ABSTRAK

Nama : Miftakhul Huda

NIM : 1147010040

Judul Skripsi : Implementasi Algoritma Latent Dirichlet Allocation pada Data Teks Terjemah Hadits Bahasa Inggris

Metode penambangan teks (*text mining*) dalam bidang ilmu komputer sangat berkembang. Terdapat banyak metode yang sudah diperkenalkan mulai dari TF-IDF (*term frequency-index document frequency*), LSI (*Latent Semantic Indexing*), LDA (*Latent Dirichlet Allocation*) dan masih banyak lagi metode lainnya. Pada penelitian ini penulis akan membahas mengenai LDA. LDA merupakan metode ekstraksi fitur berbasis topik. LDA akan diimplementasikan dengan membuat aplikasi menggunakan perangkat lunak Python. Dokumen yang digunakan merupakan terjemah hadits bahasa inggris sebanyak 903 dokumen. Proses pengolahan melibatkan *tokenizing, stemming, filtering* dan ekstraksi topik. Hasil ekstraksi berupa nilai probabilitas topik pada kata dan probabilitas topik pada dokumen. Setelah itu, probabilitas topik kata diklasifikasi dan probabilitas topik dokumen diklasifikasikan berdasarkan probabilitas terbesar. Kemudian hasil klasifikasi dievaluasi dari topik aslinya menggunakan *precision, recall, accuracy* dan *f-measure*.

Kata kunci: latent dirichlet allocation, ektraksi topik, hadits

ABSTRACT

Name **: Miftakhul Huda**

NIM **: 1147010040**

Title **: *Implementation of the Latent Dirichlet Allocation Algorithm on Text Data English Hadith Translate***

Text mining methods in computer science are highly developed. there's many methods have been introduced start from TF-IDF (term frequency-index document frequency), LSI (Latent Semantic Indexing), LDA (Latent Dirichlet Allocation) and many other methods. In this research, the writer will discuss about LDA. LDA is a topic-based feature extraction method. LDA will be implemented by creating applications using Python software. The documents used are english hadith translate totaling 903 documents. The processing involves tokenizing, stemming, filtering and topic extraction. Result of extraction are word-topic probability and document-topic probability. After that, word-topic probability classified and document-topic probability classified based on the most probability. Then, the classification results are evaluated from the original topic using precision, recall, accuracy and f-measure

Keyword: *latent dirichlet allocation, topic extraction, hadith*