INTISARI

PEMODELAN TOPIK MENGGUNAKAN LATENT DIRICHLET ALLOCATION DAN NON-NEGATIVE MATRIX FACTORIZATION PADA ABSTRAK SKRIPSI STATISTIKA FMIPA UGM

Oleh

Billyarta Wasthu Driantama
19/445715/PA/19539

Seiring dengan kemajuan teknologi, dunia telah memasuki masa ketika akses dan kontrol informasi merupakan karakteristik dari peradaban manusia. Hal tersebut menyebabkan peningkatan data sehingga muncul era *big data*. Saat ini, data tersedia dalam jumlah besar tetapi masih belum dimanfaatkan secara optimal. Salah satu contohnya yakni koleksi informasi seperti perpustakaan digital yang berisi buku, jurnal, serta karya tugas akhir para mahasiswa. Hampir setiap tahun mahasiswa menghasilkan karya tugas akhir layaknya skripsi, tesis, dan disertasi. Bahkan, dokumen tersebut terus menumpuk setiap tahun tanpa adanya pengolahan secara berkala. Salah satu metode yang populer yakni pemodelan topik, memiliki kemampuan untuk menganalisis dokumen teks dan mengidentifikasi topik utama yang terkandung di dalamnya. Penggunaan karya tugas akhir untuk pemodelan topik dapat menjadi salah satu terobosan. Penelitian ini menggunakan studi kasus abstrak skripsi mahasiswa program studi S1 Statistika FMIPA UGM pada tahun 2014-2022 dari web ETD UGM: Theses and Dissertations Repository. Digunakan pemodelan topik Latent Dirichlet Allocation (LDA) dan Non-negative Matrix Factorization (NMF) serta pembobotan kata Bag of Words (BoW) dan Term Frequency-Inverse Document Frequency (TF-IDF). Diperoleh bahwa evaluasi model terbaik yakni pemodelan topik NMF serta pembobotan kata TF-IDF menghasilkan nilai Coherence sebesar 0,6825 dan memiliki interpretabilitas yang baik. Hasil topik sebanyak 8 topik berdasarkan urutan proporsi tertinggi yaitu Topik 8: Statistical Machine Learning (45,0%), Topik 6: Variasi Model Regresi (12,9%), Topik 2: Estimasi Value at Risk dan Penentuan Harga Opsi (10,2%), Topik 5: Metode Pengendalian Kualitas (10,0%), Topik 7: Optimisasi Portofolio (8,2%), Topik 4: Estimasi Cadangan Klaim dan Perhitungan Premi (6,5%), Topik 3: Metode Statistika Klasik (3,8%), serta Topik 1: Analisis Valuasi Obligasi (3,3%).

Kata-kata kunci: pemodelan topik, LDA, NMF, pembobotan kata, BoW, TF-IDF, abstrak skripsi.

ABSTRACT

TOPIC MODELING USING LATENT DIRICHLET ALLOCATION AND NON-NEGATIVE MATRIX FACTORIZATION IN ABSTRACT THESIS STATISTICS FMIPA UGM

By

Billyarta Wasthu Driantama
19/445715/PA/19539

Along with technological advances, the world has entered a period when access to and control of information is a characteristic of human civilization. This led to an increase in data so that the big data era emerged. Currently, large amounts of data are available but are still not optimally utilized. One example is a collection of information such as a digital library that contains books, journals, and students' final assignments. Almost every year students produce final project works such as bachelor theses, theses, and dissertations. In fact, these documents continue to accumulate every year without regular processing. One popular method, topic modeling, has the ability to analyze text documents and identify the main topics contained therein. The use of final project work for topic modeling can be one of the breakthroughs. This research will use case studies of thesis abstracts of undergraduate students of the Statistics major of FMIPA UGM in 2014-2022 from the website ETD UGM: Theses and Dissertations Repository. Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) topic modeling methods are used as well as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) term weighting. It was found that the evaluation of the best model, namely NMF topic modeling and TF-IDF term weighting, resulted in a Coherence score of 0.6825 and had good interpretability. Topic results as many as 8 topics based on the order of the highest proportion, namely Topic 8: Statistical Machine Learning (45.0%), Topic 6: Variation of Regression Models (12.9%), Topic 2: Estimation of Value at Risk and Option Pricing (10.2%), Topic 5: Quality Control Methods (10.0%), Topic 7: Portfolio Optimization (8.2%), Topic 4: Claim Reserve Estimation and Premium Calculation (6.5%), Topic 3: Classical Statistical Methods (3.8%), and Topic 1: Bond Valuation Analysis (3.3%).

Keywords: topic modeling, LDA, NMF, term weighting, BoW, TF-IDF, thesis abstract.