ABSTRACT

NAMED ENTITY RECOGNITION IN LAMPUNG LANGUAGE BASED ON MULTI-CLASS CLASSIFICATION

By

Muhammad Azriel Bintang Saputra

The use of Lampung language in the rajabasa area is 50% of the 6812 total respondents, while 50% use Indonesian as their daily language. Language is a means of communication used to convey ideas, thoughts and information. Information can be accessed on various social media and online news portals. News portals are a form of unstructured data. Unstructured data can cause problems when trying to sort it and find information, so it requires tools and a long time. We can use Natural Language Processing (NLP) and Named Entity Recognition (NER) to analyze and shorten the time to get information from unstructured data. The way NER works is to recognize an entity and then give a label or name to a named entity (NE) in a text. SpaCy library is a library that performs NLP and NER with better performance than other phyton libraries. This research is done labeling and grouping named entities into several categories based on Multi Class Classification with the aim of producing an information system that can present the results of information extraction of Lampung language text data using NER based on Multi Class Classification. The method in this research uses spaCy with CNN and LSTM algorithms for machine learning implementation and extreme programming for information system implementation. An information system is obtained that can perform the NER process from the input of Lampung language text and produce extracted information and the system can classify text based on the specified entity labeling. In further research, a lot of data is needed so that there is no unbalanced data in the modeling process.

Keywords: Lampung Language, NER, SpaCy, LSTM, CNN

ABSTRACT

NAMED ENTITY RECOGNITION (NER) PADA BAHASA LAMPUNG BERBASIS *MULTI CLASS CLASSIFICATION*

Oleh

Muhammad Azriel Bintang Saputra

Penggunaan bahasa lampung didaerah Rajabasa sebanyak 50% dari 6812 jumlah responden, sedangkan 50% nya menggunakan bahasa Indonesia sebagai bahasa sehari-hari. Bahasa merupakan sarana komunikasi yang digunakan untuk menyampaikan ide, pikiran dan juga informasi. Informasi dapat diakses diberbagai sosial media dan portal berita online. Portal berita merupakan bentuk dari data yang tidak terstruktur. Data tidak terstruktur dapat menimbulkan masalah saat mencoba mengurutkannya dan mencari sebuah informasi, sehingga memerlukan alat dan waktu yang lama. Kita dapat menggunakan Natural Language Processing (NLP) dan Named Entity Recognition (NER) untuk menganalisis dan mempersingkat waktu dalam mendapatkan informasi dari data tidak terstruktur. Cara kerja NER adalah melakukan pengenalan terhadap suatu entitas lalu memberikan label atau nama pada sebuah entitas bernama atau Named Entity (NE) dalam sebuah teks. Library SpaCy merupakan library yang data melakuakn NLP dan NER dengan performa yang lebih baik daripada library phyton lainnya. Penelitian ini dilakukan pelabelan dan pengelompokkan entitas bernama menjadi beberapa kategori berbasis *Multi Class Classification* dengan tujuan menghasilkan sistem informasi yang dapat menyajikan hasil ekstraksi informasi data teks bahasa lampung menggunakan NER berbasis *Multi Class Classification*. Metode pada penelitian ini menggunakan spaCy dengan algoritma CNN dan LSTM untuk implementasi machine learning dan extreme programming untuk implementasi sistem informasi. Didapatkan sistem informasi yang dapat melakukan proses NER dari input teks bahasa lampung dan menghasilkan informasi hasil ekstraksi serta sistem dapat mengklasifikasikan teks berdasarkan pelabelan entitas yang ditetapkan. Pada penelitian selanjutnya diperlukan data yang banyak agar tidak terjadi unbalanced data pada proses pembuatan model.

Keywords: Bahasa Lampung, NER, SpaCy, LSTM, CNN