**Classification of Citation in Scientific Articles**

*Sanjoy Chowdhury1, Harsh Vardhan1, Pabitra Mitra2 and Dinabandhu Bhandari1*

*1 Heritage Institute of Technology, Kolkata*

*2 Indian Institute of Technology, Kharagpur*

Citations are very important for judging the importance of a scientific paper. The number of times a paper has been cited gives us a measure of the creditability of that paper. In this work, we have tried to find out the type of citation i.e. the citation belongs to which facet by extracting and classifying citation information from the text. Two orthogonal classes each having two subclasses are considered here. The first class tells us whether the citation is of **Juxtapositional** or **Evolutionary** type where as the second class tells us whether the citation is of **Confirmative** or **Negational** type. We have collected the raw data, manually annotated them (manually classified a particular citation belongs to which class) and used them in the investigation. 75% of the data are used for training purpose and rest is used for test. Now as a part of pre-processing of the text a few tools and in-built modules of python NLTK are used. At first, Wordnet Lemmatizer is used to lemmatize the words so that one can find the root word of a given word. Then the sentences are tokenized to get small parts namely 'tokens'. In the next step, the parts of speech of every single word are determined and put them into different pos (parts of speech) classes e.g. "NN"(for Noun), "JJ" (for Adjective), "VB" (for verb), "RB" (for adverb). Further as a part of pre-processing all the white spaces, numbers, stop words (to remove stop words we have used set of stop words from NLTK) are removed. Here we have used the notion of tf-idf (Term frequency inverse document frequency). So the weights of the words which are very frequent in their occurrence have been reduced. Primarily, two types of features those are lexical features and word level features are used. Finally we have used Multinomial Bayes theorem to train the system. At the end of the training phase we subsequently tested the system and achieved an accuracy measure of about 76%.