INTISARI

ANALISIS *TOPIC MODELING* MENGGUNAKAN LDA DAN HDP DENGAN KOMBINASI *WORD EMBEDDING* DAN K-MEANS *CLUSTERING* DARI ABSTRAK PUBLIKASI

Oleh:

Nurfirsta Idrus

22/510206/PPA/06470

Penelitian dan publikasi memegang peranan penting dalam dunia akademik dan kemajuan sebuah perguruan tinggi, dengan semakin banyaknya penelitian dan publikasi yang dihasilkan, terkadang sulit memahami keragaman topik yang dicakup. Pendekatan yang bisa dilakukan untuk mengidentifikasi topik penelitian ialah *topic modeling*.

Penelitian ini menggunakan *topic modeling* untuk mengidentifikasi topik dari abstrak publikasi FMIPA UGM. Dua metode yang digunakan adalah *Latent Dirichlet Allocation* (LDA) dan *Hierarchical Dirichlet Process* (HDP). LDA sering mengalami masalah sparsitas data pada teks pendek seperti abstrak, sehingga sulit menentukan hubungan spesifik antar topik. Sementara itu, HDP memungkinkan pembentukan hierarki topik yang lebih kompleks dan menghasilkan topik yang lebih bervariasi dibandingkan LDA. Keduanya menggunakan pendekatan *Bag of Words* (BoW) yang mengabaikan urutan dan semantik kata, sehingga penelitian ini mengusulkan model Word2Vec dan FastText sebagai teknik pembobotan kata untuk mengtasi keterbatasan LDA dan HDP, khususnya dalam hal *clustering* dokumen pada topik.

Hasil penelitian menunjukkan model HDP menghasilkan *coherence score* tertinggi (0.7035) dibanding LDA (0.4837). Selan itu, data tanpa *stemming* cenderung memberikan hasil yang lebih baik dan bermakna dibanding dengan *stemming*. Untuk *clustering* dokumen pada topik, terdapat 6 model yang dibandingkan, yaitu model LDA dan HDP *baseline* dengan model kombinasi hasil topik LDA dan HDP dengan Word2Vec dan FastText. Hasilnya menunjukkan model LDA+FastText+K-Means meraih *silhouette score* tertinggi sebesar 0.5508 dan terendah pada model HDP+K-Means dengan skor 0.2562. Model berbasis Word2Vec dan FastText mendapatkan nilai silhouette tertinggi dan menghasilkan clustering yang lebih terpisah, menunjukkan peningkatan kualitas dibandingkan model LDA dan HDP dengan pendekatan konvensional.

Kata Kunci : *topic modeling*, *word embedding*, LDA, HDP, Word2Vec, FastText

ABSTRACT

ANALYSIS OF TOPIC MODELING USING LDA AND HDP WITH COMBINATION OF WORD EMBEDDING AND K-MEANS CLUSTERING ON PUBLICATION ABSTRACTS

By:

Nurfirsta Idrus

22/510206/PPA/06470

Research and publications play an important role in the academic world and the progress of a university, with the increasing number of research and publications produced, it is sometimes difficult to understand the diversity of topics covered. An approach that can be done to identify research topics is topic modelling.

This research uses topic modelling to identify topics from abstracts of FMIPA UGM publications. Two methods used are Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP). LDA often suffers from data sparsity problems in short texts such as abstracts, making it difficult to determine specific relationships between topics. Meanwhile, HDP allows the formation of more complex topic hierarchies and produces more varied topics than LDA. Both use the Bag of Words (BoW) approach that ignores word order and semantics, so this study proposes the Word2Vec and FastText models as word weighting techniques to overcome the limitations of LDA and HDP, especially in terms of clustering documents on topics.

The results showed that the HDP model produced the highest coherence score (0.7035) compared to LDA (0.4837). In addition, data without stemming tends to give better and more meaningful results than those with stemming. For document clustering on topics, 6 models were compared, namely the baseline LDA and HDP model with the combination model of LDA and HDP topic results with Word2Vec and FastText. The results show that the LDA+FastText+K-Means model achieves the highest silhouette score of 0.5508 and the lowest is the HDP+K-Means model with a score of 0.2562. The Word2Vec and FastText-based model achieved the highest silhouette score and produced more separated clusters, showing an improvement in quality over the LDA and HDP models with conventional approaches.

Keywords : topic modeling, word embedding, LDA, HDP, Word2Vec, FastText