ABSTRAK

Nama : Yulianti Gusniar

Nim : 1157010070

Judul : *Part of Speech Tagging* Al-Qur'an Menggunakan Kombinasi Dengan Metode *K-Nearest Neighbor* dan *Naive Bayes*

Dalam linguistik, penandaan kelas kata (bahasa inggris: *Part of Speech Tagging* atau disingkat POST) adalah proses penandaan kata pada suatu teks (korpus) dalam kaitannya dengan suatu kelas kata tertentu berdasarkan definisi dan maknanya-hubungannya dengan kata yang mendampingi atau yang terkait dengannya pada suatu frasa, kalimat dan paragraf. Ini adalah proses pengelompokan setiap kata ke dalam bagian penandaan yang sesuai untuk konteks tertentu. Kerumitan teks bahasa Arab Al-Qur'an yang kini diketahui menyebabkan berbagai model penandaan POS berperfoma buruk dalam bahasa Arab Al-Qur'an. Akhirnya, ini akan menimbulkan sejumlah kesulitan untuk penandaan POS, termasuk ambiguitas yang signifikan, sedikitnya data, dan banyak kata tak dikenal. Perhatian utama di sini adalah untuk memastikan bagaimana pendekatan efektif berfungsi dalam bahasa Arab dan bagaimana korpus Qur'an dapat digunakan untuk membuat kerangka kerja yang efektif untuk penandaan POS Arab. Di sini peneliti akan menggabungkan dua metode statistik pengklasifikasi, yang pertama metode *K-Nearest Neighbor* (KNN) dan yang kedua metode *Naïve Bayes* (NB) dengan cara eksperimental untuk penandaan POS bahasa Arab. Lalu pendekatan kombinasi yang disebut Algoritma Voting Mayoritas (AVM) digunakan untuk memanfaatkan keunggulan pengklasifikasi. Peneliti telah menyelidiki banyak fitur untuk menentukan mana yang berguna dan bagaimana pengaruhnya terhadap kinerja penandaan POS Arab Qur'an. Oleh karena itu, tujuan dari penelitian ini adalah untuk menentukan set fitur mana yang berdampak dan untuk menstandarkan metode penandaan POS yang lebih tepat. Teks Al-Qur'an digunakan sebagai sumber data penelitian, dengan menggunakan Surah Al-Baqarah yang memiliki 6115 kata. Pendekatan metode *Naïve Bayes*

yang menghasilkan nilai akurasi maksimal yaitu 54,50%. Elemen yang paling efektif menghasilkan akurasi ini adalah P_0 (POS dari kata saat ini), W_{-3} (Kata dari tiga kata sebelumnya), P_0 (POS dari kata saat ini), P_1 (POS dari kata sesudahnya), S_1, S_n, Huruf Besar, dan Berisi Nomor.

Kata Kunci: Algoritma Voting Mayoritas, Klasifikasi, *K-Nearest Neighbor*, *Naïve Bayes, Part of Speech Tagging*, Pemrosesan Bahasa Alami.

ABSTRACT

Name **: Yulianti Gusniar**

Nim **: 1157010070**

Title **: *Part of Speech Tagging* Al-Qur'an Using A Combination With The *K-Nearest Neighbor* and *Naive Bayes Methods***

In linguistics, tagging word classes (English: *Part Of Speech Tagging* or abbreviated as POST) is the process of tagging words in a text (corpus) in relation to a particular class of words based on their definition and meaning-relationship with words that accompany or are related to them in a phrase, sentence or paragraph. This is the process of grouping individual words into appropriate signification sections for a particular context. The complexity of the Arabic Qur'anic text which is now known to cause various POS Tagging models to perfom poorly in Qur'anic Arabic. Finally, this will intoduce a number of difficulties for POS tagging, including significant ambiguity, little data, and many unknown words. The main concern here is to ascertain how this approach effectively functions in Arabic and how the Qur'anic corpus can be used to create an effective framework for Arabic POS marking. here the researcher will combine two statistical classifier methods, the first is the *K-Nearest Neighbor* (KNN) method and the second is the *Naïve Bayes* (NB) method in an experimental way for making Arabic POS. Then a combination approach called the *Majority Voting Algortihm* (AVM) is used to take advantage of the classifiers. Researchers have investigated many features to determine which ones are useful and how they affect the marking performance of Arabic Qur'an POS. Therefore, the aim of this study was to determine which feature sets are impactful and to standardize am more appropriate POS tagging method. The text of trhe Qur'an is used as a source of research data, using Surah Al-Baqarah which has 6115 words. The *Naïve Bayes* method approach produces a maximum accuracy value of 54.50%. The elements that most effectively produce this accuracy are, P_0 (POS of the current word), W_{-3}

(Word of the three previous words), P_0 (POS of the current word), P_1 (POS of the next word), S_1, S_n, Uppercase and Contains a Number.

Keywords: *Majority Voting Algoritm, Classification, K-Nearest Neighbor, Naïve Bayes, Part of Speech Tagging, Natural Language Processing.*