Aplikasi *Sentiment Analysis* dan *Topic Detection* pada Data *Tweet* mengenai ChatGPT pada Bulan Maret 2023

KELOMPOK 3

Topik Khusus 2 (Web Mining)

Anggota

Tulus Setiawan
2006568802

Cornelius Justin S. H.
2006529796

M. Hanif Pramudya Z.
2006487566

Gladys Nathania B
2006534556

DAFTAR ISI

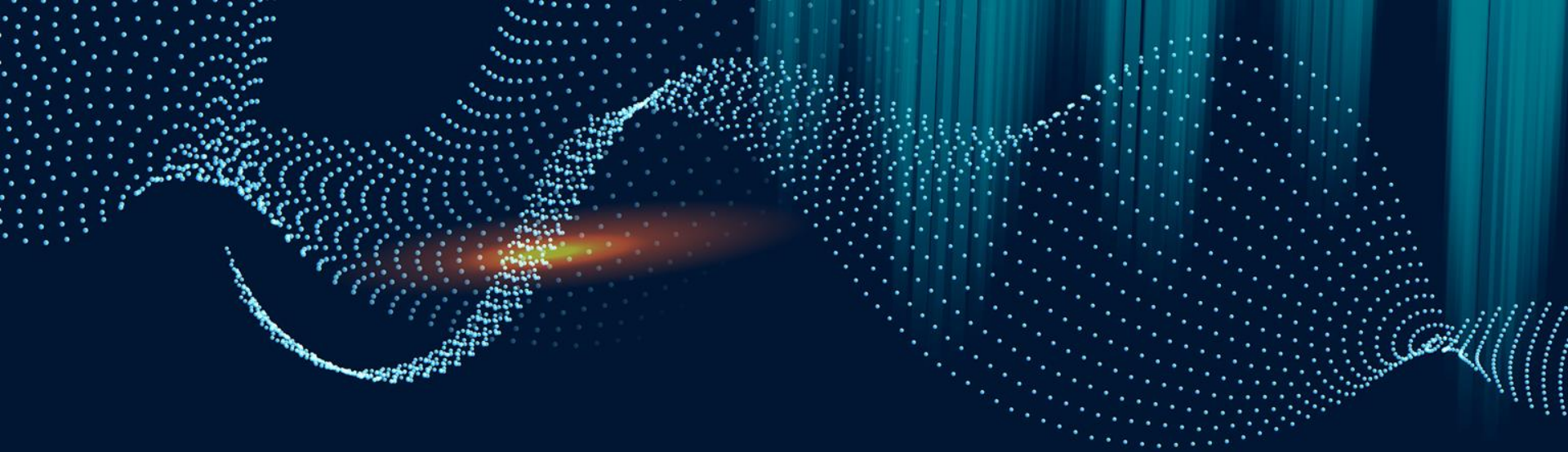
01 PENDAHULUAN

02 METODE

03 SIMULASI DAN ANALISIS HASIL

04 KESIMPULAN

05 DAFTAR PUSTAKA



01 PENDAHULUAN

Latar Belakang

ChatGPT adalah *chatbot* kecerdasan buatan (AI) yang memahami dan menghasilkan bahasa alami manusia dengan kecanggihan, kepekaan, dan kegunaan yang multi-fungsi mencakup berbagai aspek kehidupan[6].

Kemampuan dalam memberikan jawaban, solusi, dan deskripsi untuk pertanyaan-pertanyaan kompleks, termasuk cara-cara potensial untuk memecahkan masalah pada suatu situasi, menulis kode, dan menjawab pertanyaan dengan baik[7], menjadikan topik bertemakan ChatGPT ramai diperbincangkan.

Pengetahuan mengenai opini-opini yang diperbincangkan tentang ChatGPT dapat memberikan wawasan penting tentang potensi keberhasilan atau kegagalan teknologi baru ini.

Related Works

- **Mubin et al. [1]**
  - Melakukan pemodelan *topic detection* menggunakan *Latent Dirichlet Allocation* (*LDA*) serta *sentiment analysis* pada data Twitter mengenai ChatGPT pada masa awal peluncuran.
  - Memberikan gambaran respons publik awal terhadap teknologi baru tersebut.

- **Bian et al. [2]**
  - Melakukan analisis sentimen publik terhadap *Internet of Things (IoT)* melalui data Twitter.
  - Menunjukkan pengguna lebih tertarik pada aspek bisnis dan teknologi IoT dengan sentimen positif terhadap IoT.

- **Trivedi et al. [3]**
  - Menggunakan pendekatan *Robustly Optimized* BERT *Pretraining Approach* (RoBERTa) untuk menganalisis sentimen publik melalui tweet tentang pengaturan kerja hibrid.
  - Menunjukkan mayoritas pengguna memiliki sentimen positif terhadap model kerja hibrid.

Related Works

- **Studi lain. [4]**
  - Menggunakan data Twitter untuk memprediksi keberhasilan sebuah film dengan mengembangkan model prediksi peringkat dan model popularitas produk temporal.

- **Nawaz et al. [5]**
  - Mengusulkan model prediksi hasil pemilihan politik di Pakistan menggunakan data Twitter.
  - Memperoleh akurasi dan efisiensi 98%.

Tujuan

Topic Detection

Melakukan pemodelan *topic detection* untuk mengidentifikasi topik - topik yang berkaitan dengan *tweet* ChatGPT dalam Twitter pada bulan Maret 2023.

Sentiment Analysis

Melakukan analisis sentimen untuk mengidentifikasi sikap publik terhadap setiap topik yang berkaitan dengan *tweet* ChatGPT dalam Twitter pada bulan Maret 2023.

Metodologi Penelitian

Data yang akan digunakan

Data *Training*

- *tweet* bulan Desember 2022 mengenai ChatGPT
- memiliki label sentimen (negatif: 48,7%; netral: 25,5%; positif: 25,7%)
- berisi 217.622 *tweet*

Link dataset: https://www.kaggle.com/datasets/charunisa/chatgpt-sentiment-analysis

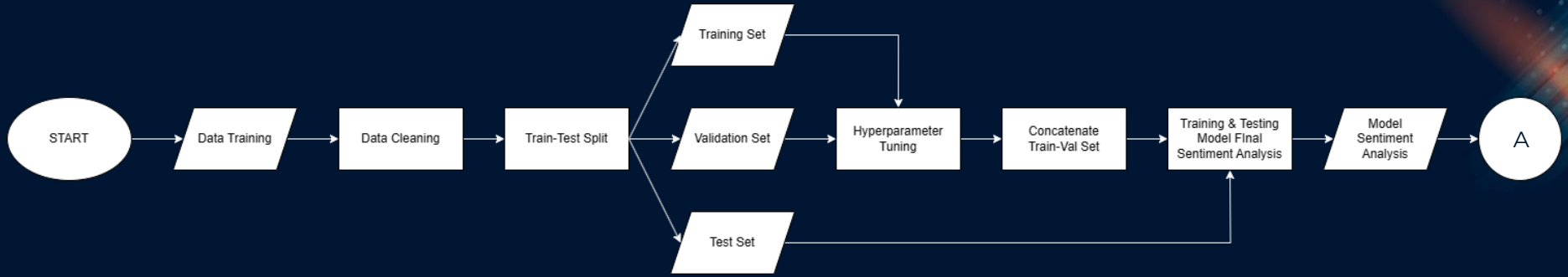
Data Prediksi

- tweet mengenai ChatGPT bulan Maret 2023
- tidak memiliki label
- berisi 92.559 *tweet*

Link dataset: https://www.kaggle.com/datasets/sanlian/tweets-about-chatgpt-march-2023

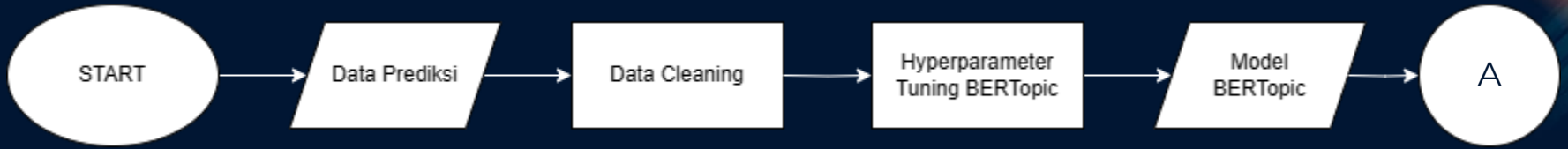
Metodologi Penelitian

Tahapan Training Model Sentiment Analysis



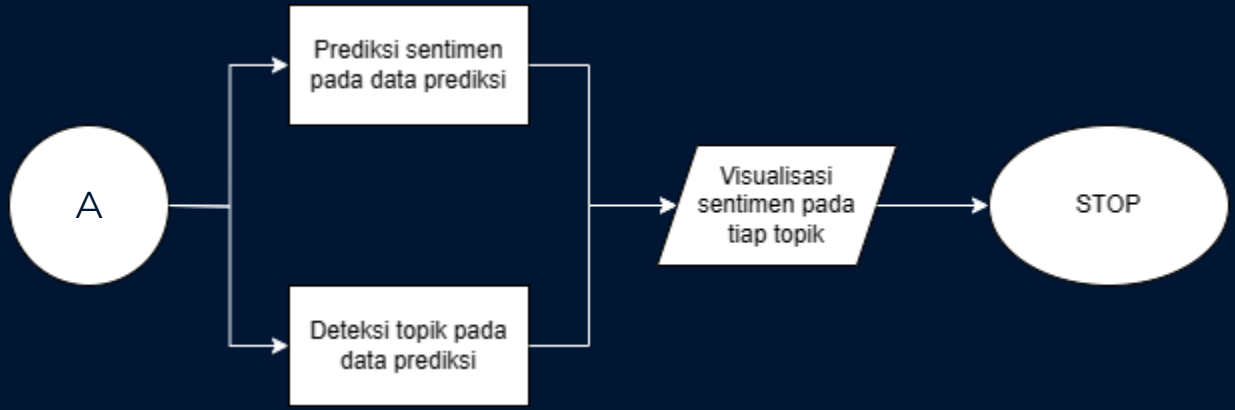
Metodologi Penelitian

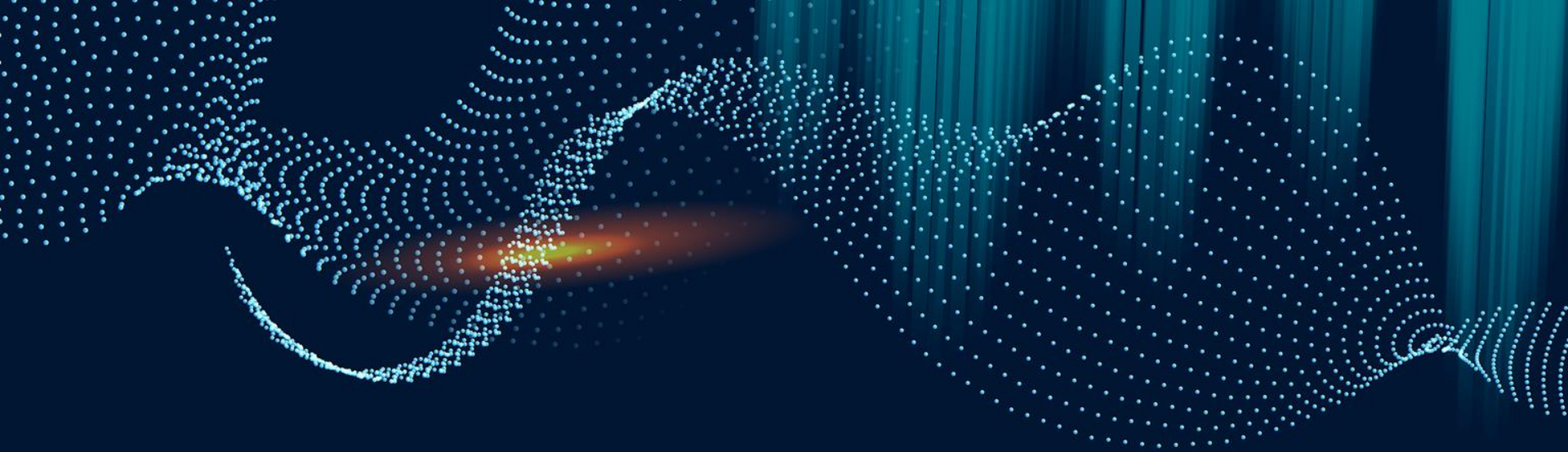
Tahapan Training Model Topic Detection (BERTopic)



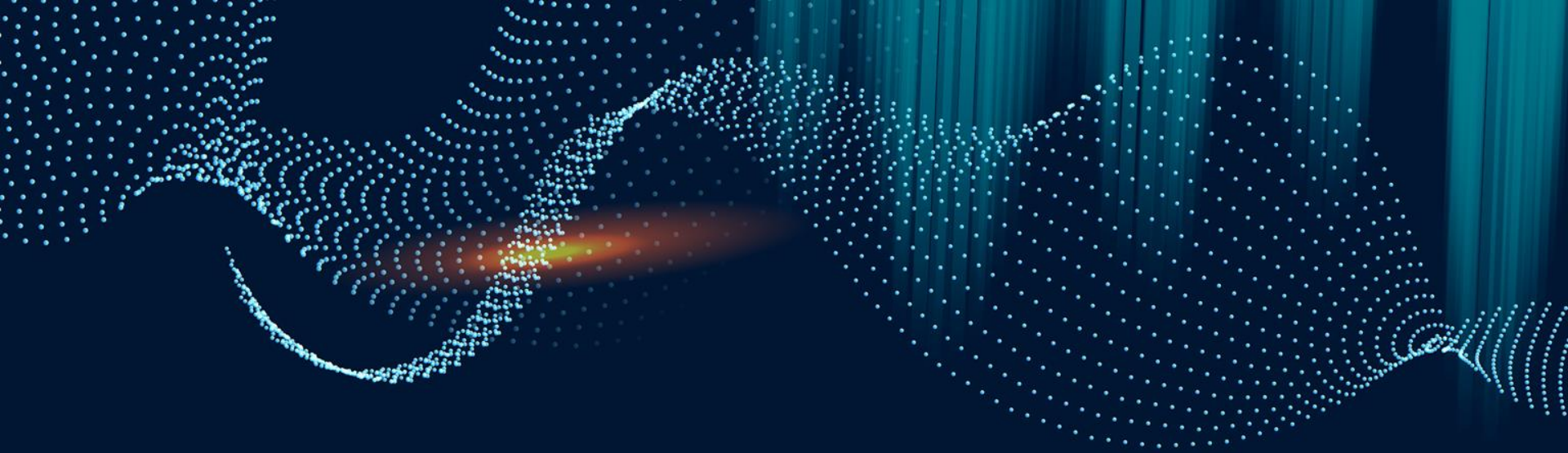
Metodologi Penelitian

Tahapan Akhir





02 METODE



Sentiment Analysis

Data Preprocessing

Label Encoding

- Melakukan encoding pada fitur labels, dimana **‘bad’ :0, ‘good’:1, ‘neutral’:2**

Data Cleaning

- Konversi kata menjadi huruf kecil semua (lower case)
- Menghapus muatan mention (@...)
- Menghapus tanda pagar (#....) menjadi kata-kata saja
- Menghapus spasi putih yang berlebih
- Menghapus muatan tautan (www. atau https://)
- Menghapus karakter unicode (*, ^,$,&,dll)
- Menghapus karakter *non-word* (a-z, 0-9)
- Menghapus baris baru
- Menghapus kata - kata yang tidak merepresentasikan sentimen dan pendeteksian topik (*stop words*)

Sentiment Analysis Model

Tokenization

- Tokenisasi pada model ini menggunakan tokenizer dari **Tensorflow Keras**. Fungsi ini akan **memvektorisasi** kumpulan teks dengan merubah setiap teks menjadi barisan bilangan bulat (setiap bilangan bulat menjadi indeks token dalam *dictionary*).
- Pada model kami, dilakukan ***Tokenizer*** dengan **fit_on_texts** pada tweets dari data train. **fit_on_texts** ini yang akan memperbaharui *vocabulary* internal berdasarkan list teks.

Model

- ***Embedding*** akan mengubah bilangan bulat positif (indexes) menjadi *dense vectors* dari *fixed size*.
- ***Bidirectional*** pembungkus dari *neural network* dalam hal ini LSTM agar dapat menangkap informasi dari dua arah.
- **LSTM** layer Long Short Term Memory untuk menghasilkan prediksi dengan mempertimbangkan informasi dari jangka waktu yang panjang dan menghapus informasi yang tidak relevan.
- **Conv1D (1-Dimension Convolution Layer)** membuat *convolution kernel* yang digabungkan dengan *input layer* melalui satu dimensi spasial (atau temporal) untuk menghasilkan tensor *output*.
- **Dropout Layer** akan secara acak membuat unit-unit input menjadi 0 dengan frekuensi kecepatan pada setiap langkah selama waktu *training*, yang membantu mencegah overfitting.
- **Dense Layer** lapisan reguler Neural Network yang *deeply connected* yang menjalankan operasi output = activation(dot(input, kernel) + bias).
- **RegularizerL1L2** merupakan regularizer yang menerapkan penalti regularisasi L1 dan L2.

Sentiment Analysis Model

Hyperparameter Tuning Sentiment Analysis model (menggunakan Bayesian Search sebanyak 15 kali)

Hyperparameter	*Possible values*
Dimensi Embedding layer	[32, 64, 96, 128]
Menggunakan Conv1D	[True, False]
Jumlah filter convolution	[32, 64, 96, 128]
Kernel size convolution	[3, 5, 7]
L1 regularization pada Conv1D	low: 0, high: 0.01, step size: 0.1
L2 regularization pada Conv1D	low: 0, high: 0.01, step size: 0.1

Hyperparameter	Possible values
Jumlah neuron pada LSTM	[16, 32, 48, 64]
L1 regularization pada LSTM	low: 0, high: 0.01, step size: 0.1
L2 regularization pada LSTM	low: 0, high: 0.01, step size: 0.1
Jumlah hidden layer dense	[1, 2]
Jumlah neuron pada hidden layer dense	low: 32, high: 256, step size: 32
L1 regularization pada dense layers	low: 0, high: 0.01, step size: 0.1
L2 regularization pada dense layers	low: 0, high: 0.01, step size: 0.1
Dropout pada dense layers	[True, False]
Learning rate	low: 10^(-5), high: 0.001

Sentiment Analysis Model

*Hyper*parameter yang tidak dilakukan *tuning* pada model *Sentiment Analysis*

Hyperparameter	*Possible Values*
Jumlah neuron pada output layer	3
Fungsi aktivasi pada hidden layer	ReLU
Fungsi aktivasi pada output layer	Softmax
Optimizer	Adam
Loss function	Categorical Crossentropy

Sentiment Analysis

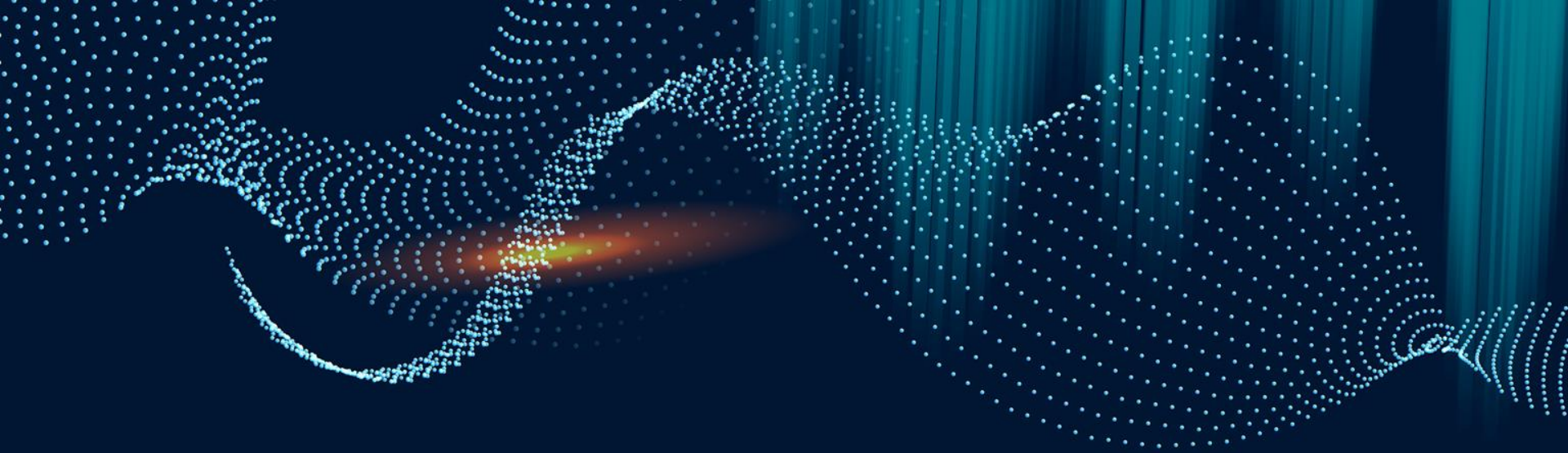
Evaluate Model

F1 SCORE

Mengukur keseimbangan antara Precision – Recall. Nilai terbaik F1-Score adalah 1.0 dan nilai terburuknya adalah 0

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

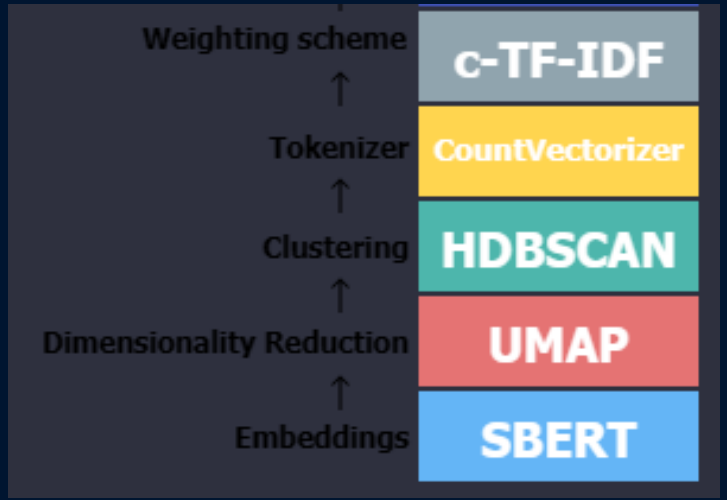
$$\text{Akurasi} = \frac{\text{jumlah data dengan kelas sesuai target}}{\text{jumlah seluruh data}} \times 100$$



Topic Detection

BERTopic

Sebuah model topik yang memanfaatkan teknik pengelompokan dan variasi TF-IDF berbasis kelas untuk menghasilkan representasi topik yang koheren. Secara lebih spesifik, BERTopic menghasilkan representasi topik melalui langkah-langkah berikut:



Sumber : The Algorithm - BERTopic (maartengr.github.io)

- Dokumen dikonversi ke representasi *embedding* menggunakan *pre-trained* model, Sentence-BERT (SBERT).
- Reduksi dimensi *embeddings* menggunakan UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) untuk mengoptimalkan proses *clustering*.
- Klustering *embedding* vektor yang sudah di reduksi dimensinya menggunakan HDBSCAN.
- Hasil dari *clustering* yang berupa Representasi topik akan diekstraksi menggunakan variasi TF-IDF berbasis kelas khusus (*custom class-based*) [8].

BERTopic

Model

*Hy*perparameter Tuning BERTopic Model dilakukan dengan metode Grid Search 2 kali simulasi dengan kombinasi parameter yang berbeda.

Grid Search Simulasi 1:

Hyperparameter	*Possible Values*
Min topic size	[50, 100, 150, 200]
Number of topics	[10, 15, 20]
Top N words	[5, 10]

BERTopic

Model

*Hy*perparameter Tuning BERTopic Model dilakukan dengan metode Grid Search 2 kali simulasi dengan kombinasi parameter yang berbeda.

Grid Search Simulasi 2:

Hyperparameter	*Possible Values*
Min topic size	[50, 150, 250]
Number of topics	[8, 12, 16]
Top N words	[10, 15]

BERTopic

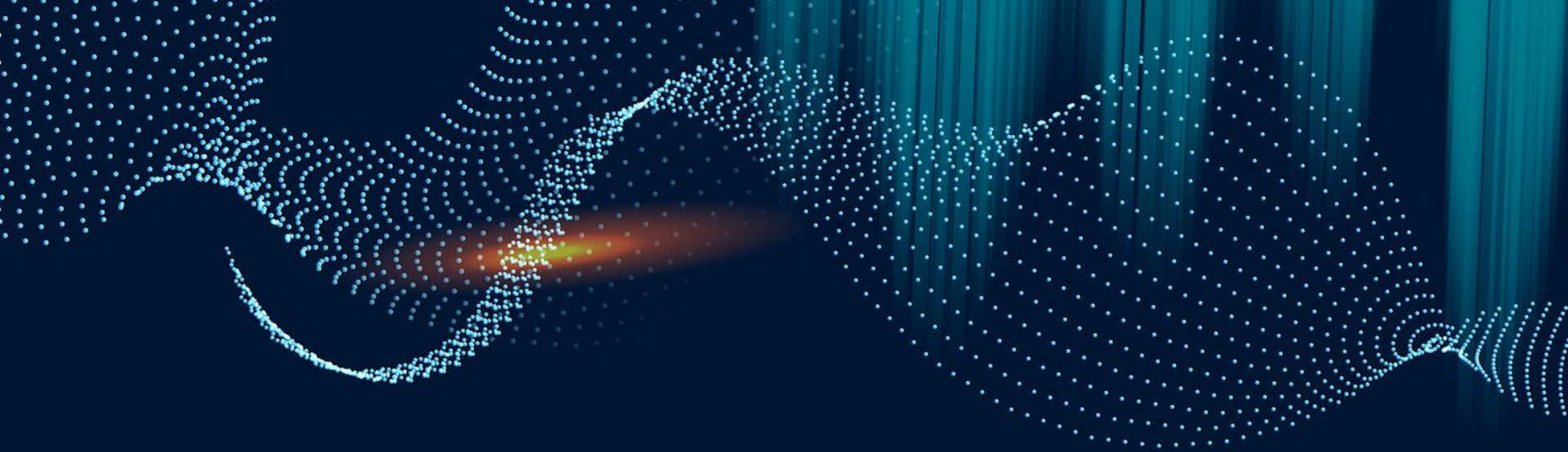
Evaluate Model

Untuk evaluasi model, kita membutuhkan ukuran kuantitatif, misal topic coherence word2vec (TC-W2V)

Misal diberikan suatu topik $\boldsymbol{t}=(t_1,t_2,....,t_N)$, maka TC-W2V dari topik $\boldsymbol{t}$ adalah:

$$TC - W2V = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} sim(t_j, t_i)$$

Dimana $sim(t_j,t_i)$ adalah *cosines similarity* antara kata t_j dan t_i pada word2vec (O'Callaghan, 2015)



03 SIMULASI dan ANALISIS HASIL

Link Colab: https://bit.ly/ColabKelompok3

Sentiment Analysis Model

Simulasi model : Menggunakan Training Set dan Validation Set dari data training, model dilatih dengan parameter teroptimal sebagai berikut ;

Hyperparameter	*Best values*
Dimensi Embedding layer	128
Menggunakan Conv1D	True
Jumlah filter convolution	32
Kernel size convolution	3
L1 regularization pada Conv1D	0.0046469
L2 regularization pada Conv1D	0.065388

Hyperparameter	Best values
Jumlah neuron pada LSTM	16
L1 regularization pada LSTM	0.0032938
L2 regularization pada LSTM	0.094458
Jumlah hidden layer dense	1
Jumlah neuron pada hidden layer dense	224
L1 regularization for dense layers	0
L2 regularization for dense layers	0
Use dropout in dense layers	True
Learning rate	0.00070334

Sentiment Analysis Model

Summary sentiment analysis model:

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 128, 128)          10460416

 conv1d (Conv1D)             (None, 126, 32)           12320

 bidirectional (Bidirectiona  (None, 32)               6272
 l)

 dense (Dense)               (None, 224)               7392

 dropout (Dropout)           (None, 224)               0

 dense_1 (Dense)             (None, 3)                 675

=================================================================
Total params: 10,487,075
Trainable params: 10,487,075
Non-trainable params: 0
_________________________________________________________________
```

Sentiment Analysis Model

Analisis Hasil

Menggunakan Test Set dari Data Training, evaluasi hasil model memprediksi label dari test set sebagai berikut ;

Label	F1-Score	Akurasi
Bad	93%	88%
Good	89%	
Neutral	79%	

BERTopic

Simulasi model

Model dengan parameter paling optimal hasil 2 kali simulasi dengan GridSearch sebagai berikut ;

Hyperparameter	Values
Min topic size	100
Number of topics	20
Top N words	10

BERTopic

Analisis Hasil

Menggunakan Data Prediksi, evaluasi hasil model sebagai berikut ;

Coherence Score	Jumlah Topik
0.194	19

BERTopic

Analisis Hasil

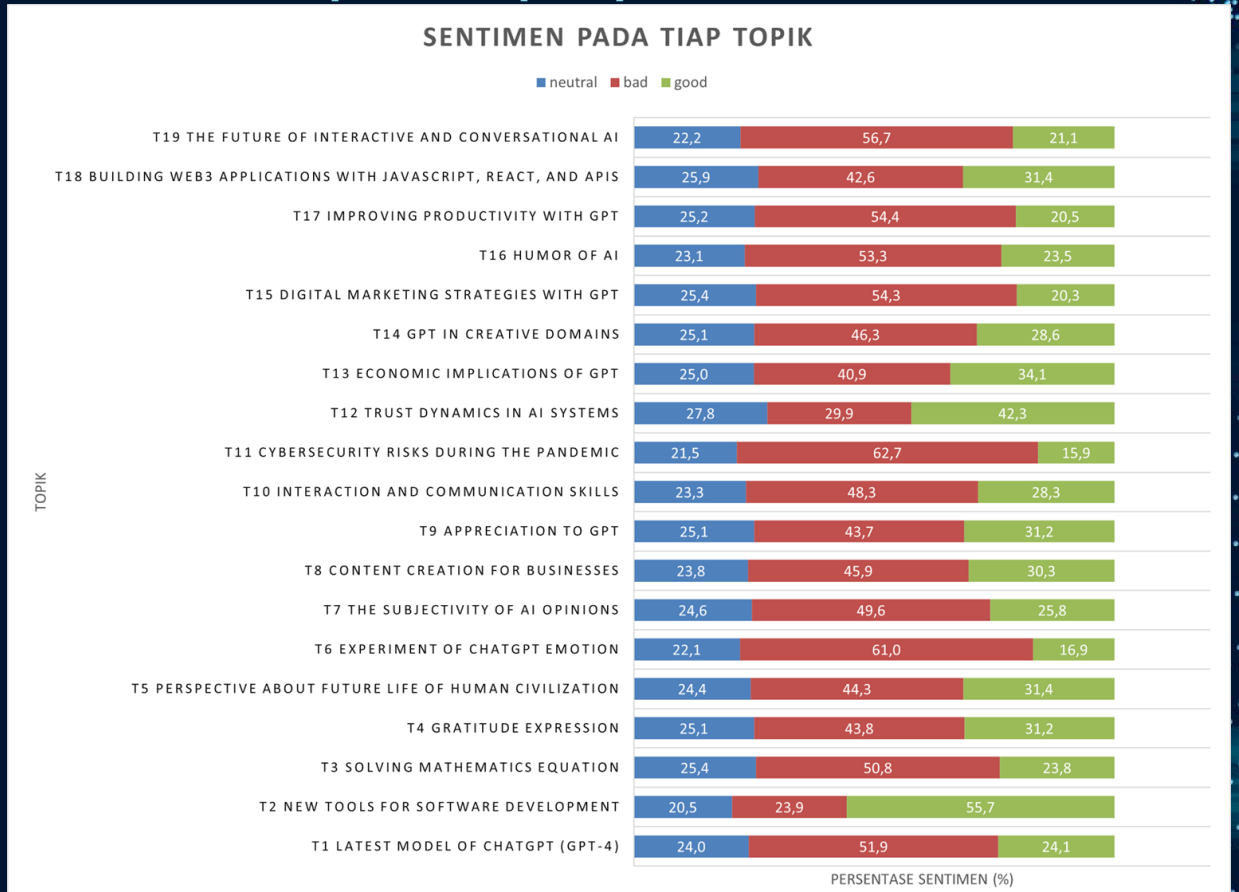
Diperoleh 19 *top words* sebagai representasi topik oleh BERTopic dari data prediksi sebagai berikut

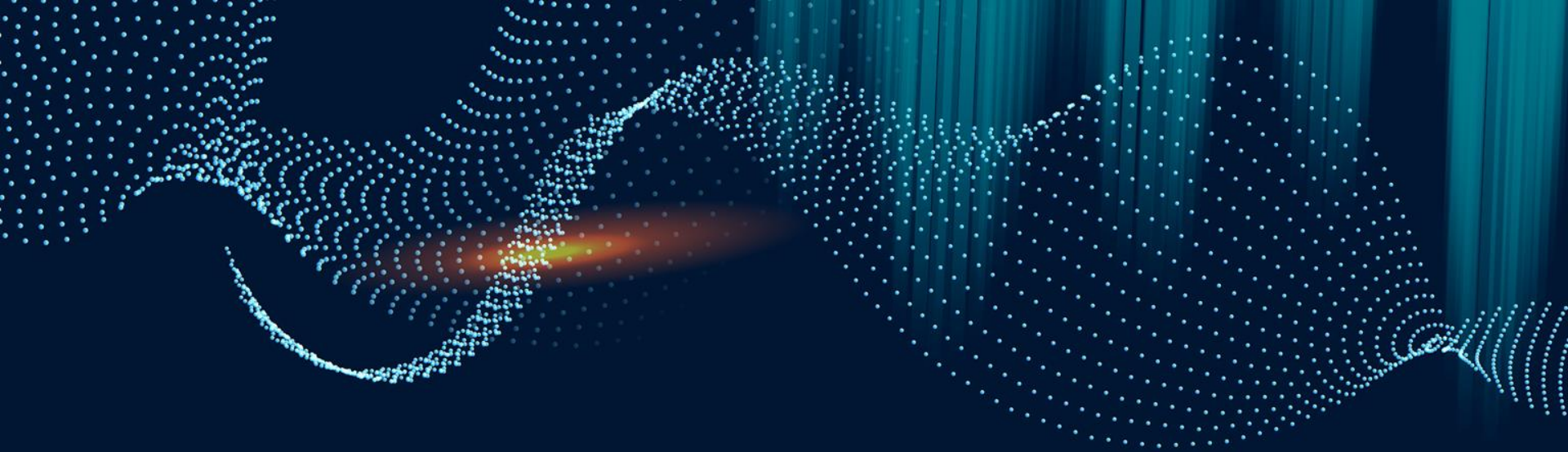
Nomor Topik	*Top Words*	Jumlah data
Topik 1	['chat', 'gpt4', 'google', 'like', 'use', 'chatgpt4', 'get', 'amp', 'using', 'new']	58968
Topik 2	['app', 'code', 'like', 'game', 'using', 'apps', 'new', 'use', 'api', 'tool']	1789
Topik 3	['math', 'answer', 'calculator', 'problem', 'question', 'student', 'using', 'solve', 'use', 'wrong']	2148
Topik 4	['love', 'thank', 'awww', 'little', 'good', 'like', 'look', 'much', 'thanks', 'dear']	3705
Topik 5	['human', 'like', 'asked', 'life', 'think', 'world', 'people', 'one', 'time', 'book']	4079
Topik 6	['light', 'one', 'gaslighting', 'like', 'night', 'asked', 'food', 'gas', 'experiment', 'would']	1719
Topik 7	['would', 'think', 'better', 'agree', 'like', 'even', 'could', 'good', 'really', 'opinion']	3116
Topik 8	['job', 'work', 'writing', 'use', 'write', 'resume', 'help', 'like', 'business', 'content']	1415
Topik 9	['congratulation', 'congrats', 'day', 'thanks', 'dear', 'result', 'today', 'week', 'good', 'im']	1317

Nomor Topik	*Top Words*	Jumlah data
Topik 10	"['apology', 'chat', 'asked', 'response', 'sorry', 'im', 'say', 'dont', 'wrong', 'said']"	3772
Topik 11	"['covid', 'vaccine', 'virus', 'threat', 'viral', 'like', 'cybersecurity', 'people', 'asked', 'get']"	983
Topik 12	"['trust', 'people', 'dont', 'like', 'think', 'know', 'would', 'could', 'one', 'say']"	1209
Topik 13	"['money', 'tax', 'pay', 'business', 'income', 'make', 'asked', 'company', 'profit', 'get']"	4018
Topik 14	"['write', 'writing', 'story', 'written', 'like', 'writer', 'wrote', 'song', 'content', 'asked']"	2807
Topik 15	"['domain', 'business', 'linkedin', 'website', 'marketing', 'content', 'name', 'prompt', 'twitter', 'like']"	1228
Topik 16	"['llm', 'meme', 'like', 'man', 'asked', 'think', 'time', 'human', 'dont', 'one']"	1050
Topik 17	"['use', 'using', 'used', 'im', 'need', 'help', 'time', 'dont', 'user', 'work']"	870
Topik 18	"['javascript', 'code', 'react', 'api', 'using', 'build', 'make', 'html', 'create', 'web3']"	1635
Topik 19	"['prompt', '2023', 'code', 'number', 'chat', 'next', 'future', 'new', '20230319', 'using']"]	1728

Hasil Akhir

Visualisasi Sentimen pada Tiap Topik





04 KESIMPULAN

Kesimpulan

- *Tweet* pada bulan Maret mengenai ChatGPT didominasi oleh topik 1 (model terbaru ChatGPT, yaitu GPT-4), diikuti oleh topik 5 (pandangan terhadap masa depan kehidupan manusia) dan topik 13 (implikasi perekonomian dari adanya GPT)
- Sebanyak 17 topik didominasi oleh sentimen negatif.
- Sebanyak 2 topik didominasi oleh sentimen positif.
- Sentimen netral dengan persentase terbesar terdapat pada topik 12 (dinamika kepercayaan terhadap sistem AI)
- Sentimen positif dengan persentase terbesar terdapat pada topik 2 (teknologi baru untuk pengembangan *software*)
- Sentimen negatif dengan persentase terbesar terdapat pada topik 11 (resiko keamanan siber pada masa pandemi)

05 DAFTAR PUSTAKA

Referensi

[1] Mubin et. al., University of Adelaide, “Exploring Sentiments of ChatGPT Early Adopters using Twitter Data.”

[2] J. Bian, K. Yoshigoe, A. Hicks, J. Yuan, Z. He, M. Xie, Y. Guo, M. Prosperi, R. Salloum, and F. Modave, “Mining twitter to assess the public perception of the “internet of things”,” PloS one, vol. 11, no. 7, p. e0158450, 2016.

[3] S. Trivedi and N. Patel, “Mining public opinion about hybrid working with roberta,” Empirical Quests for Management Essences, vol. 2, no. 1, pp. 31–44, 2022.

[4] B. Alhijawi and A. Awajan, “Prediction of movie success using twitter temporal mining,” in Proceedings of Sixth International Congress on Information and Communication Technology. Springer, 2022, pp. 105– 116.

Referensi

[5] A. Nawaz, T. Ali, Y. Hafeez, M. R. Rashid et al., “Mining public opinion: a sentiment based forecasting for democratic elections of pakistan,” Spatial Information Research, vol. 30, no. 1, pp. 169–181, 2022.

[6] S. Lock. (2022), theguardian.com, “ What is ai chatbot phenomenon chatgpt and could it replace humans?”

[7] K. Naidu. (2022), crazyengineer.in, “Chatgpt is a new ai chatbot that can find errors in your code and write you a story”

[8] Grootendorst, Maarten. (2022). “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”.


TERIMA KASIH