NLP AND WEB SCRAPING FOR NEPALI AGRICULTURAL DATA ANALYSIS

Tribhuvan University
Institute of Engineering
Thapathali Campus

Informatics and Intelligent System Engineering
Department of Computer Science and Engineering

With the guidance of:

Asst. Professor Er. Praches Acharya

Presented By:

Sameer Gautam

[THA079MSISE14]

Outline

- Motivation
- Background
- Problem Statement
- Objective of Project
- Scope of Project
- Originality of Project
- Potential Applications
- Literature Review
- Methodology
- Results
- Discussion and Analysis
- Future Enhancements
- Conclusion
- Project Schedule (Gantt Chart)
- References

Motivation

- Nepal's agricultural sector faces challenges in modernization and sustainability.
- Communication barriers and limited access to modern practices hinder progress.
- Social media platforms offer a potential solution to bridge the communication gap.
- Existing methods struggle with Nepal's linguistic diversity, overlooking Nepali language.
- This project aims to empower farmers by harnessing social media data and NLP techniques tailored to Nepal's linguistic landscape.

Background

- Agriculture employs a significant portion of Nepal’s population.
- Farmers in rural areas primarily communicate in the Nepali language.
- Limited access to agricultural information in Nepali hinders their ability to adopt modern practices.
- Lack of time to operate social media farmers lack modern knowledge, solution to that problem a project should be a able to respond with suggestion related to agriculture.
- NLP techniques can be used to analyze social media data and extract valuable insights.

Problem Statement

- Lack of tools and methodologies to understand and address the needs of Nepalese farmers.
- Existing approaches often overlook the linguistic richness of Nepali, resulting in a mismatch between interventions and the realities on the ground.
- Need for a localized approach that leverages social media data and NLP techniques to know situation of Nepali farmers and agricultural information.

Objective of Project

- Deliver actionable insights to Nepali farmers by analyzing social media data using NLP techniques.
- Overcome linguistic and technical challenges in processing Nepali social media content.

Scope of Project

- Focuses on analyzing social media data (Twitter) in the Nepali language.
- Employs NLP techniques to uncover insights into farmers' challenges and priorities.
- Develops NLP models for sentiment analysis, topic modeling, and information extraction.
- Creates user-friendly tools and resources to disseminate agricultural information for proper analysis.
- Aims to help farmers with accessible and relevant agricultural suggestions.

Originality of Project

- Addresses the unique challenge of limited access to agricultural information faced by Nepali-speaking farmers.
- Focuses on analyzing Nepali language data on social media, offering a novel method to understand farmers' needs.
- Unlike previous research, this project specifically targets Nepali-speaking farmers who have not been included in prior studies.

Potential Applications

- **Precision Farming**: Provide real-time, localized suggestions to farmers for better farming.
- **Market Engagement**: Empower farmers to strategize production and marketing efforts based on market trends.
- **Policy Decisions**: Inform evidence-based policies tailored to the needs of Nepalese farming communities.

Literature Review [1]

	1	2	3	4	5
Key Findings:	Social media reveals agricultural trends, but current sentiment analysis tools may fall short.	Social media analytics reveal agricultural stakeholders' emotional and behavioral responses during crises.	Perceived credibility, reference group influence, perceived infotainment, and perceived usefulness positively impact agriculturists' adoption of social media marketing.	Sentiment analysis has potential in agriculture but is underutilized. Machine learning is the most common approach.	Machine learning and NLP can extract valuable information from social media for natural hazard research.
Method ology	Text mining of Twitter data from farmers in Oklahoma Panhandle	Sentiment and emotion analysis of tweets related to agriculture in India during COVID-19 lockdown.	Survey of 320 agriculturalists in Tamilnadu, South India	Literature review of sentiment analysis research in agriculture.	Web scraping and Word2Vec model applied to ResearchGate data.

Literature Review [2]

1	2	3	4	5
Strength: Novel approach, data-driven insights	Real-time data, use of machine learning.	Use of established theoretical frameworks, large sample size.	Comprehensive overview, identifies future research areas.	Innovative approach, automated data collection.
Limitations: Limited to Twitter data and specific region, need for domain-specific lexicons.	Potential bias in Twitter data, focus on India, lack of qualitative analysis.	Focus on South India, potential self-selection bias, lack of qualitative data.	Excludes research in related fields, lacks in-depth analysis of individual studies	Limited to one platform and keyword, model accuracy can be improved.

Methodology[1]

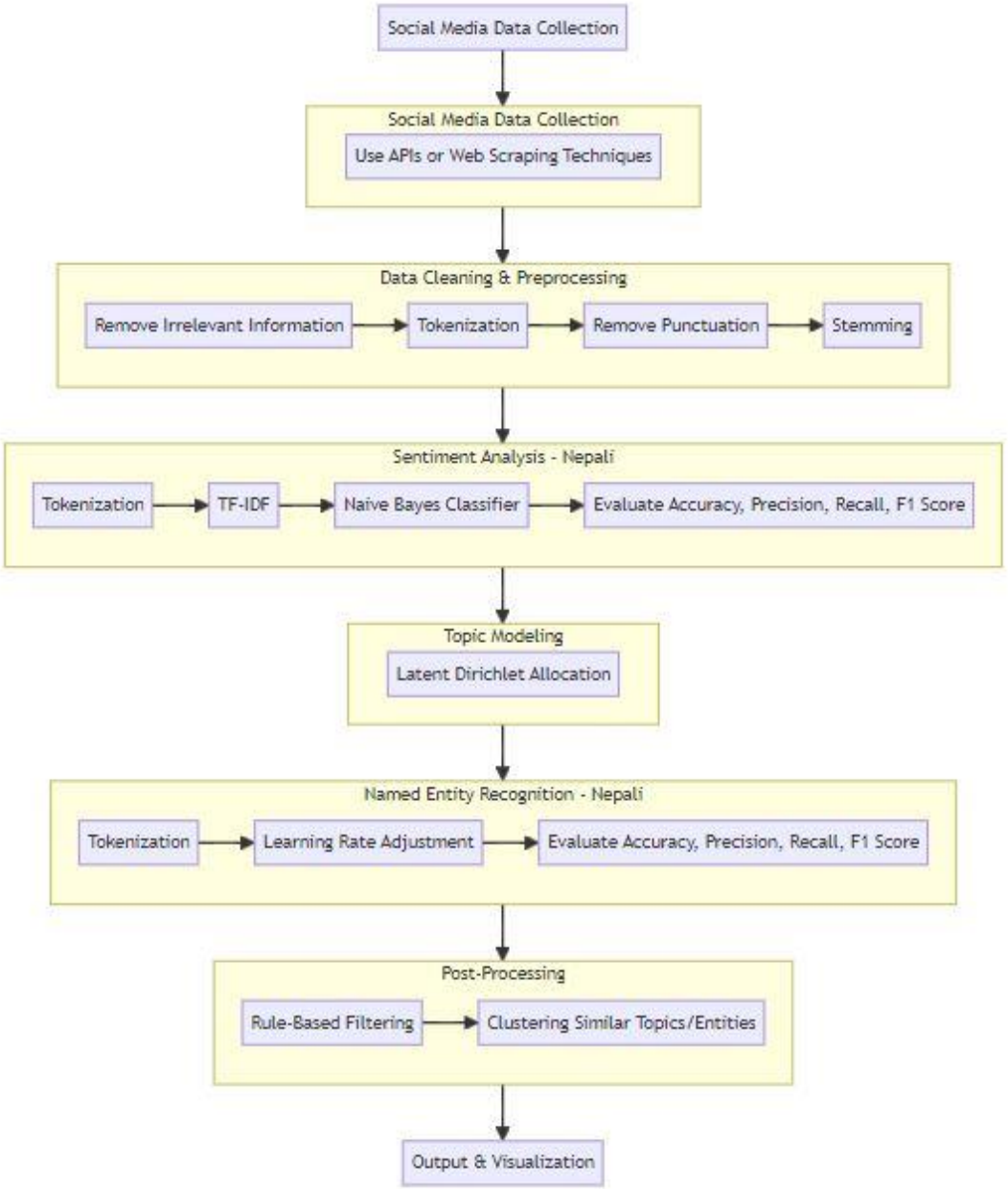
- **Data Collection:**
  - Gather Nepali language social media posts related to agriculture from platform(Twitter).
  - Utilize APIs and web scraping techniques for data collection.
- **Data Preprocessing:**
  - Clean the data by removing noise, irrelevant information, and standardizing formats.
  - Tokenize the text into words or phrases.

Methodology [2]

- **NLP Model Development:**
  - Train sentiment analysis models to classify posts as positive and negative.
  - Develop topic modeling algorithms to identify key themes and discussions.
  - Implement named entity recognition to extract relevant entities like entities, crops, and problems.
- **Tool Development:**
  - Create user-friendly tools to visualize and interpret the analysis results.
  - Develop interfaces for stakeholders to access relevant information.

Methodology [3]

- **Validation and Refinement:**
  - Evaluate model performance using metrics like accuracy, precision, recall, and F1-score.
  - Refine models based on evaluation results and iterate on the process.
- **Deployment and Dissemination:**
  - Deploy the developed tools and resources to know problem faced by Nepalese farmers.
  - Disseminate findings and recommendations to relevant stakeholders in the agricultural sector.



Block Diagram

Results[1]

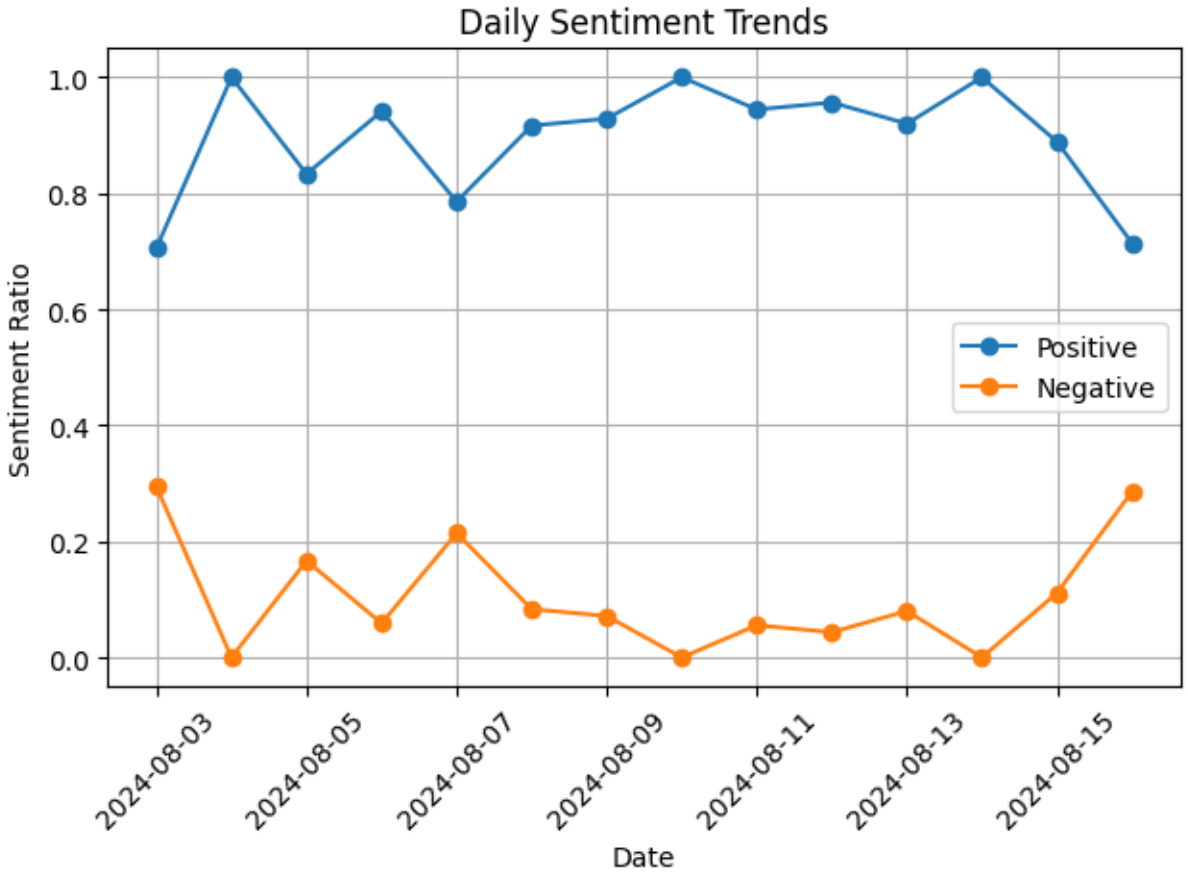
- Preprocessing
  a. Removing emoji, URLs, hashtags, usernames and punctuation.
  b. Tokenization

Text
नेपालका कृषकहरूको समस्या बढ्दै गइरहेको छ। कृषकहरूको आवाजलाई सुन्न आवश्यक छ।
बढ्दै गइरहेको कृषकहरूको समस्या छ बुझिदिने कोहि छैन। #कृषकसमस्या #नेपाल
जलवायु परिवर्तनले कृषि क्षेत्रमा ठूलो असर पुर्याएको छ। कृषकहरूलाई सहयोग आवश्यक छ। #कृषकसमस्या #नेपाल
धान उत्पादन घट्दैछ र कृषकहरू निराश छन्। #कृषकसमस्या #धान
खडेरीले गर्दा अन्न उत्पादनमा समस्या उत्पन्न भएको छ। #कृषकसमस्या

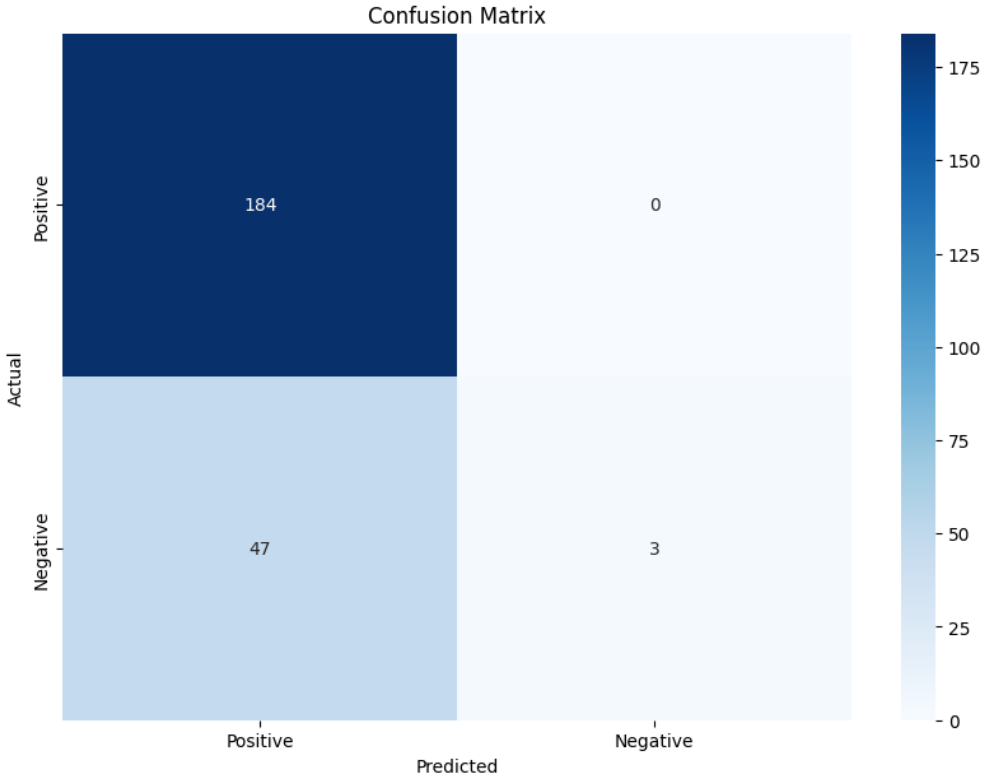
```
Dictionary size: 29
Sample dictionary tokens: {'असर': 0, 'उपज': 1, 'यस': 2, 'कम': 3, 'भएन': 4, 'थप': 5, 'आए': 6,
Corpus size: 84
```

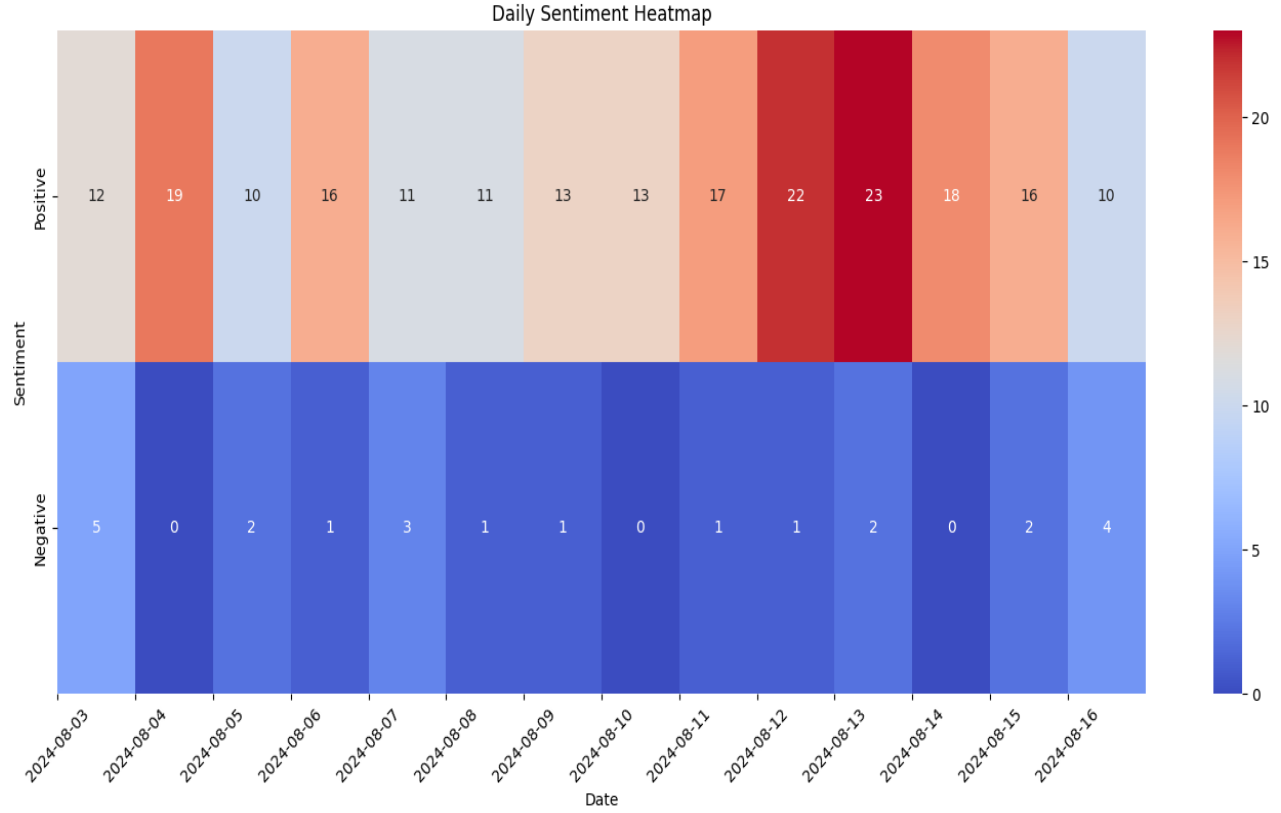
Results[2]

- Sentiment Analysis:
  a. Identify the overall sentiment of farmers towards specific agricultural practices, policies, or events.
  b. The model might classify social media posts into categories like ”positive” and ”negative,”.



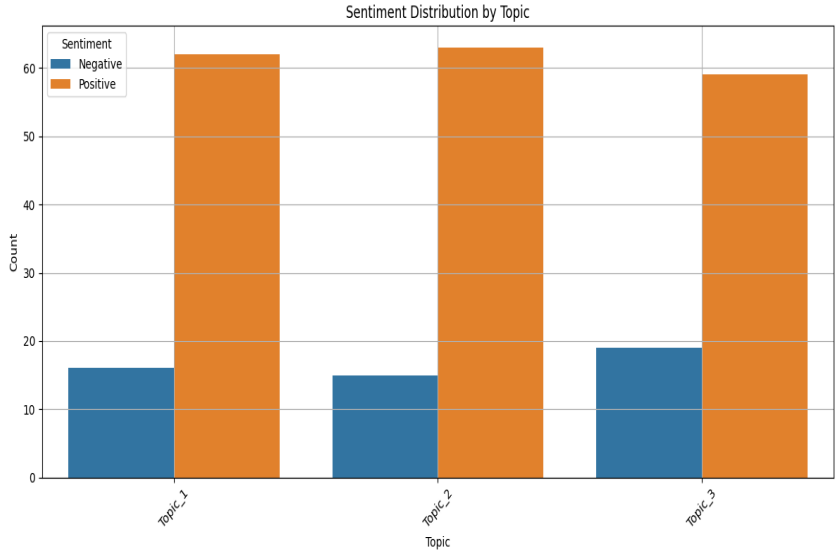
Results[2]

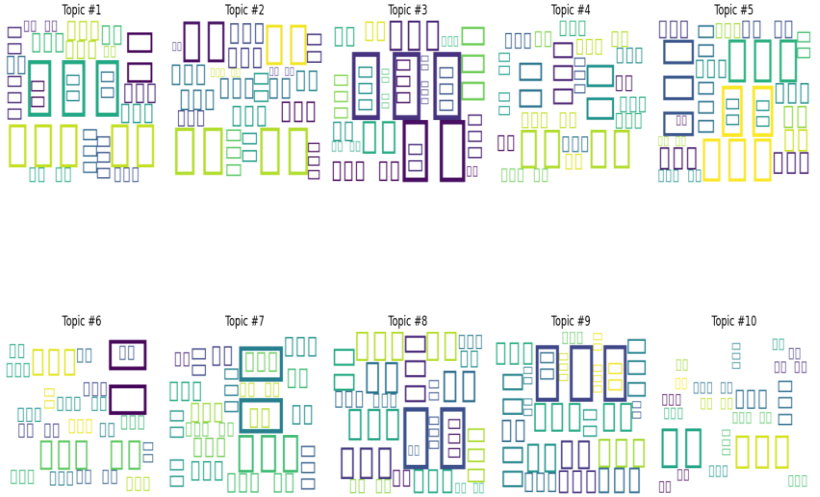

Confusion Matrix
Positive
Negative
184
0
47
3
Actual
Positive
Negative
Predicted
175
150
125
100
75
50
25
0


Daily Sentiment Heatmap
Positive
Negative
Sentiment
12 19 10 16 11 11 13 13 17 22 23 18 16 10
5 0 2 1 3 1 1 0 1 1 2 0 2 4
2024-08-03
2024-08-04
2024-08-05
2024-08-06
2024-08-07
2024-08-08
2024-08-09
2024-08-10
2024-08-11
2024-08-12
2024-08-13
2024-08-14
2024-08-15
2024-08-16
Date
20
15
10
5
0

Results[3]

- Topic Modeling:
  a. Discover the most frequently discussed topics among farmers.
  b. Example: Top topics include crop diseases, irrigation techniques, and government subsidies.





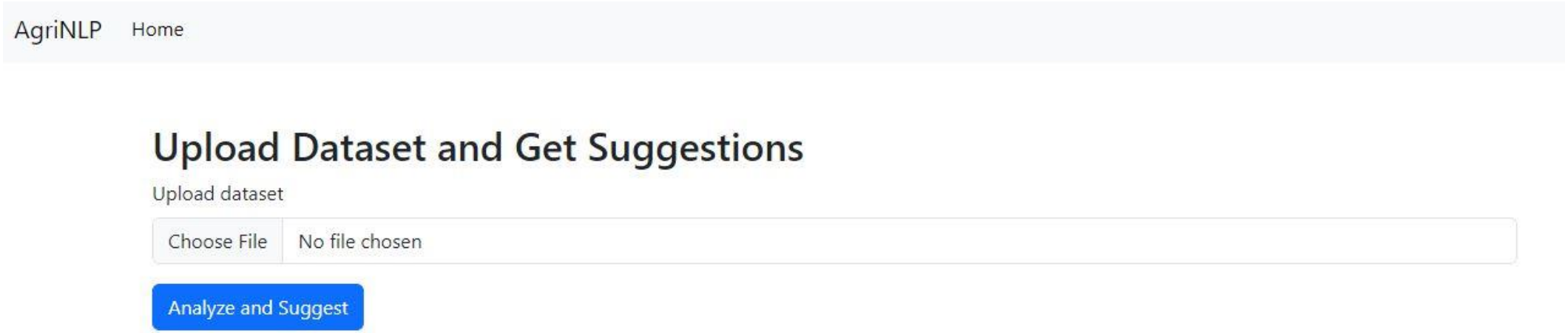
Results [4]

- Named Entity Recognition:
  a. Extract relevant entities like crop types (e.g., rice, maize) and entities (e.g. government, customer).

```
Most common entities:
ENTITY: 16
CROP: 7
```

Result[5]

- User Interface
  - Index.html from where we can upload datasets to analyze.



Result[6]

- User Interface
  - Suggestion.html, where suggestion is provided on the basis of dataset provided.

Discussion and Analysis [1]

- Project Findings: For the summary we obtained suggestion as expected, which can be used to help farmers and stakeholders.
- NLP Technique Efficacy: Evaluate the effectiveness of the Natural Language Processing techniques used is good.
- Social Media Data: the quality, relevance, and challenges of using social media data.
- Linguistic Challenges: the impact of linguistic diversity on data processing and information dissemination.
- Farmer Engagement: the level of farmer engagement and response to the provided information.

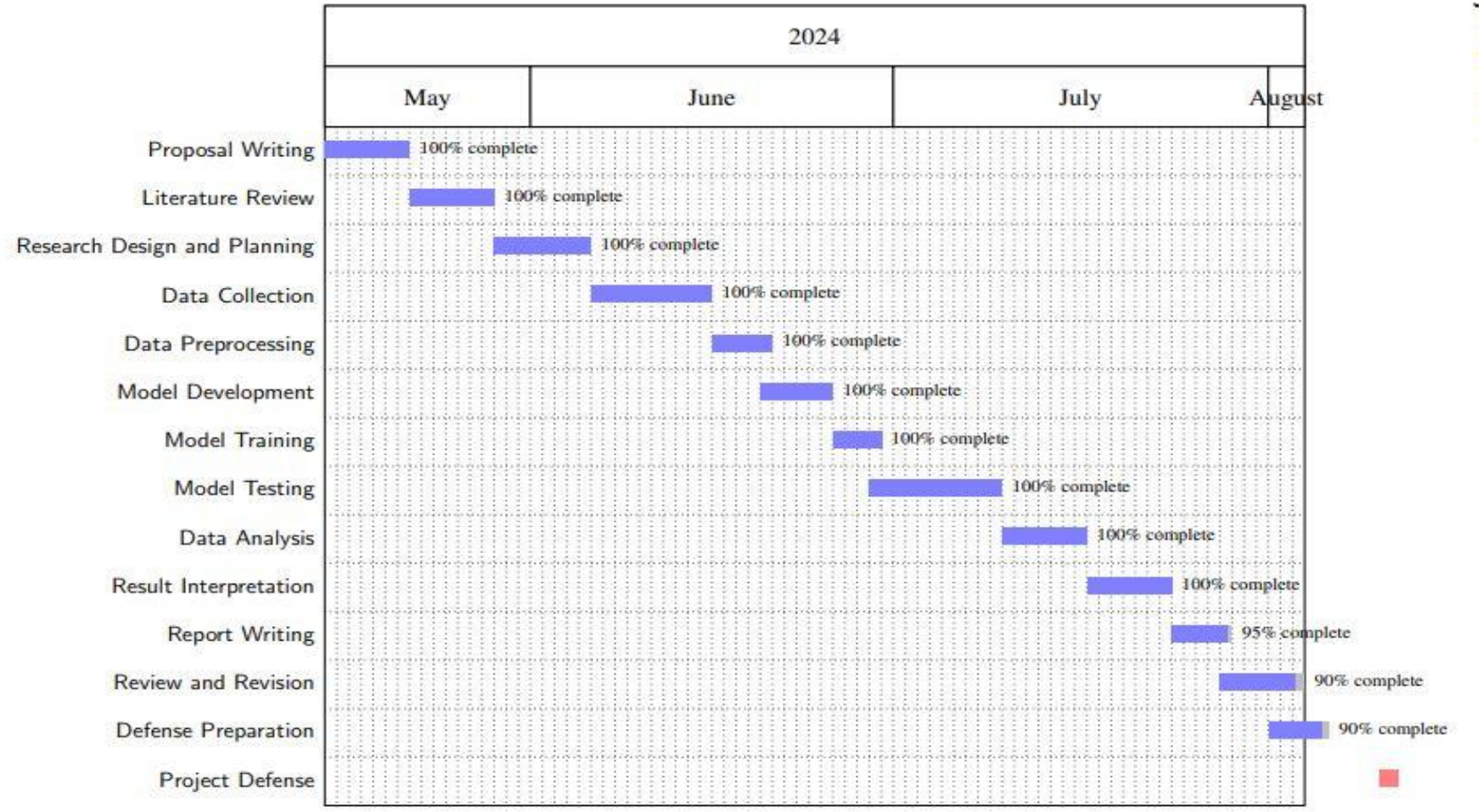
Discussion and Analysis [2]

- Comparative Analysis: this approach with traditional methods of information dissemination in agriculture is better.
- Limitations: less previous research and better NLP techniques.
- Sustainability: It can analyze the long-term sustainability of using digital platforms for agricultural knowledge dissemination.

Future Work

- Improve NLP techniques for Nepali language.
- Extend the analysis to other regional languages.
- Collaboration with local stakeholders for practical applications.

Tentative Timeline (Gantt Chart)



References

- Bagheri, A., Taghvaeian, S., & Delen, D. (2023). A text analytics model for agricultural knowledge discovery and sustainable food production: A case study from Oklahoma Panhandle. Decision Analytics Journal, 9.
- Devienne, J. A. P. M. (2023). Use of social media and natural language processing (NLP) in natural hazard research. arXiv preprint arXiv:2304.08341.
- Novak, J., Nemecek, J., & Hosek, P. (2021). Sentiment analysis in agriculture. Agris on-line Papers in Economics and Informatics, 13(1), 121–129.
- Palaniswamy, V., & Raj, K. (2022). Social media marketing adoption by agriculturists: A TAM-based study. International Journal of Professional Business Review, 7(3).
- Singh, M., Singh, A., Bharti, S., Singh, P., & Saini, M. (2022). Using social media analytics and machine learning approaches to analyze the behavioral response of agriculture stakeholders during the COVID-19 pandemic. Sustainability, 14(23).

Thank You!