Abstract

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) which trains computer to understand and interpret human language. Illustration of the whole document by a programmed process which has basic operation to collect a set of related words from the document is keyword extraction. It recognises the core information of the document. In finding the relevant information in a huge number of documents, keyword extraction is the significant approach. In NLP topic modeling (TM) is a statistical model for deriving topics from a huge collection of unlabelled text. Various models exist in literature for the purpose of topic modeling in different languages such as Arabic, English etc. But no persuasive TM is available for native Pakistani Languages particularly Punjabi. In this study, our focus is to work on existing TM like Latent Dirichlet Allocation (LDA) and Term Frequency - Inverse Document Frequency (TF-IDF) to overcome the issues of text mining in Punjabi language. LDA as an unsupervised model and TF-IDF for the Punjabi topic identication has been analyzed. LDA trained to dataset provided by TF-IDF scores and it improves the accuracy of topic extraction for Pakistani Punjabi language.