### **MODULES**

The primary objective of module design is to develop a modular program structure and represent the control relationship between modules. FAKE NEWS EXPOSURE has mainly contains four modules. These modules are described below.

* Data Pre-processing:
* Derive Relationships
* Feature Extraction
* Training Classification Model

**MODULE DESCRIPTION**

**1. Data Pre-processing:**

The dataset comprised of information (in the form of JSON files) of news articles from BuzzFeed and PolitiFact. Along with this, the dataset provided information about users and user-news interaction in the form of CSV files. Furthermore, the dataset included real and fake news content. Based on our intuition and research, we found that the “body” of the news articles, i.e, main content, best represents the news articles core information and variance. After extracting the body of all the news articles we constructed a data frame of the news-ids and the body of the text. A label was also added to the data frame to indicate whether the news article is fake or real.

**2. Derive Relationships**

Once the important aspects of the data were identified for feature extraction, the next step involved establishing relationships among the news articles that can be leveraged by classification model. The relationship between the news articles was established using a graph data structure. Here, the news articles act as the nodes of the graph and the edges between the nodes represent the relationships among the new articles. The edges between the nodes of the graphs were established using the following steps.

**3. Feature Extraction**

For the fake news detection, the actual news data (body of the news article) is being considered as features. But the data is in the form of text. It is known that for the machine learning analysis, text data does not work well. So the text data has to be converted into a numerical representation. This process is called vectorization. Every record (i.e news article in this case) should be converted into a vector. There are several techniques/algorithms which can convert text to vector. Below is the list of such techniques.

**4. Training Classification Model**

As explained in the previous sections, the adjacency matrix, feature vectors and the labels form the input for the classification model. The fake news detection task has network-based input in the form of adjacency matrix representing the relationship between news articles. As explained before, this relationship is derived from the information provided in the form of users and their association with the news articles (post, share, retweet etc). As one can observe, the data has the pattern of a graph with nodes or vertices being the news articles and edges as the relationship between them. The classification task needed a model which can utilize this network or graph-based architecture of the dataset and the feature vectors generated for the news articles using BERT. Traditional machine learning models utilize just the feature vectors for prediction and will not leverage the graph-based architecture of the data.