relevant references cited therein. The use of BERT (Bidirectional Encoder Representations from Transformers) [33] model and successive models have also played an important role for NLP.

Many researchers worked on NLP, building tools and systems which makes NLP what it is today. Tools like Sentiment Analyser, Parts of Speech (POS) Taggers, Chunking, Named Entity Recognitions (NER), Emotion detection, Semantic Role Labeling have a huge contribution made to NLP, and are good topics for research. Sentiment analysis (Nasukawaetal.,2003) [156] works by extracting sentiments about a given topic, and it consists of a topic specific feature term extraction, sentiment extraction, and association by relationship analysis. It utilizes two linguistic resources for the analysis: the sentiment lexicon and the sentiment pattern database. It analyzes the documents for positive and negative words and tries to give ratings on scale −5 to +5. The mainstream of currently used tagsets is obtained from English. The most widely used tagsets as standard guidelines are designed for Indo-European languages but it is less researched on Asian languages or middle- eastern languages. Various authors have done research on making parts of speech taggers for various languages such as Arabic (Zeroual et al., 2017) [160], Sanskrit (Tapswi & Jain, 2012) [136], Hindi (Ranjan & Basu, 2003) [105] to efficiently tag and classify words as nouns, adjectives, verbs etc. Authors in [136] have used treebank technique for creating rule-based POS Tagger for Sanskrit Language. Sanskrit sentences are parsed to assign the appropriate tag to each word using suffix stripping algorithm, wherein the longest suffix is searched from the suffix table and tags are assigned. Diab et al. (2004) [34] used supervised machine learning approach and adopted Support Vector Machines (SVMs) which were trained on the Arabic Treebank to automatically tokenize parts of speech tag and annotate base phrases in Arabic text.

Chunking is a process of separating phrases from unstructured text. Since simple tokens may not represent the actual meaning of the text, it is advisable to use phrases such as “North Africa” as a single word instead of ‘North’ and ‘Africa’ separate words. Chunking known as “Shadow Parsing” labels parts of sentences with syntactic correlated keywords like Noun Phrase (NP) and Verb Phrase (VP). Chunking is often evaluated using the CoNLL 2000 shared task. Various researchers (Sha and Pereira, 2003; McDonald et al., 2005; Sun et al., 2008) [83, 122, 130] used CoNLL test data for chunking and used features composed of words, POS tags, and tags.

There are particular words in the document that refer to specific entities or real-world objects like location, people, organizations etc. To find the words which have a unique context and are more informative, noun phrases are considered in the text documents. Named entity recognition (NER) is a technique to recognize and separate the named entities and group them under predefined classes. But in the era of the Internet, where people use slang not the traditional or standard English which cannot be processed by standard natural language processing tools. Ritter (2011) [111] proposed the classification of named entities in tweets because standard NLP tools did not perform well on tweets. They re-built NLP pipeline starting from PoS tagging, then chunking for NER. It improved the performance in comparison to standard NLP tools.

Emotion detection investigates and identifies the types of emotion from speech, facial expressions, gestures, and text. Sharma (2016) [124] analyzed the conversations in Hinglish means mix of English and Hindi languages and identified the usage patterns of PoS. Their work was based on identification of language and POS tagging of mixed script. They tried to detect emotions in mixed script by relating machine learning and human knowledge. They have categorized sentences into 6 groups based on emotions and used TLBO technique to help