Nama : Krisna Bayu Dharma Putra
NIM : 21/482071/PA/21017
Kelas : Kapita Selekta Sistem Cerdas

Analisis Paper

Judul Paper : Exploring Event-based Dynamic Topic Modeling
Link : https://ieeexplore.ieee.org/document/10438733

A. Dataset Preparation
    1) Sebanyak 9003 post platform X diambil dari tanggal 1 desember 2022 sampai 31 desember 2022. Post ini diambil menggunakan Twint, sebuah alat scraping platform X pada bahasa pemrograman python. Keywords filters yang digunakan pasti mengandung “katip, katipunan, bgc, bonifacio global city”.
    2) Pemrosesan Data : Dilakukan dengan menghapus emoji dan url, menghapus punctuation, angka, spasi berlebih, stopwords bahasa inggris dan tagalog juga dihapus. Selanjutnya, teks dilakukan *lemmatization* dan tokenisasi. *Lemmatization* dan *tokenization* tidak dilakukan pada model BERTopic.

B. Implementasi Base Model LDA

    Proses implementasi model dasar LDA digunakan sebuah parameters untuk semua base model, dimana *chunksize* diset ke 200, dan passes diset ke 10. Banyaknya topik diset dari dua sampai 9, yang kemudian dikalkulasi dan direkam untuk menentukan jumlah topik paling optimal dilakukan metode *cross validation*.

C. Implementasi Dynamic Topic Modelling

    Implementasi Dynamic Topic Modelling dilakukan dengan cara mirip dengan Base LDA, namun dengan angka beta dan alpha dalam sebuah list. List untuk pilihan angka alpha dan beta adalah [0.05, 0.1, 0.5, 1, 5, 10]. Banyaknya topik di set dari 0 sampai 10. Selanjutnya, dilakukan grid search untuk mencari kombinasi hyperparameter terbaik.

D. Evaluasi

    Evaluasi dilakukan dengan menggunakan skor koherensi.

E. Hasil dan Analisis

TABLE I Average Base Lda Model Cv Coherence Scores

	Dataset			
No. of Topics	katip	katipunan	bgc	bonifacio global city
2	0.4548	0.5863	**0.5989**	0.5136
3	**0.4943**	**0.5866**	0.4773	0.4614
4	0.4596	0.5522	0.5546	0.4643
5	0.4889	0.5100	0.5364	0.4887
6	0.4858	0.5120	0.5310	0.5124
7	0.4547	0.5006	0.5183	0.5270
8	0.4534	0.5131	0.5368	0.5191
9	0.4439	0.5060	0.5102	**0.5307**

Dapat dilihat bahwa untuk Base LDA dilihat untuk beberapa dataset mencapai nilai 0.49, 0.5866, 0.5989, dan 0.5307. Selanjutnya, ini adalah perbandingan model terbaik untuk setiap modelnya :

TABLE III Base Lda Model, Optimized Lda Model, and Bertopic Cv Coherence Scores

	CV Coherence Score		
Dataset	Base Model	Optimized Model	BERTopic
katip	0.4943	**0.5580**	0.4229
katipunan	0.5866	0.6602	**0.6637**
bgc	0.5989	**0.6625**	0.3323
bonifacio global city	0.5307	**0.6484**	0.3816

Dapat dilihat bahwa model terbaik adalah optimized model dari base LDA yang memiliki performa terbaik pada 3 buah dataset, sedangkan untuk dataset katipunan, BERTopic lebih unggul sedikit dibandingkan Optimized model.