**CHAPTER 3**

**METHODOLOGY AND SYSTEM ANALYSIS AND DESIGN**

The main objective of this research is to analyze the public perception towards 5G connectivity using natural language processing (NLP) methods. NLP is a branch of artificial intelligence that deals with the interaction between computers and human languages (Liu B. et al). We will use NLP methods to extract, process, and understand the opinions and emotions expressed in text data related to 5G technology. We will use two main NLP methods for our analysis: sentiment analysis and topic modeling. Sentiment analysis is a technique that helps identify and classify the polarity of a text, such as positive, negative, or neutral (Pang. B & Lee, 2008). Topic modeling is a technique that helps discover the hidden themes or topics in a collection of documents (Kumar, A., & Singh, S., 2020). By combining these two methods, we will be able to measure how people feel about different aspects of 5G technology and its benefits and risks. We will collect data from a popular social media platform Reddit to capture the diverse opinions and perspectives of the public. We will filter the data by keywords related to 5G technology, using “5G” as the keyword. We will also remove irrelevant or spam data, such as advertisements, promotions, or bots. We will apply sentiment analysis and topic modeling methods to the data and compare the results. We will also use visualization tools to display the results in graphs and charts.

The system analysis and design of our research consists of the following steps:

**3.1 Data collection**

Reddit’s posts were extracted for this study, from different subreddits and authors sharing their views on 5G connectivity, also in this study, Python was used for scraping the data from Reddit, preprocessing, data mining, and result visualization. This data is easily accessible for people studying public opinions, sentiments, and beliefs. We will use the popular Reddit scraper tool in python call Praw, beautifulSoup4, and Selenium to collect data, all relating to 5G from Reddit, using the keyword “5G”. We will store the data in a CSV file for further processing. The data collected have the following attributes or features like created (date and time when post was made), author’s name, post title, post content, comments, and source and with about 229 entries(rows). Additional data was collected to examine the sentiments in two subreddits r/5GDebate and r/5GDidit, All the posts in these subreddits of about 1500 posts were collected and analyzed to derive further insights and topics surrounding 5G connectivity.

**3.2 Data preprocessing**

Social media data is fun to work with but is unstructured. A piece of meaningful information can be drawn if the quality data is ensured. The unstructured and semi-structured data need pruning to get an insight from it. For this study, Python was used for data cleaning, popular libraries like pandas, natural language tool kit were used to transform our dataset. We will perform various data cleaning and normalization tasks, such as removing punctuation, stopwords, numbers, URLs, e.t.c. We will also perform tokenization, lemmatization, and stemming to reduce the dimensionality of the data.

**3.3 Sentiment analysis**

Sentiment analysis is a technique that uses natural language processing and machine learning to identify and extract the emotions and opinions expressed in a text. It can help with various applications, such as analyzing customer feedback, social media posts, product reviews, and more. Sentiment analysis can also provide insights into the polarity, intensity aspect of the sentiment in a text. For our study, we will use popular libraries like TextBlob and SentimentIntensityAnalyzer to measure sentiments’ polarity and subjectivity.

**3.4 Topic modeling**

Topic modelling provides methods for automatically organizing, understanding, searching, and summarizing large text data.

* Discovering the hidden themes in a collection.
* Classify the documents into discovered themes

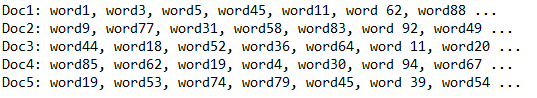
We will use an unsupervised machine learning approach to perform topic modeling on the data. We will use Latent Dirichlet Allocation (LDA), a popular probabilistic generative model that assumes that each document is a mixture of topics and each topic is a distribution of words. We will use coherence scores to determine the optimal number of topics for our model.

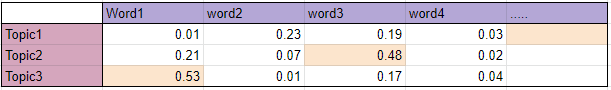
The aim of LDA is to help us discover these topics each document belong to based on the words in it.

How does LDA work?

* Go through each document and randomly assign each word in the document to one of ***k*** topics (*k* is chosen beforehand).
* For each document ***d***, go through each word ***w***and compute:
* **p(topic *t*| document *d*)**: the **proportion of words in document *d* that are assigned to topic *t****.*Tries to capture how many words belong to the topic *t*for a given document *d*. Excluding the current word.  
  If a lot of words from *d*belongs to *t*, it is more probable that word *w* belongs to *t*.
* p(word w| topic t): the proportion of assignments to topic t over all documents that come from this word w. Tries to capture how many documents are in topic t because of word w.  
  LDA represents documents as a mixture of topics. Similarly, a topic is a mixture of words. If a word has high probability of being in a topic, all the documents having w will be more strongly associated with t as well. Similarly, if w is not very probable to be in t, the documents which contain the w will be having very low probability of being in t, because rest of the words in d will belong to some other topic and hence d will have a higher probability for those topic. So even if w gets added to t, it won’t be bringing many such documents to t. (Sarkar, D, 2018)
* Update the probability for the word *w*belonging to topic *t*, as.

*p(word w with topic t) = p(topic t | document d) \* p(word w | topic t)*





**3.5 Data visualization**

We will use various visualization tools to present the results of our analysis in an intuitive way. We will use bar charts, histograms, Line charts, etc. to show the distribution of sentiments and topics across different sources and time periods. We will also use word clouds, scatter plots, etc. to show the most frequent and relevant words and phrases for each sentiment and topic.

We hope that this research will provide valuable insights into how people perceive 5G technology and what are their main concerns and expectations regarding it. We also hope that this research will contribute to the advancement of NLP methods and applications in social media analysis.

**References**

* Liu, B., Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, 5 (1), pp. 1-167, 2012.
* Pang, B. & Lee, L., Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, 2 (1-2), pp. 1-135, 2008.
* Sarkar, D. (2018, April 16). Latent Dirichlet Allocation (LDA). Towards Data Science.
* Kumar, A., & Singh, S. (2020). Topic modeling, sentiment analysis and text summarization for analyzing news headlines and articles. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE. 1