INTISARI

PEMODELAN TOPIK PENELITIAN BIDANG ILMU KOMPUTER DI INDONESIA MENGGUNAKAN *LATENT DIRICHLET ALLOCATION* (LDA) DAN *PART OF SPEECH FILTERING*

Kartika Rizqi Nastiti
21/485636/PPA/06220

Riset di bidang ilmu komputer yang dilakukan oleh peneliti Indonesia dari berbagai afiliasi semakin beragam dari tahun ke tahun dan melibatkan banyak disiplin ilmu lainnya, namun di sisi lain, tren topik penelitian pun juga semakin sulit untuk diketahui. Salah satu teknik untuk menyingkap tren topik adalah teknik pemodelan topik menggunakan algoritma *Latent Dirichlet Allocation* (LDA), namun metode ini masih memiliki masalah yang memungkinkan dihasilkannya *noise* yang dapat mempengaruhi kualitas topik yang dihasilkan.

Penelitian ini berfokus untuk melakukan efisiensi korpus sebagai input pada model topik LDA dengan melakukan filtrasi *Part of Speech* (POS) *Tagging* untuk mendapatkan korpus yang hanya terdiri dari kelas kata *noun* saja. Alur penelitian yang dilakukan dimulai dari pengumpulan data, *preprocessing,* pembentukan *n-gram* dan *filtering* berdasarkan POS *tagging.* Hasil token Filtered (hanya *noun*) dan Nonfiltered kemudian direpresentasikan menggunakan empat metode yang berbeda, yaitu *Term Frequency – Inverse Document Matrix* (TF-IDF), TF-IDF + fastText, TF-IDF + *Bidirectional Encoder Representations from Transformers* (BERT), dan TF-IDF + fastText + BERT. Selanjutnya dilakukan pembentukan model topik LDA untuk tiap korpus yang dihasilkan lalu dievaluasi dengan *topic coherence.*

Berdasarkan hasil pengujian menggunakan data judul penelitian bidang ilmu komputer selama 10 tahun (2012-2021) yang dilakukan oleh peneliti Indonesia dari 3 afiliasi, yaitu Universitas Gadjah Mada (UGM), Universitas Indonesia (UI), dan Institut Teknologi Bandung (ITB), diperoleh hasil bahwa korpus Filtered (hanya *noun*) memberikan nilai *coherence* yang lebih baik dibandingkan korpus Filtered (semua kelas kata). Korpus dengan representasi TF-IDF + BERT Filtered memberikan nilai *coherence* tertinggi, yaitu 0.533 menggunakan parameter *passes* = 10, *iterations* = 50, dan *num_topics* = 9. Hasil pemodelan pada penelitian ini menunjukkan 3 topik riset ilmu komputer yang paling banyak diangkat adalah "Data and Information Management", "Information Security", dan "Data Analytics and Machine Learning".

Kata Kunci: Pemodelan Topik, *Latent Dirichlet Allocation,* POS *Tagging,* TF-IDF, fastText, BERT, *Coherence*

ABSTRACT

TOPIC MODELING ON COMPUTER SCIENCE RESEARCHES IN INDONESIA USING LATENT DIRICHLET ALLOCATION (LDA) AND PART OF SPEECH FILTERING

Kartika Rizqi Nastiti
21/485636/PPA/06220

Research in the field of computer science conducted by Indonesian researchers from various affiliations has become increasingly diverse from year to year, involving many other disciplines or domains. However, on the other hand, the trends in research topics have also become increasingly difficult to discern. One technique for uncovering topic trends is topic modeling using the Latent Dirichlet Allocation (LDA) algorithm. However, this method still has issues that may result in noise that can affect the quality of the topics produced.

This study focused on carrying out corpus efficiency as input to the LDA topic model by filtering Part of Speech (POS) Tagging to get a corpus that only consists of noun word classes. The research flow started from data collection, preprocessing, n-gram formation and filtering based on POS tagging. The results of Filtered (noun only) and Nonfiltered tokens were then represented using four different methods; Term Frequency – Inverse Document Matrix (TF-IDF), TF-IDF + fastText, TF-IDF + Bidirectional Encoder Representations from Transformers (BERT), and TF-IDF + fastText + BERT. Furthermore, the formation of the LDA topic model for each corpus produced was then evaluated with topic coherence.

Based on the results of evaluation using research title data in the field of computer science for 10 years (2012-2021) conducted by Indonesian researchers from 3 affiliations; Gadjah Mada University (UGM), University of Indonesia (UI), and Bandung Institute of Technology (ITB), the result was that Filtered corpus (only nouns) gave a better coherence value than Filtered corpus (all word classes). The corpus with TF-IDF + BERT Filtered representation obtained the highest coherence value, which was 0.533 using passes = 10 parameters, iterations = 50, and num_topics = 9. The modeling results in this study show that the 3 computer science research topics that are most frequently raised were "Data and Information Management", "Information Security", and "Data Analytics and Machine Learning".

Keywords: *Topic Modeling, Latent Dirichlet Allocation, POS Tagging, TF-IDF, fastText, BERT, Coherence*