# Entity Recognition for Quran English Version with Supervised Learning Approach

Muhammad Aris Maulana [#1], Moch Arif Bijaksana [#2], Arief Fatchul Huda [*3]

[#] *School of Computing, Telkom University*
*Bandung, Indonesia*

[1] muharismaulana@student.telkomuniversity.ac.id

[2] arifbijaksana@telkomuniversity.ac.id

[*] *Faculty of Science and Technology, State Islamic University (UIN)*
*Bandung, Indonesia*

[3] afhuda@uinsgd.ac.id

**Abstract**

The Quran is a Muslim holy book that consists of 6236 ayat or verses which divides into 144 surahs or chapters. In each chapter, there are many entities scattered in each verse. For a person, finding a particular entity will be difficult without a classification process, Resulting in difficulties in understanding the Quran. A system can be modeled to extract the information on entities in the Quran to solve this problem. Therefore, we want to offer a method to identify and classify entities using Entity recognition. The system will use the SVM techniques where the system will be given various entities from the Quran as an input to be able to identify correct entities. We are using the dataset obtained from website tanzil.net consists of 19.473 tokens and 720 entities. The classification scenario using a linear kernel with unigram produces the highest f-measure value of 0.75.

**Keywords:** Named-Entity Recognition, Quran, Supervised Learning.

**Abstrak**

Al-Quran merupakan kitab suci Muslim yang terdiri dari 6236 ayat atau bait yang dibagi menjadi 144 surah atau bab. Di setiap bab, ada banyak entitas yang tersebar di setiap ayat. Bagi seorang individu, menemukan entitas tertentu akan sulit tanpa proses klasifikasi yang membuat kesulitan dalam memahami Quran. Sebuah sistem dapat dimodelkan untuk mengekstrak informasi tentang entitas dalam Al-Quran untuk menyelesaikan masalah ini. Oleh karena itu, kami menawarkan sistem untuk mengidentifikasi dan mengklasifikasikan entitas menggunakan Entity Recognition. Sistem akan menggunakan teknik SVM di mana sistem akan diberikan berbagai entitas dari Quran sebagai input untuk dapat mengidentifikasi entitas yang benar. Kami menggunakan dataset yang diperoleh dari situs web tanzil.net terdiri dari 19.473 tokens dan 720 entitas. Skenario klasifikasi yang menggunakan linear kernel dengan unigram memperoleh nilai f-measure tertinggi sebesar 0,75.

**Kata Kunci:** Named-Entity Recognition, Quran, Supervised Learning.