c) Paper Reviews: It provides reviews of computing and informatics conferences written in English and Spanish languages. It has 405 reviews which are evaluated on a 5-point scale ranging from very negative to very positive.
d) IMDB: For natural language processing, text analytics, and sentiment analysis, this dataset offers thousands of movie reviews split into training and test datasets. This dataset was introduced in by Mass et al. in 2011 [73].
e) G.Rama Rohit Reddy of the Language Technologies Research Centre, KCIS, IIIT Hyderabad, generated the corpus "Sentiraama." The corpus is divided into four datasets, each of which is annotated with a two-value scale that distinguishes between positive and negative sentiment at the document level. The corpus contains data from a variety of fields, including book reviews, product reviews, movie reviews, and song lyrics. The annotators meticulously followed the annotation technique for each of them. The folder "Song Lyrics" in the corpus contains 339 Telugu song lyrics written in Telugu script [121].

2. Language Modelling: Language models analyse text data to calculate word probability. They use an algorithm to interpret the data, which establishes rules for context in natural language. The model then uses these rules to accurately predict or construct new sentences. The model basically learns the basic characteristics and features of language and then applies them to new phrases. Majorly used datasets for Language modeling are as follows:

a) Salesforce's WikiText-103 dataset has 103 million tokens collected from 28,475 featured articles from Wikipedia.
b) WikiText-2 is a scaled-down version of WikiText-103. It contains 2 million tokens with a 33,278 jargon size.
c) Penn Treebank piece of the Wall Street Diary corpus includes 929,000 tokens for training, 73,000 tokens for validation, and 82,000 tokens for testing purposes. Its context is limited since it comprises sentences rather than paragraphs [76].
d) The Ministry of Electronics and Information Technology's Technology Development Programme for Indian Languages (TDIL) launched its own data distribution portal (www.tdil-dc.in) which has cataloged datasets [24].

3. Machine Translation: The task of converting the text of one natural language into another language while keeping the sense of the input text is known as machine translation. Majorly used datasets are as follows:

a) Tatoeba is a collection of multilingual sentence pairings. A tab-delimited pair of an English text sequence and the translated French text sequence appears on each line of the dataset. Each text sequence might be as simple as a single sentence or as complex as a paragraph of many sentences.
b) The Europarl parallel corpus is derived from the European Parliament's proceedings. It is available in 21 European languages [40].
c) WMT14 provides machine translation pairs for English-German and English-French. Separately, these datasets comprise 4.5 million and 35 million sentence sets. Byte-Pair Encoding with 32 K tasks is used to encode the phrases.