ABSTRACT

Crime analysis is a methodical approach to identifying and analyzing patterns and trends in crime. By using crime analysis and text mining, we can analyze the modus operandi of crimes. One of the most important stages of text mining is feature extraction. In text mining, feature extraction can select important words and rank the importance of words. Examples of this feature extraction are TF-IDF and BM25. These two methods are often found in the information retrieval but have begun to be developed into the realm of short text. BM25 is claimed to be better than the TF-IDF. Therefore, it will be proved by comparing these two methods. To get the optimal value, the BM25 will be adjusted using a manual search for tuning hyperparameter values for *k1* and *b*. Then one-dimensional clustering will be carried out with Jenks Natural Break and the GVF value will be calculated.

TF-IDF and BM25 as feature extraction methods in text mining to determine the importance of a word. Both of these methods use the IDF function, however, the IDF used in these two methods is different. The data used is data from social media Twitter with 39,964 data that has the keyword “penipuan” in Indonesian. The previous data was cleaned through preprocessing, then the sparse term was removed, then the TF-IDF and BM25 feature extraction and tuning hyperparameter of BM25 were carried out, then the words were grouped using Jenks Natural Break and evaluated using the Goodness Variance of Fit. From this experiment, it was found that the value of GVF is directly proportional to the value of the parameter *k1* on BM25. The highest GVF value is in the BM25 hyperparameter tuning with a value of *k1*=3.2 which the value is 0,.9176. This conclude that valie of GVF of Jenks Natural Break in BM25 feature extraction of Twitter data with hyperparameter tuning reach an optimum value with increasing *k1*.

Keywords : bm25, tfidf, *hyperparameter*, fraud, twitter.

INTISARI

Analisis kejahatan adalah pendekatan metodis untuk mengidentifikasi dan menganalisis pola dan tren dalam kejahatan. Dengan menggunakan analisis kejahatan dan *text mining*, kita dapat menganalisis modus operandi kejahatan. Salah satu tahapan *text mining* yang sangat penting adalah ekstraksi fitur. Dalam *text mining,* ekstraksi fitur dapat menyeleksi kata penting dan memberikan peringkat pentingnya kata. Contoh dari ekstraksi fitur ini adalah TF-IDF dan BM25. Kedua metode ini sering dijumpai pada sistem temu kembali tetapi mulai dikembangkan ke ranah *short text*. BM25 diklaim lebih baik dari TF-IDF. Oleh karena itu, akan dibuktikan dengan cara membandingkan kedua metode ini. Untuk mendapatkan nilai optimal, maka BM25 akan dilakukan penyetelan *hyperparameter* dengan *manual search* untuk nilai *k1* dan *b*. Kemudian akan dilakukan klasterisasi satu dimensi dengan *Jenks Natural Break* dan dihitung nilai GVF.

TF-IDF dan BM25 sebagai metode ekstraksi fitur pada *text mining* untuk mengetahui pentingnya suatu kata. Kedua metode ini menggunakan fungsi IDF tetapi, IDF yang digunakan pada kedua metode ini berbeda. Data yang digunakan adalah data dari media sosial Twitter dengan 39.964 data yang memiliki kata kunci "penipuan" dalam bahasa Indonesia. Data sebelumnya dibersihkan melalui praproses, lalu penghilangan *sparse term*, kemudian dilakukan ekstraksi fitur TF-IDF dan BM25 serta penyetelan *hyperparameter* BM25, selanjutnya kata dikelompokkan menggunakan *Jenks Natural Break* dan di evaluasi menggunakan fungsi *Goodness Variance of Fit*. Dari eksperimen ini, ditemukan bahwa nilai GVF berbanding lurus pada nilai parameter *k1* pada BM25. Nilai tertinggi GVF ada pada penyetelan *hyperparameter* BM25 dengan nilai *k1*=3.2 yaitu 0,9176. Hal ini menunjukkan bahwa nilai GVF pada *Jenks Natural Break* dalam metode ekstraksi fitur BM25 pada data Twitter dengan penyetelan *hyperparameter* mencapai nilai optimal dengan menaikan nilai *k1*.

Kata kunci – bm25, tfidf, *hyperparameter*, penipuan, twitter.