# Natural Language Processing on SNLI dataset

Name: Arun kumar S

Department: UG

S.R.No: 15656

SNLI dataset : The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labelled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE). Training set, validation set and test set have sizes respectively 550152, 10000, 10000.

## Text cleaning:

In the get\_sentences\_and\_labels(data\_corpus) function, we are cleaning the given corpus to obtain the premise and hypothesis sentences along with labels. So now we have cleaned premise and hypothesis sentences along with their labels.

## Text Preprocessing:

In the first model, we were supposed to use tf-idf features. I have obtained the same using TfidfVectorizer of sklearn.

For the second deep model specific for text, to represent the words as vectors in so that we can get the semantics of the words (like similar words being together in the vector space, unrelated words being separated by large distance), I have used Word embeddings , pretrained glove and Word2vec models. So, I have considered all the prominent ways of text preprocessing. Now, for each of these methods of feature extraction, I need to create various Deep models specific for text like RNN, LSTM, GRU, etc.

1)Logistic Regression Classifier using tf-idf features.

We instantiated TfidfVectorizer. This vectorizer was fitted on the training text (i.e, both the premise and the hypothesis)