ABSTRAK

Pencarian informasi secara konvensional kurang memberikan kemudahan, seperti penyampaian informasi di Jurusan Informatika UPN "Veteran" Yogyakarta melalui surat edaran, media sosial dan media lainnya. Seringkali pertanyaan mahasiswa ditanyakan secara berulang dan terbatas karena jam kerja. Oleh karena itu, salah satu solusi teknologi yang dapat menciptakan komunikasi interaktif, cepat dan informatif yaitu *chatbot* dengan pendekatan NLP berbasis *retrieval*. Pengenalan *intent* atau klasifikasi pertanyaan merupakan salah satu tahapan penting NLP berbasis *retrieval*. Klasifikasi biasanya dilakukan menggunakan metode *machine learning*, salah satunya yaitu dengan *K-Nearest Neighbor* (KNN). Namun KNN menghabiskan waktu relatif lama dan kurang akurat dalam perhitungan jarak, karena ruang vektor sparse (memiliki banyak nol) dengan dimensi yang besar, serta keterbatasan dalam memperhatikan semantik. Maka, untuk mengatasi kelemahan KNN dibutuhkan sebuah algoritma pemilihan fitur yang mampu memperhatikan semantik. *Latent Semantic Analysis* (LSA) merupakan salah satu algoritma pemelihan fitur dengan reduksi fitur yang memetakan ruang fitur ke ruang tereduksi.

Penelitian akan dilakukan untuk klasifikasi pertanyaan mahasiswa kedalam *intent* pertanyaan berdasarkan topik informasi yang ada di Jurusan Informatika UPN "Veteran" Yogyakarta. Penelitian ini menggunakan data yang diperoleh pengumpulan data pertanyaan mahasiswa dengan formullir, dimana data tersebut diberi label sesusai topiknya. Kemudian data akan dilakukan text preprocessing, ekstraksi fitur dengan *Term Frequency-Inverse Document Frequency* (TF-IDF), pemilihan fitur LSA, selanjutnya proses *training* dan *testing* dengan menggunakan algoritma KNN. Pengujian yang dilakukan menggunakan data sebanyak 1410 dengan perbandingan data latih 90% dan data uji 10% dengan *K-Fold Cross Validation*.

Penerapan LSA pada klasifikasi pertanyaan membuat klasifikasi pertanyaan lebih efisien waktu, serta meningkatkan akurasi, presisi dan *recall* dari algoritma KNN dengan menghilangkan fitur yang berlebihan, menghubungkan antar pertanyaan dan melihat relasi antar kata. Hasil pengujian yang telah dilakukan dapat diketahui bahwa terjadi efisiensi waktu ditunjukan dengan waktu komputasi yang semula 27.49 detik menjadi 17,97 detik, peningkatan nilai rata-rata akurasi dari 92,1% menjadi 93,2%, peningkatan rata-rata presisi dari 76,7% menjadi 78,8%, peningkatan rata-rata *recall* dari 74.9% menjadi 77,9%. Model ini juga mengalami peningkatan nilai rata-rata AUC dari 85% *(good classification)* menjadi 87% (*good classification*).

ABSTRACT

Conventional search for information does not provide convenience, such as the delivery of information at the Department of Informatics UPN "Veteran" Yogyakarta through circulars, social media and other media. Often student questions are asked repeatedly and are limited due to working hours. Therefore, one of the technology solutions that can create interactive, fast and informative communication is a chatbot with a retrieval-based NLP approach. The introduction of intent or question classification is one of the important stages of retrieval-based NLP. Classification is usually done using machine learning methods, one of which is K-Nearest Neighbor (KNN). However, KNN takes a relatively long time and is less accurate in calculating distances, because the sparse vector space (has many zeros) with large dimensions, and limitations in paying attention to semantics. So, to overcome the weakness of KNN, we need a feature selection algorithm that is able to pay attention to semantics. Latent Semantic Analysis (LSA) is a feature selection algorithm with feature reduction that maps the feature space to the reduced space.

Research will be conducted to classify student questions into question intent based on the topic of information in the Department of Informatics UPN "Veteran" Yogyakarta. This study uses data obtained by collecting student question data using a form, where the data is labeled according to the topic. Then the data will be text preprocessing, feature extraction with Term Frequency-Inverse Document Frequency (TF-IDF), LSA feature selection, then the training and testing process using the KNN algorithm. The tests were carried out using 1410 data with a comparison of 90% training data and 10% test data with K-Fold Cross Validation.

The application of LSA to the classification of questions makes the classification of questions more time efficient, and improves the accuracy, precision and recall of the KNN algorithm by eliminating redundant features, connecting between questions and seeing the relationships between words. The results of the tests that have been carried out can be seen that there is time efficiency as indicated by the computation time from 27.49 seconds to 17.97 seconds, an increase in the average value of accuracy from 92.1% to 93.2%, an increase in the average precision from 76,7% to 78.8%, an increase in the average recall from 74.9% to 77.9%. This model also experienced an increase in the average AUC value from 85% (good classification) to 87% (good classification).