ABSTRAK

Pertumbuhan teknologi menjadikan internet sebagai tempat masyarakat mengekspresikan opini mereka. Salah satunya dalam bentuk ulasan produk kecantikan yang dibagikan melalui *platform* digital. Forum Female Daily adalah contoh *platform* populer di dunia kecantikan yang menyediakan ribuan ulasan. Banyaknya ulasan yang diunggah pengguna menghadirkan tantangan tersendiri dalam memahami opini dan tren secara menyeluruh. Oleh karena itu, dibutuhkan pendekatan otomatis, seperti pemodelan topik yang mampu menyaring informasi dari ulasan dalam skala besar. Beberapa studi telah membandingkan metode pemodelan topik untuk menemukan pendekatan terbaik dalam menganalisis opini pengguna. Salah satunya adalah penelitian oleh Egger & Yu (2022) yang menunjukkan bahwa NMF unggul dalam menghasilkan topik yang tidak tumpang tindih, mudah diinterpretasikan, dan cocok untuk teks pendek seperti ulasan. Di sisi lain, BERTopic dinilai mampu menangkap wawasan baru melalui pendekatan *embedding* berbasis transformer dan menawarkan fleksibilitas tinggi dalam pemodelan topik. Kedua metode tersebut menunjukkan performa yang kompetitif dan mewakili dua pendekatan yang berbeda, yaitu pendekatan konvensional berbasis aljabar (NMF) dan pendekatan modern berbasis semantik (BERTopic). Pemilihan keduanya dalam penelitian ini juga didasarkan pada relevansi metode terhadap karakteristik data ulasan skincare yang pendek, subjektif, dan ekspresif (Egger & Yu, 2022b).

Penelitian ini menggunakan 2.500 ulasan pengguna terhadap produk Avoskin Miraculous Refining Toner yang dikumpulkan melalui teknik *web scraping* dari forum Female Daily Talk. Data ulasan diproses melalui tahapan praproses teks yang sama untuk kedua metode, mencakup *data cleaning, case folding, tokenization*, *normalization, stopwords removal,* dan *stemming*. Pemodelan dengan BERTopic dilakukan melalui *embedding* dokumen menggunakan *Sentence*-BERT, reduksi dimensi dengan UMAP, *clustering* menggunakan HDBSCAN, dan representasi topik dengan pendekatan class-based TF-IDF. Sementara itu, pemodelan NMF dilakukan melalui pembobotan TF-IDF dan dekomposisi matriks. Evaluasi performa model dilakukan dengan dua metrik utama, yaitu *topic coherence* dan *topic diversity*. Sistem juga dikembangkan dalam bentuk antarmuka berbasis web menggunakan Streamlit untuk mendukung proses visualisasi dan eksplorasi hasil topik.

Hasil evaluasi menunjukkan bahwa BERTopic memiliki skor *topic coherence* yang lebih tinggi dibandingkan NMF. Hal ini menunjukan topik-topik yang dihasilkan BERTopic memiliki keterkaitan semantik antar kata yang lebih kuat. Selain itu, nilai *topic diversity* pada BERTopic lebih tinggi dibandingkan dengan NMF. Hal ini menunjukkan BERTopic mampu menghasilkan topik yang lebih bervariasi dan tidak tumpang tindih.

Kata Kunci: BERTopic, NMF, pemodelan topik*,* TF-IDF*, Sentence*-BERT

ABSTRACT

The growth of technology has made the internet a place for people to express their opinions. One of them is beauty product reviews shared through digital platforms. The Female Daily forum is an example of a popular platform in the beauty world that provides thousands of reviews. The sheer number of user-uploaded reviews makes it challenging to understand opinions and trends thoroughly. Therefore, there is a need for automated approaches, such as topic modeling, that are able to filter information from reviews on a large scale. Several studies have compared topic modeling methods to find the best approach for analyzing user opinions. One of them is a study by Egger & Yu (2022) *which shows that NMF outperforms in generating topics that are non-overlapping, easy to interpret, and suitable for short texts such as reviews. On the other hand, BERTopic is considered capable of capturing new insights through a transformer-based embedding approach and offers high flexibility in topic modeling. Both methods show competitive performance and represent two different approaches, which are the conventional algebra-based approach (NMF) and the modern semantic-based approach (BERTopic). Their selection in this study is also based on their relevance to the short, subjective and expressive characteristics of skincare review data.*

This research uses 2,500 user reviews of Avoskin Miraculous Refining Toner products collected through web scraping techniques from the Female Daily Talk forum. The review data was processed through the same text preprocessing stages for both methods, including data cleaning, case folding, tokenization, normalization, stopwords removal, and stemming. Modeling with BERTopic is performed through document embedding using Sentence-BERT, dimension reduction with UMAP, clustering using HDBSCAN, and topic representation with class-based TF-IDF. Meanwhile, NMF modeling is performed through TF-IDF weighting and matrix decomposition. The model performance evaluation is conducted using two main metrics, namely topic coherence and topic diversity. The system is also developed in the form of a web-based interface using Streamlit to support the process of visualization and exploration of topic results.

The evaluation results show that BERTopic has a higher topic coherence score than NMF. This indicates that the topics generated by BERTopic have stronger semantic relationships between words. In addition, the topic diversity value in BERTopic is higher than that in NMF. This shows that BERTopic is capable of generating more varied topics that do not overlap.

Keywords:*, BERTopic, NMF, topic modelling, TF-IDF. Sentence-BERT*