prepare the text data for analysis. These steps include tokenization, removal of stopwords and punctuation, spell-checking, replacing non-ASCII characters, normalization, stemming, lemmatization, and parallel processing using Dask.

- **Tokenization and Lowercasing**: The text is first converted to lowercase and tokenized into individual words.
- **Removing Non-ASCII Characters**: Non-ASCII characters are replaced using the **unidecode** library to standard-

ize the text.

- **Stopwords and Punctuation Removal**: Stopwords, including custom stopwords relevant to the dataset, and punctuation are removed to reduce noise. The NLTK library is used to handle standard stopwords, and additional stopwords and punctuation are specified.
- **Handling Special Cases**: Months, single characters, and abnormal tokens are removed. Specific patterns are identified and excluded using regular expressions.
- **Spell Checking**: Words are checked for spelling errors using the **SpellChecker** library, and misspelled words are excluded.
- **Stemming and Lemmatization**: Words are reduced to their root forms using the Porter Stemmer and WordNet Lemmatizer from NLTK to standardize the text.
- **Sentence Splitting**: The text is split into sentences to facilitate detailed analysis at a finer granularity.
- **Parallel Processing with Dask**: To handle the large dataset efficiently, the Dask library is used to distribute the Pre-processing workload across multiple processors. Batches of files are processed in parallel to enhance performance.
- **Verification**: The file length and samples of file were printed out for verification at last.

This methodical approach ensures the text data is cleaned, standardized, and ready for subsequent analysis, leveraging parallel processing to handle large datasets efficiently.

C. Topic Detection

In this subsection we will introduce all the methods we explored for hot topic detection in code 'Part 4: Topics Detection By Different Methods'. In the results analysis section we will give the detailed analysis and comparison of all the methods.

1) Counting Method: In code 4.1 and 4.2, we have tried two different dimension of Counting method, one dimension is all nouns while the other is all subjects, direct objects, and prepositional objects, following below steps.

- **Vectorization**: A **CountVectorizer** is used to transform the text data into a matrix of token counts, focusing on the top 10,000 features.
- **Extraction**: The most frequent words in the above dimension are then identified by summing the counts across all documents.
- **Output**: The top 20 and their counts are printed, providing insights into the most common themes within the sampled text data.

2) TF-IDF Method: In code 4.3, we use the TF-IDF counting instead of term frequency only in the following steps:

- **Vectorization**: A **TfidfVectorizer** is used to transform the text data into a matrix of TF-IDF scores, focusing on the top 10,000 features.
- **Extraction**: The most frequent words are identified by summing the TF-IDF scores across all documents.
- **Output**: The top 20 words and their TF-IDF scores are printed, providing insights into the most common themes within the sampled text data.

3) nGram Method: In code 4.4, we use both biGram and triGram to extract most frequent phrase in the following steps:

- **N-Gram Generation**: The cleaned text is processed to generate n-grams (biGram and triGram). N-grams are sequences of n contiguous words from the text.
- **Frequency Counting**: The most frequent n-grams are identified by counting their occurrences across all documents. The **Counter** from the **collections** module is used to tally the n-grams.
- **Output**: The top 20 biGram and triGram, along with their counts, are printed, providing insights into the most common multi-word expressions within the sampled text data.

4) LDA with CountVectorizor: Latent Dirichlet Allocation (LDA) is a generative statistical model used to identify topics in a collection of documents. In the Gensim library we are using, LDA is implemented to extract hidden thematic structures from a large corpus. It assumes each document is a mixture of topics and each topic is a mixture of words, which enables the discovery of abstract topics within the text data. By iteratively refining topic-word and document-topic distributions, LDA provides a probabilistic framework for uncovering underlying topics. This is particularly useful for hot topic detection as it reveals the most prominent topics discussed within a set of documents.

The LDA model is based on the Dirichlet distribution and can be described by the following generative process:

$$p(\theta_d \mid \alpha) \cdot \prod_{n=1}^{N_d} \left(\sum_{k=1}^{K} p(z_{d,n} = k \mid \theta_d) \cdot p(w_{d,n} \mid z_{d,n} = k, \beta) \right) \tag{1}$$

where:

- θ_d is the topic distribution for document d.
- α is the parameter of the Dirichlet prior on the per-document topic distributions.
- N_d is the number of words in document d.
- $z_{d,n}$ is the topic assignment for the n-th word in document d.
- $w_{d,n}$ is the n-th word in document d.
- β is the parameter of the Dirichlet prior on the per-topic word distribution.

In our code 'Part 5: LDA With CountiVectorizer', the processing steps of LDA is described as below:

- **Concatenation**: All sentences within each document were concatenated into a single string.
- **Character Count Filtering**: Documents with a character count exceeding 100,000 were removed to ensure manageable processing.
- **Dictionary Creation**: A dictionary representation of the documents was created using the **gensim.corpora.Dictionary** class.
- **Filtering Extremes**: Extreme values were filtered out by specifying thresholds for document frequency to limit the number of features.

- Words appearing in fewer than 10% of the documents were excluded.
- Words appearing in more than 50% of the documents were excluded.
- The dictionary size was capped at the 100,000 most frequent words.

The LDA model is famous with its tunable parameters, which allows for greater flexibility and more accurate topics. In our code, we use the following parameters:

- **alpha='auto'**: The Dirichlet prior for the per-document topic distributions was set to 'auto' to allow the model to learn an asymmetric prior from the data.
- **eta='auto'**: The Dirichlet prior for the per-topic word distributions was set to 'auto' to enable the model to learn an asymmetric prior from the data.
- **num_topics=20**: The model was configured to extract 50 distinct topics from the corpus.
- **num_keywords=15**: For each topic, the top 15 most significant keywords were extracted.

We have tested different number of topics for a more accurate topics detection, before the human judgement, we rely on the coherence score as a metric to indicate if the topics are well classified. Coherence score is a metric used to evaluate the quality of topics generated by topic modeling algorithms like Latent Dirichlet Allocation (LDA). It measures the semantic similarity between high-scoring words in a topic. In the Gensim library, coherence score helps to determine the interpretability of topics by comparing them to a large corpus of human-annotated texts. Higher coherence scores indicate more interpretable and meaningful topics. The coherence score is typically computed using methods such as C_V, UMass, and C_W2V, which capture the degree of pairwise word co-occurrences within a sliding window or based on external reference corpora.

One common method to compute coherence is the C_V measure, which combines a sliding window, a one-set segmentation of the top words, and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. The coherence score C_V can be expressed as follows:

$$\text{Coherence}(T, V) = \frac{1}{|T|} \sum_{t \in T} \frac{1}{|W_t|(|W_t| - 1)} \sum_{w_i, w_j \in W_t, i \neq j} \text{NPMI}(w_i, w_j) \tag{2}$$

where:

- T is the set of topics.
- t is a topic in T.
- W_t is the set of top words in topic t.
- w_i, w_j are words in W_t.
- $\text{NPMI}(w_i, w_j)$ is the normalized pointwise mutual information between words w_i and w_j.

As shown in Figure 2, when the topics number is set as 20, the coherence score is the highest, which is a novel finding while the common publications believe the higher topics number the higher coherence score. Coherence score doesn't necessarily increase with the topics number is also proven in the following NMF methods, which has the best coherence score with 15 topics number. Higher topics number does consume more running time.

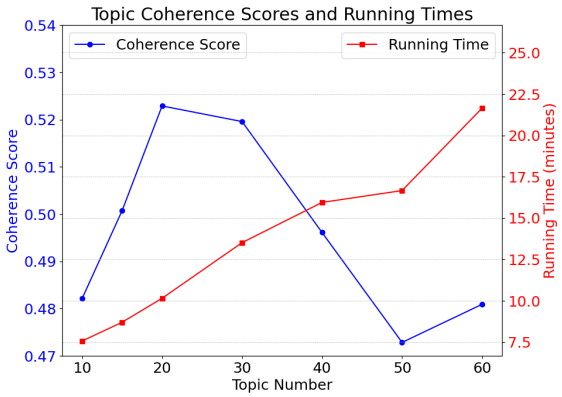

Fig. 2. Coherence score & running time with various number of topics

5) NMF with TfidfVectorizer: Non-negative Matrix Factorization (NMF) is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. In the sklearn library, NMF is implemented to decompose data into interpretable components, which is particularly useful for topic detection and image processing.

The basic form of NMF is given by:

$$V \approx WH \tag{3}$$

where:

- $V \in \mathbb{R}^{m \times n}$ is the original non-negative matrix.
- $W \in \mathbb{R}^{m \times r}$ is the non-negative matrix containing the basis vectors.
- $H \in \mathbb{R}^{r \times n}$ is the non-negative matrix containing the coefficients.
- r is the number of components or topics.

The objective of NMF is to minimize the reconstruction error, which can be measured using various loss functions. One common approach is to minimize the Frobenius norm of the difference between V and WH:

$$\min_{W,H} \|V - WH\|_F^2 \tag{4}$$

subject to the constraints that $W \geq 0$ and $H \geq 0$.

In the code 'Part 3: NMF with tfidfVec' and 'Part 6: NMF With TFIDFVectorizer', We employe the NMF model with tfidfVectorizer using the **sklearn** library for topic modeling. The NMF processing steps and parameters is shown below:

- **TF-IDF Vectorization**: Instead of using a dictionary, we used the **TfidfVectorizer** from **sklearn** to convert the text data into a TF-IDF weighted term-document matrix. The parameters were set to a maximum of 100,000 features, with terms appearing in more than 50% of the documents or fewer than 10% being excluded.
- **NMF Model Training**: The NMF model was trained on the TF-IDF matrix. The number of topics was set to 20, and the model was initialized with a random state of 42 for reproducibility.
  - **num_topics=20**: The model was configured to extract 20 distinct topics from the corpus.
  - **num_top_words=15**: For each topic, the top 15 most significant keywords were extracted.
  - **max_features=100000**: The maximum number of features to consider when building the TF-IDF matrix.
  - **max_df=0.5**: Terms appearing in more than 50% of the documents were excluded.
  - **min_df=0.1**: Terms appearing in fewer than 10% of the documents were excluded.
  - **random_state=42**: Ensured reproducibility of the results.

D. Clause Extraction

Two counting methods and TF-IDF method all extract the top 20 keywords. The keywords are not grouped into different topics, so we put them in a list and search all the documents, if any document contains more than 4 of the keywords, we output the 4 or more keywords followed by the sentence.

The LDA with CountVectorizer and NMF with TfidfVectorizer methods assume multiple topics is distributed across the documents. These methods identify the required number of topics within the documents, and then output the two most dominant topic each with a list of 15 keywords. To better understand the true topics behind these keywords, we extracted the top 10 documents most dominated by each dominant topic. Figure 3 illustrates the topic weights for 15 topics across these 10 documents from NMF method, with topic 12 being the most dominant. By reading through these 10 documents, we can interpret the underlying themes of the topics.

The following steps are performed for extracting the top 10 documents:

- **Document Topic Distribution**: For each document in the corpus, the topic distribution is obtained using the LDA or NMF model's **get_document_topics** method.
- **Dominant Topic Identification**: The dominant topic for each document is identified by selecting the topic with the highest probability.
- **Clause Extraction**: Top 10 documents containing the dominant topics are extracted and stored for further analysis.

E. Topic Interpretation

Topic interpretation relies heavily on human judgment. To aid in better interpretation by humans, word clouds were proposed. A word cloud is a visual representation of text data where the size of each word indicates its frequency or importance in the dataset. This visualization helps people quickly grasp the most prominent terms associated with each topic, making it easier to understand and interpret the underlying themes.

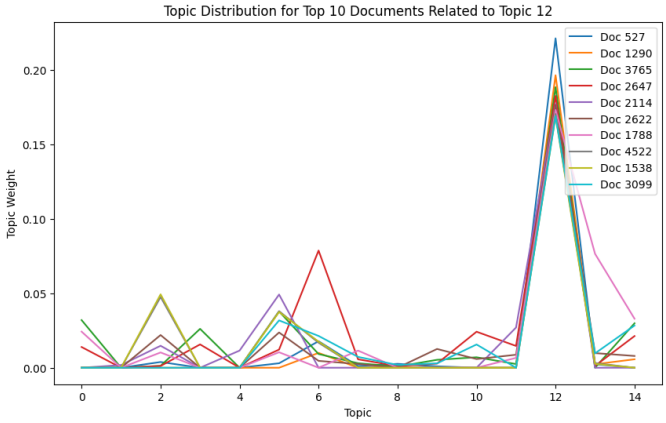

Fig. 3. Topic distribution in top 10 topic 12 dominated documents

The following steps are performed to create the word clouds:

- **Color Generation**: A set of random colors is generated for the word cloud using a custom function.
- **Word Frequency Calculation**: For each topic, the top 15 keywords and their corresponding probabilities are extracted using the LDA model's **show_topic** method.
- **Word Cloud Creation**: The **WordCloud** class from the **wordcloud** library is used to generate word clouds. The word frequencies are used to size the words proportionally in the word cloud.

The Figure 4 shows the word cloud of the most dominant topic in students demography. We can easily find that this topic is about computer, website and internet stuffs, the keywords are quite close to each other, which also reflects on the coherence score 0.5986.

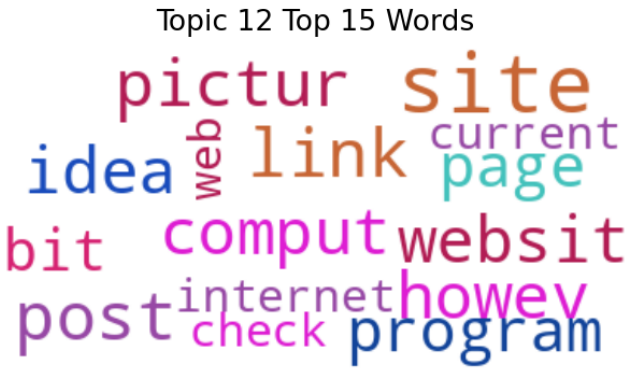

Fig. 4. Word cloud of the most dominant topic

To verify the interpretation from the word cloud, we further analyze the top documents extracted from the 'Clause

Extraction' step. The keyword of topic and summary of per document is given below:

1) **Internet**: The most dominant topic is the democratic nature and influence of the internet on society, including its impact on governance, equality, cultural dynamics, and user experience. It talks about the internet's role as a platform for free speech, its democratizing potential, and the challenges and opportunities it presents in various contexts such as education, technology, and social interaction.
2) **Internet**: It talks about the influence and impact of technology and the internet on human life and society. It discusses how technology shapes interactions, communication, and the broader societal structures, reflecting on both positive and negative aspects of technological advancements.
3) **Gadgets**: This document is about the continuous advancements and trends in consumer technology and gadgets, with a focus on new, innovative, and compact electronic devices that cater to various aspects of everyday life.
4) **Technology and others**: This document is a personal commentary on various recent experiences and reflections, spanning a range of subjects including movies, technology, social issues, personal anecdotes, and broader societal observations.
5) **Web Development**: The main topic is the personal narrative of the author's experiences and reflections on university life, web development, and the challenges and frustrations associated with balancing personal interests and academic responsibilities.
6) **Website and others**: The most dominant topic of this document is the detailed updates and personal reflections of the author regarding their involvement with various tech-related activities and platforms, such as working on a website, discussing experiences with different software and hardware, and providing information and tips on computer-related issues.
7) **Technical Issues and others**: Various topics are talked in this document, from personal details, activities to technical issues with computers, software updates, and fixing problems with internet connections and devices
8) **Programming and others**: The most dominant topic of this sentence is the organizational and administrative activities related to a school or club, including appointments, events, member roles, programming, quizzes, and general announcements and updates.
9) **Severs and others**: The topics of this document is various, computer servers and new systems are mainly talked about, as well as club events.
10) **Technology**: This sentence is about technology and related developments, specifically covering legal music downloads, handheld devices, computer systems, software updates, space exploration, security breaches, and market trends in the tech industry.

From the analysis of these 10 documents, we are confident to interpret the first dominant topic of students demography is computer related technology.

IV. Results Analysis

The results analysis presented below uses student demographics as an example. We conducted a thorough analysis of all the methods employed and demonstrate why we discarded some methods, spending more time on LDA and NMF, as shown in table III. We explain why we ultimately chose the NMF method. The results for all demographics will be provided at the end.

A. Why we give up counting and tf-idf

Taking counting all nouns strategy as example, we use the spaCy natural language processing library and the CountVectorizer from scikit-learn to count all nouns across all students documents to identify the top 20 most frequently occurring nouns. Below are the top 20 nouns identified, along with their respective counts: **time**: 68,042, **day**: 54,644, **thing**: 52,259, **today**: 39,800, **love**: 31,882, **people**: 29,799, **friend**: 29,189, **way**: 28,815, **life**: 27,909, **work**: 26,073, **school**: 24,986, **year**: 24,387, **night**: 22,791, **talk**: 22,686, **guy**: 20,869, **week**: 18,139, **lot**: 18,130, **fun**: 17,911, **person**: 17,133, **man**: 17,004.

These identified most common nouns do not clearly reveal their corresponding themes. If we were to categorize them into two themes, it might look something like this: Words such as "time," "day," "life," "work," and "school" indicate a focus on daily routines and personal experiences. Terms like "love," "friend," and "person" point to discussions about relationships and social interactions. Together, these nouns likely form topics around everyday life, relationships, and individual experiences. We are not saying these topics are wrong, just this broad categorization is not very useful for an innovation company that has acquired these blogs with the objective of innovating new products or services. The themes are too general and lack the specificity needed to generate actionable insights for product or service innovation.

When we employ the TF-IDF strategy, we get a consistent result with counting all nouns if we only check the nouns in the results. The same issue exists, it's hard to segment topics from the keywords and the topics are too general. Because of this, we give up these two methods.

like: 500.5750, **know**: 378.6944, **think**: 357.8869, **really**: 319.8933, **time**: 317.9421, **day**: 290.4602, **good**: 281.3713, **thing**: 261.0636, **love**: 255.9845, **want**: 254.9999, **feel**: 244.8542, **people**: 236.3834, **come**: 230.0447, **today**: 229.6730, **friend**: 212.0538, **tell**: 181.2683, **life**: 179.5089, **school**: 178.4504, **work**: 178.0631, **look**: 174.7575.

B. Why we give up N-gram

N-grams can identify common word pairs or triplets that frequently appear together, providing more context than single words. For example, in social media analysis, bigrams like

TABLE III
ANALYSIS AND COMPARISON OF VARIOUS METHODS FOR TOPIC EXTRACTION

Methods	**Top occurring keywords**	**Interpreted topics from extracted clauses**	**Results analysis and method cons**
Counting nouns	time, day, thing, today, love, peopl, friend, way, life, work, school, year, night, talk, guy, week, lot, fun, person, man	Social relationships and interactions, personal reflections and emotions, daily life and activities	The extracted clauses are tend to reflect the same topics on daily life things, it's not easy to pick up the second topic. The daily life topic is just too general, it covers a lot of things.
Counting subjects	thing, realli, go, life, peopl, friend, tri, person, time, guy, one, love, man, wan, god, hope, work, day, someon, girl	Emotion challenges of teenage life, emotion experiences, daily life reflections	Similar the counting nouns.
Counting direct objects	thing, time, peopl, realli, life, love, work, friend, stuff, school, place, way, fun, talk, hope, reason, guy, person, money, one	Emotional, daily life reflections, mental health	Similar to the counting nouns.
Counting prepositional objects	time, day, thing, realli, peopl, way, friend, life, hour, love, week, year, man, guy, school, know, person, work, think, stuff	Illness, coping with emotions, daily life reflections	Similar to the counting nouns.
TF-IDF	like, know, think, realli, time, day, good, thing, love, want, feel, peopl, come, today, friend, tell, life, school, work, look	Mental health and personal growth, social and ethical reflections, relationships and daily routines	Kind of the same with the counting nouns.
Bigram	('feel', 'like'), ('love', 'love'), ('look', 'like'), ('get', 'to'), ('good', 'friend')	Daily life experiences, personal growth	Kind of the same with the counting nouns.
Trigram	('love', 'love', 'love'), ('love', 'oh', 'love'), ('oh', 'love', 'oh'), ('blah', 'blah', 'blah'), ('josh', 'josh', 'josh')	Travel and personal updates, daily experiences, personal reflections	Kind of the same with the counting nouns.
LDA countVectorizer	summer, tomorrow, probabl, weekend, tonight, excit, famili, job, move, realiz, car, month, drive, abl, dad	Reflections of life and plans, feelings of upcoming holiday	When analyzing the extracted clauses, we found that the clauses under the second topic of LDA had poor relevance, with multiple themes that were not very related. We find that the topics in LDA tend to revolve around a major theme, with each article representing a sub-theme in this hierarchical structure. This sometimes results in poor relevance
	hand, woman, eye, face, door, stori, black, three, side, white, light, point, child, open, foot	Observations of nature and animals, personal reflections and experiences, fantasy and adventure	
NMF tfidfVectorizer	**site link post pictur howev comput idea websit program page bit internet web current check**	**Internet technologies computer, website, programming**	**Analysis of the extracted top 10 documents of each topics give us very clear topics. The topics are also revealed from the keywords set obviously. Thus making this our best choice when extracting blog information for an innovation business.**
	heart dream eye pain cri smile hurt face tear beauti hand fall soul hold die	**Daily life reflection**	

”climate change” or ”artificial intelligence” can capture specific topics more effectively than individual words. However, in our code, the extracted bigrams and trigrams, as shown below, do not suggest meaningful topics. One strategy is to remove the nonsensical combinations to let the meaningful ones emerge, but there are just too many nonsensical combinations, making it too labor-intensive to optimize. Section of results are shown as below.

Top 20 most common bigrams: (’feel’, ’like’), (’love’, ’love’), (’look’, ’like’), (’get’, ’to’), (’good’, ’friend’), (’come’, ’home’), (’long’, ’time’), (’will’, ’not’), (’year’, ’old’), (’realli’, ’want’), (’realli’, ’good’)

Top 20 most common trigrams: (’love’, ’love’, ’love’), (’love’, ’oh’, ’love’), (’oh’, ’love’, ’oh’), (’blah’, ’blah’, ’blah’), (’josh’, ’josh’, ’josh’), (’la’, ’la’, ’la’), (’realli’, ’realli’, ’realli’), (’realli’, ’feel’, ’like’), (’problem’, ’problem’, ’problem’), (’ha’, ’ha’, ’ha’)

C. LDA with counting vectorization

In the example of student demography, LDA with CountVectorizer determines two most dominant topics which are topic 17 and topic 3. Figure 5 are the visual representations of the topic 17 in the format of word clouds.

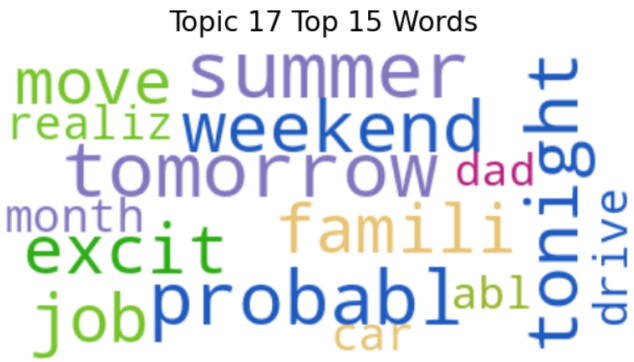

Fig. 5. LDA Topic 17 Word Cloud

We extract the top 10 document dominated by topic 17, and analyze these files to identify the main themes of this topic.

1) **Daily life and connections**: This file discusses keeping in touch with friends through blogging, covering life events such as graduation, work, church, achievements, relationships, and maintaining connections with friends in different countries.
2) **Feelings**: This file talks about the author’s feelings of loneliness, frustration, and sadness during summer, feeling misunderstood or judged by family, and struggling with depressive feelings and lack of plans.
3) **Reflections on activities and family**: The dominant topic is the author’s reflections on recent and upcoming activities, family dynamics, and personal feelings, including weekend experiences, family interactions, struggles with sleep, and thoughts on independence.
4) **Summer plans and relationships**: This file reflects on the author’s experiences and feelings about babysitting, relationships, and summer plans, discussing their enjoyment of babysitting, dynamics with their boyfriend, and hopes for the summer.
5) **Life challenges and reflections**: The dominant topic is the author’s reflections and frustrations about life, including their upcoming birthday, school and nursing program challenges, family dynamics, and experiences with different churches.
6) **Transitions and aspirations**: This file reflects on the author’s current life challenges and transitions, including moving to a new apartment, starting a new job, dealing with family issues, and considering starting a business, expressing personal frustrations and aspirations.
7) **Heather’s life changes**: This file covers Heather’s recent life changes and activities, including moving, starting a job, attending weddings, and considering a youth ministry position.
8) **Exciting Summer plans**: This file describes the narrator’s exciting summer plans and experiences, including trips to San Francisco, Six Flags, a water park, Norway, and Colorado, touching on the joy of spending time with best friends and maintaining close friendships.
9) **Emotional struggles**: This file addresses the narrator’s emotional struggle with a challenging friendship and related social and personal issues, describing feelings of frustration, betrayal, disappointment, and moments of hope.
10) **Weekly reflections**: This file reflects on the narrator’s week, highlighting creative pursuits, work experiences, financial concerns, and anticipation for upcoming events.

The topic keywords of topic 17 is ”summer, tomorrow, probably, weekend, tonight, excited, family, job, move, realize, car, month, drive, able, dad”. Topic 17 revolves around personal life events, plans, and transitions. It includes reflections on daily activities, emotional experiences, and interactions with family and friends. LDA identified a very meaningful topic; however, the scope of this topic is still somewhat too broad. It is like having many branches on a tree that all belong to the same theme but are somewhat independent of each other.

Similarly, for the purpose of innovation, the LDA method may not be very suitable. The broad themes identified are too general and lack the specificity needed to generate actionable insights for developing new products or services.

D. NMF with TF-IDF vectorization

In the ’Clause Extraction’ section, we have reviewed the most dominant topic determined by the NMF method, and confidently interprete the topic is primarily centered around technology, internet, and personal experiences with digital tools, this is very suggestive for a company to make decision. Comparing all these methods, the NMF results in the most convincing dominant topics, so all the demographics are then being processed by NMF methods and the results are shown in Table IV.

TABLE IV
TWO DOMINANT TOPICS PER DEMOGRAPHY WITH NMF-TFIDF TOPIC MODELING

Demography	Keywords per topic	Interpreted Topics	Coherence
Student	**Set 1:** site, link, post, picture, however, computer, idea, website, program, page, bit, internet, web, current, check	**Internet technologies**	0.5986
	Set 2: heart, dream, eye, pain, cry, smile, hurt, face, tear, beauty, hand, fall, soul, hold, die	**Daily life reflection**	
Male	**Set 1:** job, weekend, money, company, tomorrow, pay, car, business, tonight, phone, couple, plan, office, drive, manage	**Job and mental health**	0.6036
	Set 2: car, water, city, picture, road, bike, train, ride, trip, park, room, minute, street, bus, drink	**Travel experiences**	
Female	**Set 1:** heart, dream, pain, hurt, cry, smile, tear, soul, alone, felt, fall, hold, inside, true, moment	**Love and emotional journey**	0.5766
	Set 2: water, car, room, food, dog, door, hair, cat, picture, minute, red, hot, foot, wear, drink	**Urban life and reflections**	
Age $>$ 20	**Set 1:** student, class, experience, human, study, education, teacher, community, university, question, culture, however, teach, develop, individual	**Pressure and political critique**	0.6602
	Set 2: heart, dream, pain, smile, soul, hurt, cry, wish, moment, tear, felt, alone, beauty, dark, inside	**Heartbreak and emotional struggles**	
Age $\leq$ 20	**Set 1:** heart, dream, hurt, pain, cry, eye, cause, woe, smile, alone, believe, tear, fall, true, hold	**Heartbreak and love**	0.6034
	Set 2: mom, dad, movie, tomorrow, summer, car, sister, tonight, brother, family, awesome, excited, yesterday, drive, hang	**Daily activities and reflections**	
Everyone	**Set 1:** job, weekend, tomorrow, tonight, plan, money, excited, business, meet, yesterday, couple, phone, pay, party, office	**Life events and reflections**	0.6101
	Set 2: class, book, student, learn, experience, human, question, study, teacher, however, become, fact, perhaps, university, common	**Citizenship and societal expectations**	

V. SUMMARY AND REFLECTIONS

A. Summary

In our study, we developed a comprehensive methodology for hot topic detection by processing and analyzing a large dataset of student files. Our pipeline included data preparation, pre-processing, topic modeling, clause extraction, and result evaluation. We began by quickly setting up the entire pipeline with a subset of 100 files, then extended our analysis to the full dataset of 5120 student files, taking approximately 10-20 minutes per task.

The dataset was segmented by various demographics, and the files were cleaned and prepared through a series of pre-processing steps such as tokenization, stopwords removal, spell-checking, and parallel processing using Dask. Different topic detection methods were explored, including counting methods, TF-IDF, n-grams, LDA with CountVectorizer, and NMF with TfidfVectorizer.

We found that the Non-negative Matrix Factorization (NMF) method provided the most reasonable and interpretable topics. While counting and TF-IDF methods resulted in broad and general themes, and n-grams produced too many nonsensical combinations, NMF effectively identified meaningful topics. The LDA method, although useful, also resulted in broader themes that were less actionable for product or service innovation.

Our analysis demonstrated that NMF with TfidfVectorizer yielded the highest coherence scores and the most specific topics, making it the preferred method for our study. The final results across different demographics showed distinct and actionable themes, providing valuable insights for further exploration and innovation.

By leveraging the strengths of NMF and TF-IDF, we successfully identified relevant topics that can inform the development of new products or services, addressing the specific needs and interests of different demographic groups. This approach not only enhanced our understanding of the data but also demonstrated the practical applications of advanced topic modeling techniques in real-world scenarios.

B. Reflections

During the course of this project, we initially built the pipeline and ran it end-to-end. We then proceeded to modify the LDA parameters in hopes of achieving better results. After numerous futile attempts, we gradually came to understand the principles of LDA. This experience highlights an important lesson: instead of diving straight into parameter adjustments and testing, we should first thoroughly discuss and understand the underlying principles of the methods we intend to use.

We find that when using the counting method to extract topics, it is initially difficult to identify multiple themes. However, by observing which top keywords frequently appear together, we can group those keywords into sets. We then extract articles containing these related keyword sets and analyze the themes discussed in those articles. If we redo the assignment, we will investigate deeper on this approach.

APPENDIX

The code url for hot topic detection is : https://colab.research.google.com/drive/1m1lXYo8eZPRmBuT3rEpfpdNpRladN9yo?usp=sharing

REFERENCES

[1] David M Blei, Andrew Y Ng, and Jordan@cs Berkeley Edu. Latent dirichlet allocation michael i. jordan, 2003.

[2] Amar Jeet Rawat, Rakesh Semwal, and Rajeev Kumar. Topic modeling based consumer behavior analysis using latent dirichlet allocation. pages 582–587. Institute of Electrical and Electronics Engineers Inc., 2023.

[3] Aziida Nanyonga, Hassan Wasswa, and Graham Wild. Topic modeling analysis of aviation accident reports: A comparative study between lda and nmf models. Institute of Electrical and Electronics Engineers Inc., 2023.

[4] Bin Yin and Chih Hung Yuan. Detecting latent topics and trends in blended learning using lda topic modeling. *Education and Information Technologies*, 27:12689–12712, 11 2022.

[5] Hamza H.M. Altarturi, Muntadher Saadoon, and Nor Badrul Anuar. Web content topic modeling using lda and html tags. *PeerJ Computer Science*, 9, 2023.

[6] Indra, Edi Winarko, and Reza Pulungan. Trending topics detection of indonesian tweets using bn-grams and doc-p. *Journal of King Saud University - Computer and Information Sciences*, 31:266–274, 4 2019.

[7] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. pages 399–408. Association for Computing Machinery, 2 2015.

[8] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning*, New York, NY, USA, 2009. ACM.

[9] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, Red Hook, NY, USA, 2009. Curran Associates, Inc.

[10] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

APPENDIX

Parma Nand

Marking rubric		
Criteria	**Ratings**	**Points**
Research Question and Rationale Description view longer description	**Comments** Done well and the language is coherent.	9 / 10 pts
Data Description and Analysis view longer description	**Comments** Data described well, including noise processing. How about long and short blogs? Did this have any impact?	13 / 15 pts
Research Design view longer description	**Comments** Described well. Rationale for chaning from TF/TFIDF to LDA explained. However what were the two most common topics in each of the demographics? this was the core of the project. In table IV, How would the topic "Daily life reflection", help the manager build strategies?. It seems too general.	26 / 30 pts
Implementation (Code) submitted as Appendix view longer description	**Comments** submitted and well documented.	13 / 15 pts
Analysis and Evaluation view longer description	**Comments** Done well, except no evidence of which one is better out of TF/TFIDF. Did you do any manual evaluation of the topics to see how good your results were compated to your evaluation?	16 / 20 pts
Conclusion, Formatting, Language and References view longer description	**Comments** Done well. however could contain some info about the acutal topics retrieved as an answer to the RQ.	9 / 10 pts
		Total points: 86

Fig. 6. Marking details of this paper