INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Akshay Parekh
Roll Number	:	166101008
Programme of Study	:	Ph.D.
Thesis Title:		Understanding and Mitigation of Noise in Crowd-Sourced Relation Classification Dataset
Name of Thesis Supervisor(s)	:	Dr Amit Awekar & Prof Ashish Anand
Thesis Submitted to the Department/ Center	:	Computer Science & Engineering
Date of completion of Thesis Viva-Voce Exam	:	31-01-2023
Key words for description of Thesis Work	:	Information Extraction, Relation Classification, & Learning from Noisy Dataset.

SHORT ABSTRACT

Relation classification (RC), a task of classifying the relation between a given pair of entities in a sentence to a relation label is fundamental to IE systems. The identified structured triple (*subject_entity*, *relation*, *object_entity*) from the unstructured text can vastly help in knowledge base completion. This organized relational knowledge can further be used for other downstream tasks like question-answering, and common-sense reasoning. A large RC dataset TACRED has been widely used for benchmarking modern deep neural models. However, RC at a large scale is restricted mainly due to the presence of noise in the training dataset. Hence, the performance of such advanced deep neural models, which have shown excellent improvement on other NLP tasks, has been held back for RC.

This dissertation attempts to analyse noise present in a large-scale RC dataset and propose automated methods to mitigate some of the noise. In particular, the thesis focused on three main objectives:

1. Characterizing Noise present in the RC dataset.
2. Exploring automatic and cost-sensitive approaches to reduce noise from the RC dataset.
3. Analyzing the cost of reannotating them.