METHODOLOGY AND SYSTEM ARCHITECTURE

**3.1 Introduction**

This chapter outlines the methodology and system architecture used to analyze public opinions and sentiment towards organic food. The study leverages sentiment analysis and topic modelling techniques to provide a comprehensive understanding of consumer behavior and sentiment.

The aim of this project is to analyze the public opinions and sentiment towards organic food using big data and machine learning techniques. Organic food is food produced without the use of synthetic pesticides, fertilizers, hormones, antibiotics, or genetically modified organisms (GMOs) (Journal of Big Data, 2021). Organic food has been gaining popularity in recent years due to its perceived health, environmental, and ethical benefits (SpringerLink, 2022). However, organic food also faces challenges such as higher prices, lower availability, and inconsistent quality (ScienceDirect, 2019). Therefore, it is important to understand how consumers perceive and react to organic food in order to design effective marketing strategies and policies (ScienceDirect, 2020).

The methodology of this project consists of three main stages: data collection, data preprocessing, and sentiment analysis. The data source for this project is extracted for a popular social media platform known as Reddit, data collected contain opinions on organic food. The data covers a time span from 2009 to 2023 to capture the current trends and sentiments of the public (Journal of Big Data, 2021). Data preprocessing involves cleaning, tokenizing, normalizing, stop-word removal, and stemming the text data to make it suitable for analysis (SpringerLink, 2021). Sentiment analysis involves applying sentiments classification techniques using sentiment analysis libraries in python to detect sentiments like TextBlob to the preprocessed data in order to classify the opinions into positive, negative, or neutral. Topic modelling was also implemented using Latent Dirichlet Allocation to identify trending topics or famous topics regarding organic foods. The methodology of this project provides a comprehensive framework for analyzing public sentiment towards organic food using advanced computational techniques. The results of this project can provide valuable insights into consumer behavior and preferences regarding organic food. The results can also help identify the strengths and weaknesses of organic food as well as the opportunities and threats for its market growth (ScienceDirect, 2018).

**3.2 Research Design**

The research design is a mixed-method approach that combines qualitative and quantitative data analysis. The study uses a three-stage process: data collection, data preprocessing, and sentiment analysis.

**3.3 Data Collection**

Acknowledging that most of the previous works made on this project topic had it data source from Twitter and Online forums. However, the data used for this project was from Reddit. The data was collected using web scraping techniques in python and using Reddit API called PRAW, and python libraries like beautifulsoup4 and Selenium. The time frame for data collection spans from January to December 2021 to ensure a representative sample of current public sentiment.

**3.4 Data Preprocessing**

Once the data has been collected, it needs to be preprocessed before it can be analyzed. This may involve removing stop words, stemming or lemmatizing words, and correcting spelling errors. Stop words are common words that do not add much information to the text, such as "the", "is", and "are". Stemming and lemmatizing are techniques for reducing words to their root form. This can help to improve the accuracy of the analysis.

This involves cleaning and preparing the data for analysis. This may include removing stop words, stemming or lemmatizing words, and correcting spelling errors

* **Cleaning:** Removal of irrelevant content, such as ads and spam.
* **Tokenization:** Breaking down text into individual words or phrases.
* **Normalization:** Converting text to a standard format (e.g., lowercase).
* **Stop-word Removal:** Eliminating common words that add no value to sentiment analysis.
* **Stemming:** Reducing words to their root form.

**3.5 Sentiment Analysis**

The third step in the methodology is sentiment analysis. The preprocessed data was analyzed using a sentiment analysis algorithm. The algorithm used in this project is the VADER (Valence Aware Dictionary and sEntiment Reasoner) algorithm embedded in the Textblob and SentimentIntensityAnalyzer python libraries. The VADER algorithm is a rule-based sentiment analysis tool that is specifically designed for social media text. These libraries will help classify posts from Reddit into positive, neutral or negative sentiments.

**3.6 Topic Modelling**

Topic modeling is a statistical technique for identifying the latent (hidden) topics in a corpus of text. It is a powerful tool for analyzing large amounts of text data, such as social media posts, news articles, and product reviews.

In the context of the research topic "Analyzing public opinions and sentiment analysis towards organic food", topic modeling can be used to:

* Identify the most common topics that people are discussing about organic food.
* Identify the sentiment of people's opinions on organic food.
* Track changes in public opinion on organic food over time.
* Identify groups of people who share similar opinions on organic food.

For example, topic modeling could be used to identify the following topics in a corpus of Reddit posts about organic food:

* Health benefits of organic food
* Sustainability of organic farming
* Taste and quality of organic food
* Price of organic food
* Availability of organic food

Topic modeling could then be used to identify the sentiment of people's opinions on each of these topics. For example, the algorithm could calculate the percentage of Reddit posts in each topic that are positive, negative, or neutral.

Topic modeling can also be used to track changes in public opinion on organic food over time. For example, the algorithm could be used to compare the sentiment of Reddit posts about organic food from different time periods. This could reveal whether public opinion on organic food is becoming more positive or more negative over time.

Finally, topic modeling can be used to identify groups of people who share similar opinions on organic food. For example, the algorithm could be used to identify clusters of users who are all posting about the health benefits of organic food. This information could be used by marketers to target their advertising campaigns to specific groups of people.

Overall, topic modeling is a powerful tool for analyzing public opinions and sentiment analysis towards organic food. It can be used to identify the most common topics that people are discussing about organic food, the sentiment of people's opinions on organic food, changes in public opinion on organic food over time, and groups of people who share similar opinions on organic food.

Latent Dirichlet Allocation (LDA) is a topic modeling algorithm that is widely used in natural language processing (NLP) tasks such as text classification, sentiment analysis, and document summarization. LDA works by identifying the latent topics that are present in a corpus of text. A latent topic is a group of words that are frequently associated with each other. For example, a latent topic on organic food might include words such as "organic", "natural", "healthy", and "sustainable".

LDA works by assuming that each document in a corpus is a mixture of latent topics. The document-topic mixture is represented by a probability distribution over topics. The algorithm then learns the probability distribution of topics for each document and the probability distribution of words for each topic.

To learn the topic model, LDA uses a variational inference algorithm. The algorithm starts by initializing the topic model parameters. Then, it iteratively updates the parameters until they converge. The algorithm converges when the change in the parameters is below a certain threshold.

Once the topic model has been learned, it can be used to identify the latent topics in a new document. To do this, the algorithm calculates the probability of each topic for the new document. The topic with the highest probability is the most likely topic for the document.

LDA can be used to analyze public opinions and sentiment analysis towards organic food in a number of ways. For example, LDA can be used to identify the most common topics that people are discussing about organic food. LDA can also be used to identify the sentiment of people's opinions on organic food.

LDA can be a powerful tool for analyzing public opinions and sentiment analysis towards organic food. However, it is important to note that LDA is an unsupervised learning algorithm. This means that it does not require any labeled data to train the model. However, this also means that LDA can be more difficult to interpret than supervised learning algorithms.

**3.6 System Architecture**

The system architecture comprises the following components:

* **Input Layer:** Receives raw data from social media platforms.
* **Preprocessing Module:** Processes the data as per the steps outlined in section 3.4.
* **Analysis Module:** Applies sentiment analysis techniques to the preprocessed data.
* **Topic Modelling Module:** Get different topics surrounding public opinions towards organic food.
* **Output Layer:** Generates reports on sentiment analysis results.

**3.7 Ethical Considerations**

The study adheres to ethical standards by anonymizing personal information and ensuring data security.

**3.8 Conclusion**

This methodology provides a robust framework for analyzing public sentiment towards organic food using advanced computational techniques.

References:

1. Journal of Big Data (2021). [Modeling the public attitude towards organic foods: a big data and text mining approach 1](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00551-6).
2. SpringerLink (2022). [Study of Public Sentiment Using Social Media for Organic Foods in Pre-Covid and Post-Covid Times Worldwide 2](https://link.springer.com/chapter/10.1007/978-3-031-17746-0_11).
3. SpringerLink (2019). [Sentiment analysis using deep learning architectures: a review 3](https://link.springer.com/article/10.1007/s10462-019-09794-5).
4. SpringerLink (2023). [Sentiment Analysis Using an Improved LSTM Deep Learning Model 4](https://link.springer.com/chapter/10.1007/978-981-19-9225-4_19).
5. ACL Anthology (2023). [A Hybrid Deep Learning Architecture for Sentiment Analysis](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00551-6) [5](https://aclanthology.org/C16-1047/).
6. SpringerLink (2021). [An efficient sentiment analysis methodology based on long 6](https://link.springer.com/article/10.1007/s40747-021-00436-4).
7. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(4-5), 993-1022.
8. Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of using big data for social science research. Princeton University Press.
9. Ho, T. K. (2013). Topic modeling: Foundations and extensions. CRC Press.
10. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109/8122