INTRODUCTION

Quran is the Muslim holy book consisting of 114 surahs and around 6200 verses [1]. Because of the large number of surahs and verses, as well as various phrases in the Quran that have a complicated structure, such as nested entities, it is crucial to develop a system that can accurately, rapidly, and automatically extract information, particularly in the form of people entities. The application of the Named Entity Recognition system, which is dedicated to extract people entities, can aid in the automatic extraction of people entities and can also be used for future related research, making it valuable for a better comprehension of the Quran and its Tafseer.

Named Entity Recognition, or NER is a sub-task in information extraction that aims to identify a specific entity in a text, such as a person's name, organization, or geographic location [2]. In this research, what will be emphasized is the use of NER in extracting people entities in the Quran. People entity extraction is a NER task that extracts only people entities rather than the complete named entity, such as a person's name, a group's name, and so on. The text of the verse from the Quran serves as the input to the people entity extraction system, and the people entity extracted by the system serves as the output. Figure 1, shows an example of input from Surah Al-Anfal verses 46-47 the system is expected to identify people entities in the input sentences or ayat. Extracted sentences three and four are examples of people entity extraction from nested people entities, where green sentences are level one entities and blue sentences are level two entities. Only nested entities up to two levels is used in this research. The development of a NER system that can automatically identify and extract people entities in the Quran is important for furthering our understanding of the Quran; moreover, the extracted people entity may be useful for future research that requires people entities to obtain specific information such as a question answering system for the Quran.

In its application, NER system can be divided into several approaches. Earlier NER systems were built with handcrafted rules, lexicons, orthographic features, and ontologies. The system is then followed up with NER based on feature engineering and machine learning [3]. Then, in recent years, NER systems based on neural networks with minimal feature engineering are becoming more popular [4]. The purpose of this study is to extract people entities with several algorithms, then compare their performance. The systems are built based on machine learning and neural networks. For machine learning, a probabilistic model, Conditional Random Field (CRF) algorithm is used [5], and for neural networks, BiLSTM-CRF and pre-trained deep learning model IndoBERT is used. BiLSTM-CRF which was first introduced by Zhiheng Huang *et al*, is a combination of Bidirectional LSTM and CRF [6]. IndoBERT is a pre-trained BERT model for Indonesian language which are trained on indo4B dataset that consists of social media text, blog, news, and website [7], BERT itself is a transformer based based model that uses attention mechanism [8]. The three models were chosen because they were often used at the time for NER and sequence labeling problems.

The main contribution of this research is the extraction of people entities using the aforementioned algorithms and providing a full comparative analysis of the performance benchmarks of the algorithms used. The performance of each algorithm will be measured by the evaluation metric, namely F1-score. This research of several supervised learning algorithms will provide insight for researchers who will use these approaches in the future.

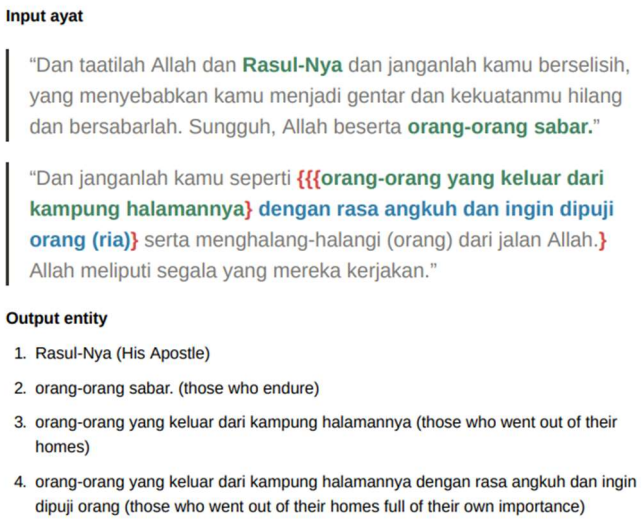

Figure 1. An example of the input paragraph with the intended output of people entities; note that in the second paragraph there are examples of nested entities wrapped in red brackets; entities extracted at the output are only nested entities up to two levels, namely those colored green and blue.