

## INTISARI

### PEMODELAN TOPIK *TWEET* VAKSIN COVID-19 DI INDONESIA MENGUNAKAN KLUSTER *EMBEDDINGS FROM LANGUAGE MODELS* (*ELMo*) DENGAN *CLASS-BASED* TF-IDF

Oleh

Masti Fatchiyah Maharani

20/466422/PPA/05988

Kehadiran vaksin COVID-19 ramai dibicarakan oleh berbagai kalangan masyarakat Indonesia salah satunya melalui media sosial *twitter*. Banyaknya opini dan berita yang dibagikan oleh para pengguna *twitter* menjadikannya sulit untuk menemukan topik utama yang sedang dibahas. Salah satu metode yang dapat mengatasi permasalahan ini adalah pemodelan topik.

Pemodelan topik merupakan metode yang berguna untuk menganalisis dan menemukan topik tersembunyi dalam sebuah kumpulan dokumen seperti halnya kumpulan *tweet*. Saat ini, penelitian tentang pemodelan topik pada *tweet* berbahasa Indonesia telah banyak dilakukan dengan dominasi penggunaan algoritma *Latent Dirichlet Allocation* (LDA). Namun, algoritma ini masih memiliki keterbatasan dalam menangkap unsur semantik dan konteks kalimat

Pada penelitian ini dilakukan pemodelan topik pada *tweet* berbahasa Indonesia dengan cara mengintegrasikan antara *Embeddings from Language Models* (ELMo) dengan algoritma klustering K-Means. ELMo digunakan untuk mengekstraksi fitur *tweet*. Sementara algoritma K-Means digunakan untuk mengelompokkan *tweet* berdasarkan kemiripannya. Representasi topik ditentukan dengan mencari nilai bobot kata menggunakan metode *class-based* TF-IDF. Berdasarkan hasil evaluasi topik, didapatkan model dapat melampaui performa LDA dengan skor koherensi mencapai 0.085 dan 412.55955744 detik untuk waktu eksekusi.

Kata kunci: ELMo, K-Means, *class-based* TF-IDF, pemodelan topik

## ABSTRACT

*TOPIC MODELING OF COVID-19 VACCINE-RELATED TWEET IN  
INDONESIA USING CLUSTER EMBEDDING FROM LANGUAGE MODELS  
(ELMo) WITH CLASS-BASED TF-IDF*

By

Masti Fatchiyah Maharani

20/466422/PPA/05988

Indonesians have widely discussed the COVID-19 vaccine on Twitter. The sheer number of opinions and news Twitter shares makes finding the main topics being discussed difficult. One method that can overcome this problem is topic modeling.

Topic modeling is a useful method for analyzing and finding hidden topics in collections of documents such as collections of tweets. Nowadays, research on topic modeling in Indonesian tweets has been carried out predominantly by the Latent Dirichlet Allocation (LDA) algorithm. However, this algorithm still has limitations in capturing semantic elements and the context of sentences.

In this research, topic modeling has been applied in Indonesian tweets by integrating Embeddings from Language Models (ELMo) with the K-Means clustering algorithm. ELMo is used to extract tweet features. Meanwhile, the K-Means algorithm is used to cluster tweets based on their similarities. Topic representation is determined by finding the word weight value using the class-based TF-IDF method. Based on the topic evaluation results, the model shows the outperformers of LDA based on coherence scores of 0.085 and 412.55955744 seconds for execution time.

Keywords: ELMo, K-Means, *class-based TF-IDF*, *topic modeling*