Recent advancements in the disciplines of text mining and natural language processing have helped text classification (TC), a task of basic importance, gain traction. Text classification techniques all aim to assign a predefined label to a given input text, though this term can apply to a range of specialized techniques used in many fields. As from the article **on Text Classification Algorithms: From Text to Predictions [1],** the standard preprocessingtechniques includetokenization, stop word and noise removal.

After standard preprocessing the model, it is recommended to apply preprocessing deep models, these include models such as Byte Pair Encoding, which is now regarded as the breakthrough method for sub-word tokenization. The most common vocabulary term is combined into a new vocabulary word via this algorithm, which iteratively calculates the occurrences of subsequent pairs of vocabulary terms. The word Piece tokenizer employs a data-driven strategy for breaking words into sub-words. It was originally developed as a solution to Japanese text segmentation issues. It uses n-gram-based language models to identify recurrent prefixes, word fragments, and syllables in corpora. But unlike Word piece, UnigramLM takes the opposite approach, initializing at a vocabulary size that is significantly greater than the intended number of sub-words and then going on to iteratively delete them. The expectation-maximization procedure is applied in each iteration to eliminate the items with the lowest likelihood, cycling until the vocabulary reaches the required size. We apply the data preprocessing techniques suggested in this article.

We then use topic modelling for classification. A topic model is a model that can identify themes in a document based on the words that appear there. We use Latent Dirichlet Allocation [2] technique which is one the method for topic modelling. LDA assumes that each document is made up of a combination of subjects and words, and that each topic is made up of a combination of topics.

Graphical user interface, diagram, application, Teams

Description automatically generated

The above flow chart shows the steps involved in LDA. It uses three different parameters. The number of topics anticipated in the text is controlled by the α hyperparameter. K determines how many topics we need to extract, and the β hyperparameter determines how many words per topic there are in the document. The major limitation while applying this model is one must manually assign the themes to the given topics, which could lead to mistakes. LDA performs poorly on small texts. Since the description column in our dataset has sufficient information, we feel that LDA is better for applying

[1]  *Andrea, G; Matteo, M; Alessandro, Z; Andrea, A.* A Survey on Text Classification Algorithms: From Text to Predictions **2022**

<https://www.mdpi.com/2078-2489/13/2/83/htm>

[2] Topic Modelling using LDA

<https://www.mdpi.com/2078-2489/13/2/83/htm>