

**9/5/2022**

**RESEARCH PROPOSAL**

**RSAD**

**Research Skills and Dissertation**

**Mahathi MONAGIGARI- 31050513**

**TABLE OF CONTENTS**

1. INTRODUCTION AND JUSTIFICATION ………………….. 2
2. RESEARCH QUESTION, AIM AND OBJECTIVES, DELIVERABLES………………………………………………2
3. LITERATURE REVIEW……………………………………… 3
   1. Models used for Analysis………………………………….. 3
   2. Data sources, Collection and Analysis…………………….. 3
   3. Split of dataset into Training and Test Data………….……. 4
   4. Evaluation of models used in Literature…………………… 5
4. RESEARCH DESIGN………………………………….………7
5. ETHICS, RISKS AND ISSUES………………………………..9
6. TIME PLAN…………………………………………………….9
7. REFERENCES…………………………………..……………..10

Sentiment Analysis using Deep Leaning approach of Customer Reviews on UK Energy companies on Twitter

Mahathi Monagigari

31050503

# Introduction and justification

A massive volume of both structured and unstructured multimedia data is being uploaded on the Internet due to rapidly growing ubiquitous web access over the world. (Woldemariam 2016). Reviews, opinions, emotions from the customer play a major role in the analysing the growth of company and helps in further decision for any organisation. This field of study that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from written language is known as Sentiment Analysis(Liu 2012).

ChatGPT is designed to understand and generate human like text responses. As more businesses and users interact with ChatGPT, it becomes crucial to gauge the user sentiment and experiences. For tools like ChatGPT the significance of customer review provides valuable insight and this project helps to understand the customer insight on ChatGPT. This Analysis helps the organisation to identify area of improvement, optimise user interactions therefore leading to informed decisions on implementation.

# Research question, aims & objectives

## Research question

How accurate is the deep learning approach effective on finding Sentiment polarity analysis of text in Reviews of customer in Twitter on ChatGPT?

## Objectives

* Main objective of this project is to determine the Sentiment polarity and classify the public opinion on ChatGPT AI through Twitter Reviews.
* While calculating sentiment polarity, the other objective is to understand different word Embedding techniques and deep learning methods.
* Understand the processing time and performance of different deep learning methodologies.

## Deliverable

This project delivers a Deep Leaning Model with Calculated Accuracy, Precision, Recall and F1 score and shows the public opinion of positive, Negative and Neutral based on Twitter Reviews for ChatGPT AI.

# Literature review

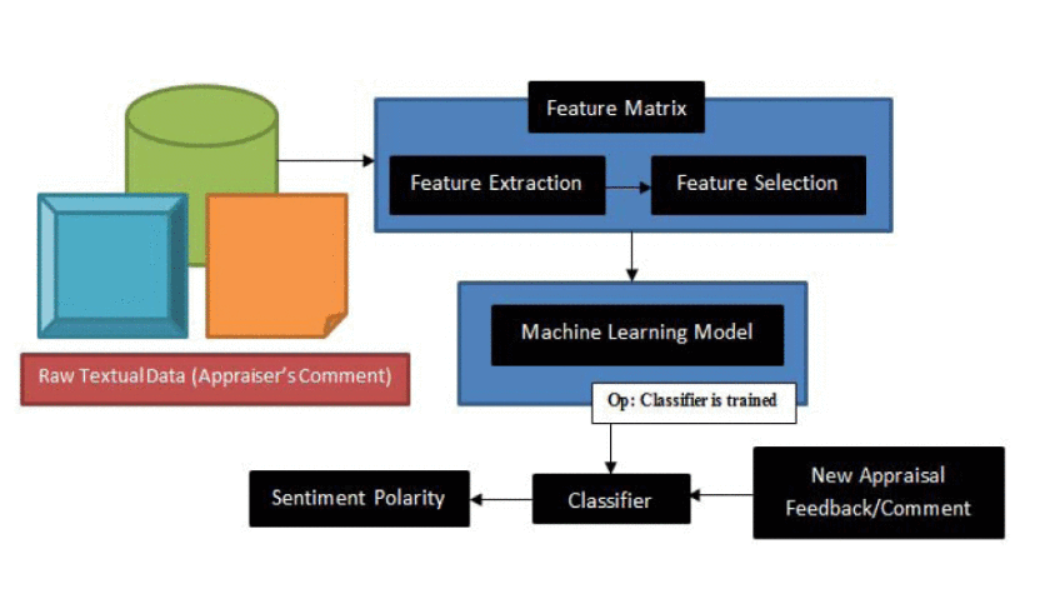
**Models used for Analysis**

From the current literature, Studies - (Liu 2012) , (Niu and Wu 2019), (Woldemariam 2016), (Zharmagambetov and Pak 2015) have used natural language processing for sentiment analysis of their own area of interest. At present, the sentiment analysis mainly includes the method based on sentiment dictionary, the traditional machine learning method and the deep learning method based on neural network (Niu and Wu 2019) . ELM has been an important research topic due to its high efficiency, easy-implementation, unification of classification and regression and therefore might be capable to be implemented in sentiment analysis field (Zhang and Zheng 2016).

In general, sentiment analysis methods are classified into lexicon-based and machine learning-based. Machine learning methods make use of learning algorithm and classifier models trained on a known dataset. The lexicon-based approach involves calculating sentiment polarity using dictionaries of words annotated with sentiment scores (Woldemariam 2016). Woldemariam (2016) study compared lexicon- based approach and Machine learning method Recursive Neural Tensor Network (RNTN). The difficulty of sentiment analysis is determined by the emotional language enriched by sleng, polysemy, ambiguity, sarcasm, all this factors are misleading for both humans and computers. Word2Vec algorithm helps in in reducing the space of thesaurus into the meaning's space (Zharmagambetov and Pak 2015). Supervised Model cannot be used when sufficient training data is present for analysis and thereby lexicon based algorithm can be used for optimisation (Ikoro, Sharmina et al. 2018). Many giant business organizations are getting benefitted from NLP (Sajib, Shargo et al. 2019). Relying two issues, opinion extremity and estimation score , (Choudhary and Chhabra 2021) study classified the texts of a client into a positive or negative and The anticipated model appropriately detaches service and product review, notwithstanding this.

**Data sources, Collection and Analysis:**

All the mentioned studies in the document used various data sources like Twitter, Google, IMDB Movie reviews. The data from twitter is in the form of API and from the API’s data is extracted using some of the tools like PostMan. One of the advantages of using this approach to retrieve the data for sentiment analysis was that we were able to retrieve a conversation thread between the consumers and the energy suppliers within a specific time frame. One of the challenges in sentiment analysis is the lack of annotated data sets that can be used to train a model capable of adjusting to differences in multiple domains . (Sajib, Shargo et al. 2019) study’s dataset is collected from a huge corpus of people’s tweet sentiments on tech companies collected by Niek Sanders. Flowchart of sentiment analysis in Study of (Choudhary and Chhabra 2021) is depicted by below figure.



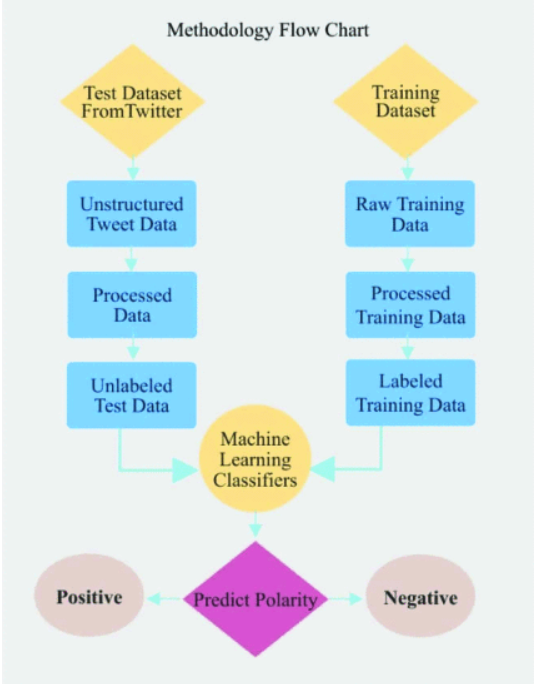
*Figure 1*

However studies like (Zhang and Zheng 2016) do not have a clear mention of data source but there is proper mention of usage of data.

After data collection, there is need of data pre-processing is required. Data pre-processing steps include Denoising, Segmentation, Filtering the Stop-Word. Segmentation is done either by words or numeric. Study (Ikoro, Sharmina et al. 2018) of used sentiment lexica for our analysis. There are some domain-specific lexica that take into account the variations in the use of words and the context or community in which a word is being assessed for polarity. Other studies classified the data to either positive/negative words and some classified to 1 or 0 based on the customer review.

**Split of Data into Training and Test Data:**

Pre Processed Dataset is classified into training and testing data. Machine learning algorithms are trained by training dataset and in order to see if the model developed is accurate or not, algorithm is tested with test dataset. Naive Bayes classifiers is a set of classification algorithm based on Bayes’ Theorem. Using Bayes theorem, we can find the probability of X happening, given that C has occurred. More precisely, given labelled training data, the SVM algorithm outputs an optimal hyperplane which draws a boundary between two different set of classes.  Raw data from the dataset is used for training the machine learning algorithm and later the actual dataset from datasource is used to test the algorithm. Below figure illustrates us the training and testing of an algorithm (Sajib, Shargo et al. 2019).



*Figure 2*

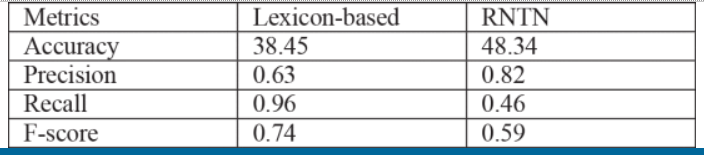
Study of (Zharmagambetov and Pak 2015) have used the 50 percent of data for training and 50 percent for testing. Some other studies like (Sajib, Shargo et al. 2019) training dataset from other data source (Example- tweets is collected from a huge corpus of people’s tweet sentiments on tech companies collected by Niek Sanders) and thetest dataset is from other data source (Example- Twitter API).

**Evaluation of models used in Literature:**

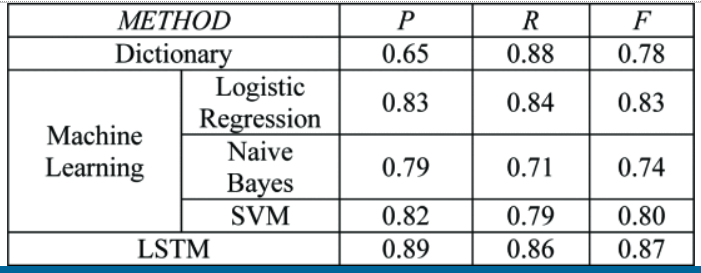
The huge volume of information has to be technically processed for segregating the relevant knowledge. Sentiment analysis is the popular method extensively used for this purpose (Kumar and Garg 2020). Therefore there is need in using the best algorithm with high accuracy rate for analysing the sentiment polarity between the customer reviews. Study of (Choudhary and Chhabra 2021) helped in comparison of machine learning models using confusion matrix which was not done by any of other papers in this Literature review. The result for the confusion matrix could not correctly asses the negative reviews. Study of (Zhang and Zheng 2016) compared one of the machine learning Model to extreme learning model. Accuracy of machine learning model and extreme learning models is almost same but the performance of ELM is quite effective comparing to ML Model. On clear evaluation of (Niu and Wu 2019) study, the result of each algorithm depends on large number of new words and unregistered words. From this paper, deep learning algorithm is better when compared to machine learning and dictionary based methods. It is stated that with the deep learning, it is prone to over fitting which needs to be taken care. In a study Lexicon based and RNTN models are compared and RNTN has high accuracy over lexicon and worked well for negative data whereas Lexicon based model worked for positive well. From the Literature review from many papers, the process that is followed for better accuracy are below:

1. Data collection from the data source
2. Data pre- Processing and Assuring data Quality
3. Encoding of data into positive or negative or neutral based on the context
4. Human annotation before applying the machine learning models
5. Split of the dataset into training and testing dataset.
6. Applying the Model according to their interest of study.

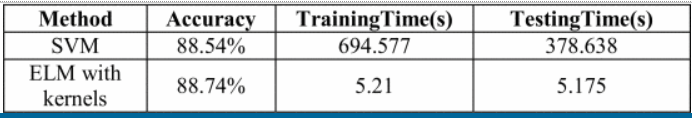
Below are some figures from various papers showing the accuracy of models based on their own dataset to test the accuracy.



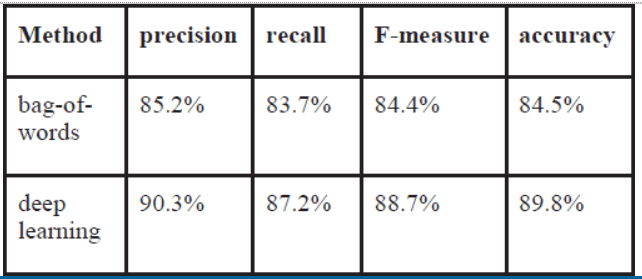
*Table 1*



*Table 2*



*Table 3*



*Table 4*

# Research Design

In order to carry on this research, Concept of Research Onion (Saunders et. al, 2016) is followed and the dissertation research is done at each level.

1. **Research Philosophy:** Pragmatic approach of research will be adopted to do the research of Sentiment Analysis on Twitter Reviews of customers on ChatGPT. As in Pragmatic approach, Knowledge while doing the research is not fixed to what we have, but constantly questioned and interpreted. Conclusions of the research comes out as an answer to the of the problem statement that is found in previous research on questioning the outcome.
2. **Research Approach:** Abductive approach will be followed from the research onion concept. Here for the current dissertation research, Data is main key point that is collected from Twitter. Using this data from Twitter, a model is developed that generates a pattern to explain our research question. Subsequently some data is used for pattern generation or analysis and rest is of Data collected is used for testing the generated patterns.
3. **Research Methodoly:** The above mentioned Research question in this document can be answered through Mixed Method simple design. We aim to start with quantitative data from Twitter API in order to be sufficient for analysis. Later the collected data go through the quality check that is cleaning and categorisation which is qualitivate method.
4. **Research Strategy:** Rather than doing research as an experimental or Action research, I would like go with Case study strategy as this is a detailed, in-depth study of a single subject. The objective in selecting Case study is to gain depth understanding of context. In this case of research, Customer reviews on Twitter are narrowed down from a statement to emotion and depth Sentiment analysis on data will be carried out to understand the actual emotion of customer. While we use case study research strategy it helps in extending the experience on Sentiment analysis and adds strength to already known subject. Research is carried out as a case study as it can comprise of various methodologies to investigate the above mentioned research problem.
5. **Time Horizon:**  The time over which the research to answer the research question is cross sectional. Within the limited time for research, the data of one year would really be helpful.
6. **Data Collection:** The next layer and is a key point is Data collection for analysis. For the research that answers Big Data research questions, the main role is within the Data collection. The Data will be collected from Twitter API. The API from twitter for ChatGPT reviews by customers will be extracted though PostMan tool and the data will be copied to an excel file.

Once the data is collected to an excel file, Below are the steps that will be carried out

1. Data cleaning using python environment and libraries like Numpy, Pandas.
2. Categorisation: For each review of customer on ChatGPT, Category of Positive, Negative and neutral categorisation will be done using Python script.
3. Once the data in excel file is clean and categorisation, The data is loaded as a Data frame to Python and using machine learning libraries like scikit etc. for deep learning model.
4. While developing the Deep Learning approach, 80% of the Data collected will be used for training the Model and the remaining 20% of Data collected will be for testing the accuracy of Developed Model.
5. Based on the accuracy, Precision, Recall and F1 score of test data recommendations will be made as a result.

# Ethics, risks and Issues

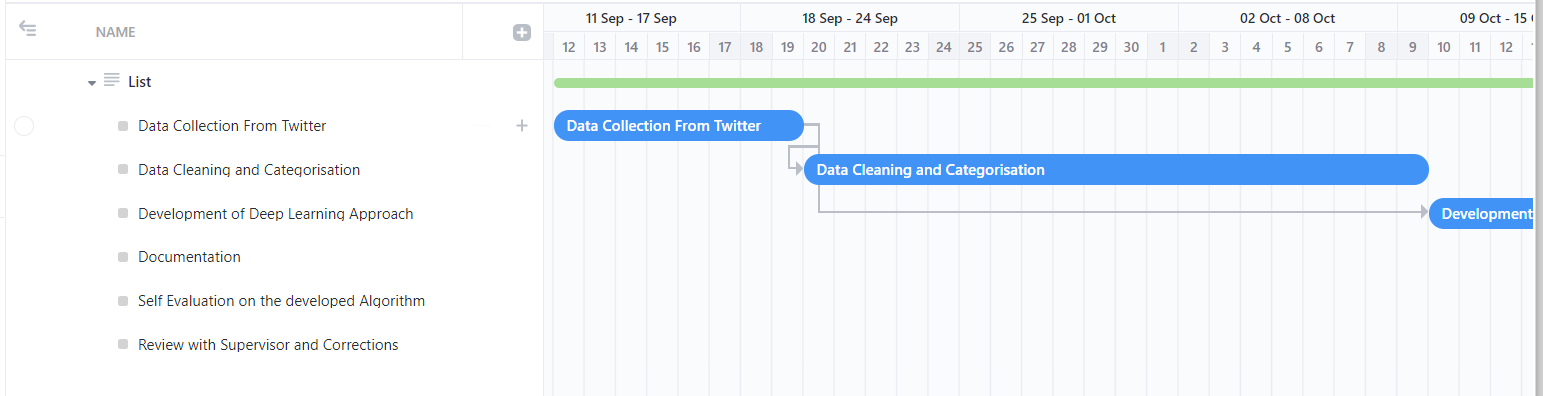
As part of this research, Data will be collected from Twitter API and as it is a free open source data there is no particular list of ethics to be followed. Despite there are challenges or issues in carrying out this research.

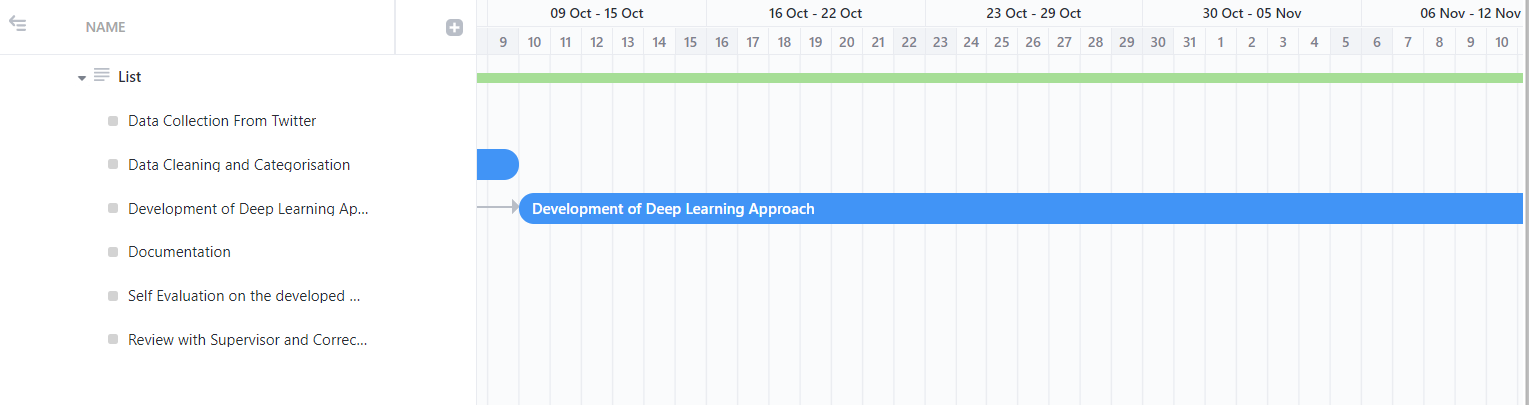
In collecting and retrieving data to form large datasets it may not be possible to obtain informed consent from all of the participants, simply due to the volume of tweets retrieved. There will be an ethical issue when we try to reproduce the tweets.

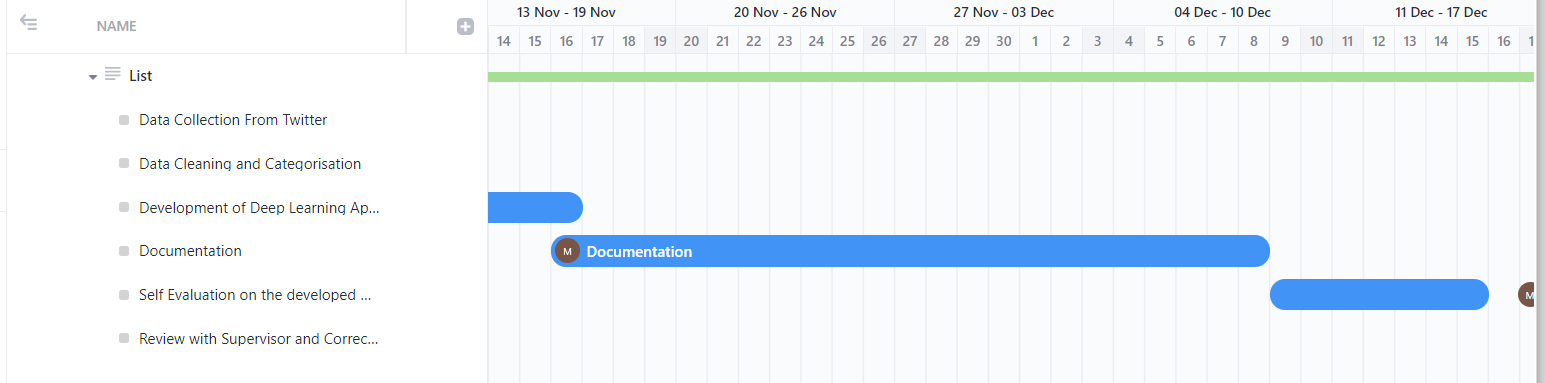
Use of certain keywords or hashtags to retrieve data sometimes may not give us the entire data related to the topic. Proper brainstorming helps us in search queries.

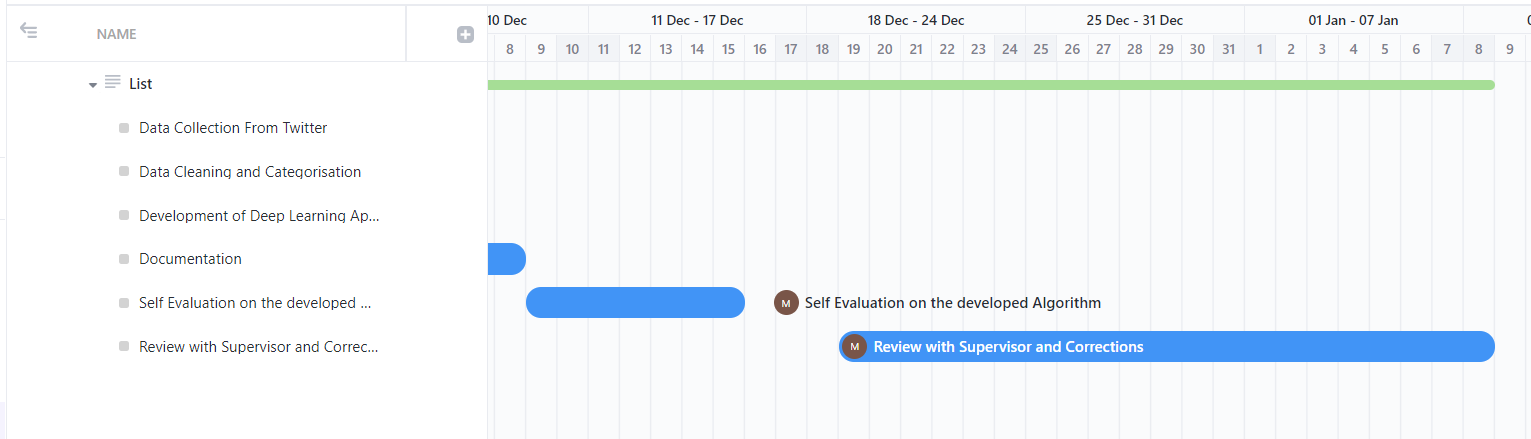
# Time Plan

Below gannt chart is pictorial representation of different stages of research and it shows the time line for each stage of research. Overall three months of duration for research is divided into different stages based on the complexity of the task.









# REFERENCES

1. Ikoro, V., Sharmina, M., Malik, K., & Batista-Navarro, R. (2018, October). Analyzing sentiments expressed on Twitter by UK energy company consumers. In *2018 Fifth international conference on social networks analysis, management and security (SNAMS)* (pp. 95-98). IEEE.
2. Kumar, A., Garg, G. Systematic literature review on context-based sentiment analysis in social multimedia. Multimed Tools Appl 79, 15349–15380 (2020). https://doi.org/10.1007/s11042-019-7346-5
3. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.
4. Woldemariam, Y. (2016, March). Sentiment analysis in a cross-media analysis framework. In *2016 IEEE international conference on big data analysis (ICBDA)* (pp. 1-5). IEEE.
5. Mansouri, I., Newborough, M., & Probert, D. (1996). Energy consumption in UK households: impact of domestic electrical appliances. *Applied Energy*, *54*(3), 211-285.
6. Zhang, X., & Zheng, X. (2016, July). Comparison of text sentiment analysis based on machine learning. In *2016 15th international symposium on parallel and distributed computing (ISPDC)* (pp. 230-233). IEEE.
7. W. Niu and L. Wu, "Sentiment Analysis and Contrastive Experiments of Long News Texts," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2019, pp. 1331-1335, doi: 10.1109/IAEAC47372.2019.8997550
8. S. Choudhary and C. Chhabra, "Sentiment Analysis of Amazon Food Review Data," 2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2021, pp. 116-120, doi: 10.1109/CCICT53244.2021.00033.
9. M. I. Sajib, S. Mahmud Shargo and M. A. Hossain, "Comparison of the efficiency of Machine Learning algorithms on Twitter Sentiment Analysis of Pathao," 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038208.
10. S. Zharmagambetov and A. A. Pak, "Sentiment analysis of a document using deep learning approach and decision trees," 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO), 2015, pp. 1-4, doi: 10.1109/ICECCO.2015.7416902.
11. Saunders, M., Lewis, P. & Thornhill A. (2016). Research methods for business students (7th ed.) Essex: Pearson Education.