Observations and insights made from the assignment

- The dataset given to me in assignment was an excel sheet which contained initially of 9 columns or features. The main focus of this data as provided in the problem statement was to detect if the news told was fake or true. The solution of this, to put it on a high or surface level is that of a classification. This means there were 4-5 types of classes like ‘pants-fire’, ‘true’, ‘false’, ‘partially true’ etc. Thus, our main aim is to create a text mining algorithm which can automatically identify whether a given piece of statement is true or not. If yes, how much truth is there? Thus to solve this, we needed to extract keywords/relevant from the “statement” column of the dataset which will help us understand the truth in that sentence. There were columns like “Resource” , “Party”, “Speaker”, “Topic” etc which along with the lemmitized or processed data would help train our algorithm. For example, if the “Resource” was from online and the “Speaker” was **someone not known** to provide information initially, giving a factual statement about “Topic” abortion and the it is labelled as false , then in the future if we get a text about someone making a statement from an online platform about abortion and he/she is not someone not recognised as a public figure, chances are higher that this piece of information is false too. Thus having many attributes or dimensionalities will help us develop and train our text mining algorithm to produce more precise prediction.

- We first needed to clean all the rows in the “Statement” column in order to make the algorithm more efficient and precise. Thus we first removed all the special characters from the text and converted the rest of the text in lower case. Then we separated the statements in the double quotes as they can help us identify the key words or assist us in some way along the process. Removing all the special characters from the text required a bit of knowledge on regular expressions and after the process, the data was presented in a much trivial and machine understandable format. After each and every data processing step, the complexity of language was decreasing and it was turning into more computer friendly text. By the time I was completed with the lemmatization it there were only root words of some key words remaining and the sentence in general had lost all it’s meaning to me as a human. But the computer unlike humans, can understand a lower level of syntax and semantics of a language. The stop word removal will work for cases where the context of the statement is not required which happens to be true in this problem as we are mainly trying to only classify the text in to two or three groups. For a more complex text mining job, we might need to skip this procedure. Using NTLK package to perform tokenization will help us in most classification problem and it will help us get a good algorithm but in more contextual based cases using spacy package for tokenization would be better.

- Another thing which I would do in this text cleaning process is something which is not being done on the “Statement” column but other features. Especially the label column where the values are among [“Pants Fire”, “ False”, “barely true”, “mostly true”, “true”] which can be given an integer value so that it would be more convenient for the machine or computer to understand as it is more trivial than a text. Also I think Pants fire annotation is contextual and it is not literally meat that pants are on fire. The same numerical labelling can be done to “Resource ” columns and “Speaker” column.