**Manuscript title:** NLP-based approach to classify heterogeneous terms for unambiguous exchange of roadway data

**Manuscript #**: CPENG-1952

**Abstract**

The inconsistency of data terminology due to the fragmented nature of the highway industry has imposed big challenges on integrating digital data from distinct sources. The issue of semantic heterogeneity may lead to the lack of common understanding of the same data between the sender and receiver. Explicit semantic relations in digital dictionaries, such as lexicons, and ontologies can enable the meaning of a concept name to be unambiguously understood by computer systems. However, the current manual process of identifying these relations to support developing semantic resources for the highway sector is laborious and time-consuming due to the lack of an effective automated method. This paper presents a novel methodology that leverages recent advances in Natural Language Processing (NLP) techniques to extract English-American roadway terms and their semantic relations from various design guidelines used in different government agencies. Natural Language Processing (NLP) techniques and the C-value method are first implemented to extract commonly used technical terms from a corpus of roadway design guidelines collected from across the State Departments of Transportation. A model for measuring semantic similarity is then trained on the data of context words in the corpus using the Skip-gram neural network model. This semantic model is then utilized by a proposed term classification algorithm that measures the semantic similarity between terms and assigns relation types (synonyms, hyponyms, and attributes) to pairs of related terms. The proposed methodology was evaluated by conducting an experiment comparing the automatically-identified synonyms by the proposed system with a human-constructed golden standard dataset obtained from Wikipedia. The result shows that the proposed model achieves a precision of over 80 percent.