TOPIC MODELING OF SOCIAL MEDIA X USERS' PERCEPTIONS ON THE KAMPUS MERDEKA INTERNSHIP PROGRAM USING BERTOPIC

KAMILAH NURUL AZIZAH

STUDY PROGRAM OF STATISTICS AND DATA SCIENCE
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
BOGOR
2025

STATEMENT REGARDING SKRIPSI AND SOURCES OF INFORMATION, AS WELL AS COPYRIGHT TRANSFER

As the author, I affirm that the skripsi titled "Topic Modeling of Social Media X Users' Perceptions on The Kampus Merdeka Internship Program Using BERTopic" represents my work under the guidance of my supervisors and has not been presented in any format to any other university. Any sources from published or unpublished works by other authors are duly acknowledged within the text and documented in the references section after this skripsi.

The author hereby transfers the copyright of this skripsi to IPB University.

Bogor, June 2025

Kamilah Nurul Azizah
G1401211073

ABSTRAK

KAMILAH NURUL AZIZAH. Pemodelan Topik Persepsi Pengguna Media Sosial X terhadap Program Magang Kampus Merdeka Menggunakan BERTopic. Dibimbing oleh KHAIRIL ANWAR NOTODIPUTRO dan LAILY NISSA ATUL MUALIFAH.

Program Merdeka Belajar Kampus Merdeka (MBKM) yang diinisiasi oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi bertujuan menjembatani kesenjangan antara teori akademik dan kebutuhan industri. Salah satu program utamanya adalah Magang Kampus Merdeka yang dirancang untuk membekali mahasiswa dengan pengalaman kerja praktis. Sebagai program berskala nasional, pelaksanaannya menimbulkan berbagai respons dan diskusi publik yang masif di platform media sosial, khususnya X. Penelitian ini bertujuan menerapkan metode pemodelan topik berbasis transformer yaitu BERTopic untuk mengidentifikasi topik utama dalam diskusi mengenai program tersebut. Analisis dilakukan pada 16.943 data dari platform X untuk periode 21 Mei 2021–28 Februari 2025. Proses pemodelan BERTopic meliputi IndoSBERT untuk *embedding* teks, UMAP untuk reduksi dimensi, HDBSCAN untuk klasterisasi, dan c-TF-IDF untuk representasi topik. Model dioptimalkan menggunakan Optuna serta penerapan *Maximal Marginal Relevance*, yang menghasilkan delapan topik dengan skor koherensi 0,48 dan diversitas 0,96. Topik yang teridentifikasi mencakup manfaat seperti pengembangan profesional dan dukungan finansial, serta tantangan berupa kendala administratif hingga perdebatan ideologis. Temuan ini menjadi landasan rekomendasi untuk perbaikan implementasi, tata kelola, dan keberlanjutan program.

Kata kunci: BERTopic, magang kampus merdeka, media sosial X, opini publik, pemodelan topik

ABSTRACT

KAMILAH NURUL AZIZAH. Topic Modeling of Social Media X Users' Perceptions on The Kampus Merdeka Internship Program Using BERTopic. Supervised by KHAIRIL ANWAR NOTODIPUTRO. and LAILY NISSA ATUL MUALIFAH.

The Merdeka Belajar Kampus Merdeka (MBKM) program, initiated by the Ministry of Education, Culture, Research, and Technology, aims to bridge the gap between academic theory and industry needs. One of its flagship programs is the Kampus Merdeka Internship, designed to equip students with practical work experience. As a large-scale national program with widespread impact, its implementation has generated massive public discussion and diverse responses on social media platforms, particularly X. This study aims to implement BERTopic, a transformer-based topic modeling method, to identify the main topics in the discourse surrounding the program. The analysis was performed on 16,943 data collected from the X platform, covering the period from May 21, 2021, to February 28, 2025. The BERTopic modeling process involves IndoSBERT for text embedding, UMAP for dimensionality reduction, HDBSCAN for clustering, and c-TF-IDF for topic representation. The model was optimized using Optuna and the application of Maximal Marginal Relevance, yielding eight topics with a coherence score of 0.48 and a diversity score of 0.96. The identified topics cover benefits, such as professional development and financial support, as well as challenges ranging from administrative hurdles to ideological debates. These findings provide a basis for recommendations aimed at improving the program's implementation, governance, and sustainability.

Keywords: BERTopic, kampus merdeka internship, public opinion, social media X, topic modeling

TOPIC MODELING OF SOCIAL MEDIA X USERS' PERCEPTIONS ON THE KAMPUS MERDEKA INTERNSHIP PROGRAM USING BERTOPIC

KAMILAH NURUL AZIZAH

Skripsi
as a partial fulfilment for the degree of
Bachelor in
Study Program of Statistics and Data Science

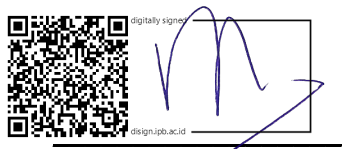
STUDY PROGRAM OF STATISTICS AND DATA SCIENCE
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
BOGOR
2025

External Skripsi Examiner:
Rahma Anisa, S.Stat., M.Si.

Title : Topic Modeling of Social Media X Users' Perceptions on The Kampus Merdeka Internship Program Using BERTopic
Name : Kamilah Nurul Azizah
Student ID : G1401211073

Approved by

Main Supervisor:
Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S.



Co-Supervisor:
Laily Nissa Atul Mualifah, M.Si.

Acknowledge by

Head of Statistics and Data Science Study Program:
Dr. Bagus Sartono, S.Si, M.Si.
NIP 19780411 200501 1002

Examination Date:
June 19th, 2025

Date of Graduation:

PREFACE

All praise and gratitude are extended to Allah subhanahu wa ta'ala for His abundant blessings, which have enabled the successful completion of this scientific work. This project conducted from January 2025 to June 2025 by exploring the field of text analysis, with the title "Topic Modeling of Social Media X Users' Perceptions on The Kampus Merdeka Internship Program Using BERTopic".

The completion of this project would not have been possible without the invaluable assistance and support from various parties who have made significant contributions throughout this process. Therefore, the author wishes to express the deepest gratitude to:

1. The author's beloved parents, Muhamad Soleh and Ade Nuraeni, as well as the author's siblings and entire extended family. Their continuous support, heartfelt prayers, and unconditional love have been a constant source of motivation and strength throughout this journey.
2. The author's esteemed advisors, Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S. and Laily Nissa Atul Mualifah, M.Si.. Their profound guidance, insightful direction, and unwavering support have been instrumental in the preparation and execution of this project.
3. The external examiner, Rahma Anisa, S.Stat. M.Si., for her valuable comments, constructive criticism, and insightful suggestions which contributed to the improvement and refinement of this project.
4. The colloquium and seminar moderators, Akbar Rizki, S.Stat. M.Si. and Gerry Alfa Dito, S.Si. M.Si., for their thoughtful feedback and engaging questions during the presentations which helped sharpen the project focus.
5. All the lecturers and administrative staff of the Study Program of Statistics and Data Science. Their dedication to imparting knowledge and their assistance in meeting the author's needs throughout the academic journey and the preparation of this scientific work are gratefully acknowledged.
6. The author's dearest friends, especially Alfidhia Rahman Nasa Juhanda, Anis, Aida, Dinda, Nadila, Diva, and Dhiya. Their encouragement, companionship, and spirited discussions made this challenging journey.
7. All other parties who provided moral and material support, whose names cannot be mentioned individually.

It is the author's sincere hope that this scientific paper will be beneficial to those in need and contribute to the advancement of knowledge.

Bogor, June 2025

Kamilah Nurul Azizah

TABLE OF CONTENTS

LIST OF TABLES		x
LIST OF FIGURES		x
I	INTRODUCTION	1
	1.1 Background	1
	1.2 Objectives	2
II	LITERATURE REVIEW	3
	2.1 Kampus Merdeka Internship	3
	2.2 Web Crawling	3
	2.3 BERTopic	3
	2.4 Text Embedding with IndoSBERT	4
	2.5 Uniform Manifold Approximation and Projection (UMAP)	6
	2.6 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)	7
	2.7 CountVectorizer	7
	2.8 Class-Based Term Frequency-Inverse Document Frequency (c-TF-IDF)	7
	2.9 Fine-Tune Topic Representation	8
	2.10 Optuna	8
	2.11 Evaluation Metrics: Topic Coherence	8
	2.12 Evaluation Metrics: Topic Diversity	9
III	METHODOLOGY	10
	3.1 Data Source	10
	3.2 Data Analysis Procedure	10
IV	RESULT AND DISCUSSION	13
	4.1 Data Selection	13
	4.2 Data Pre-processing	14
	4.3 Exploratory Data Analysis	14
	4.4 BERTopic	16
	4.5 Hyperparameter Tuning	17
	4.6 Fine-Tuning Topic Representation	17
	4.7 Topic Modeling Results	18
	4.8 Topic Interpretation	21
	4.9 Topic Trend Analysis and Recommendations	24
V	CONCLUSION AND SUGGESTION	26
	5.1 Conclusion	26
	5.2 Suggestion	26
REFERENCES		27
APPENDICES		29
BIOGRAPHY		31

LIST OF TABLES

1	Example of dataset's structures	10
2	Examples of non-opinion tweets	13
3	Example of data pre-processing steps	14
4	Numerical representation of text	16
5	Comparison of representative keyword	18
6	Summary of topics identified	19
7	Examples of documents classified as outliers	20

LIST OF FIGURES

1	BERT input representation (Devlin *et al.* 2019)	5
2	UMAP framework	6
3	Number of tweets (a) by half year and (b) over time	15
4	Optimization contour plot	17
5	Documents by topic	18
6	Wordcloud of (a) Topic 1 and (b) Topic 2	21
7	Wordcloud of (a) Topic 3 and (b) Topic 4	22
8	Wordcloud of (a) Topic 5 and (b) Topic 6	23
9	Wordcloud of (a) Topic 7 and (b) Topic 8	23
10	Trends over time (Topic 3, Topic 4, and Topic 6)	24