DWIT COLLEGE

DEERWALK INSTITUTE OF TECHNOLOGY

A WEB BASED SENTIMENT-CENTRIC NEWS PORTAL USING SVM AND CNN WITH TIME SERIES ANALYSIS THROUGH LDA

A FINAL YEAR PROJECT - PROPOSAL REPORT

Submitted to

Department of Computer Science

DWIT College

Submitted by

Prayusha Acharya

06/08/2024

TABLE OF CONTENTS

LIST OF FIGURES: III

LIST OF TABLES IV

LIST OF ABBREVIATIONS V

1. INTRODUCTION 1

2. PROBLEM STATEMENT 2

3. OBJECTIVES 3

4. METHODOLOGY 4

4.1. REQUIREMENT IDENTIFICATION 4

4.1.1. STUDY OF EXISTING SYSTEM 4

4.1.2. LITERATURE REVIEW 6

4.1.3. REQUIREMENT ANALYSIS 7

4.2. FEASIBILITY STUDY 8

4.2.1. TECHNICAL FEASIBILITY 8

4.2.2. OPERATIONAL FEASIBILITY 8

4.2.3. ECONOMIC FEASIBILITY 8

4.2.4. SCHEDULE FEASIBILITY 8

4.3. HIGH LEVEL DESIGN OF SYSTEM 9

4.3.1. SYSTEM DEVELOPMENT MODEL 9

4.3.2. FLOWCHART 10

4.3.3. WORKING MECHANISM OF PROPOSED SYSTEM 12

4.3.4. ALGORITHMS 13

5. EXPECTED OUTCOMES 16

REFERENCES 17

LIST OF FIGURES

Figure 1: Home Page of The Rising Nepal 4
Figure 2: Home Page of Online Khabar 5
Figure 3: Use Case Diagram of the Sentiment-centric News Portal 7
Figure 4: Gantt Chart for the Sentiment-centric News Portal 9
Figure 5: Waterfall Model 10
Figure 6: Flow Chart for Users of the Sentiment-centric News Portal 10
Figure 7: Flowchart for Implementation of Algorithms 11
Figure 8: ER Diagram for Sentiment-centric News Portal 12
Figure 9: Working of SVM 13
Figure 10: Working of CNN 13
Figure 11: Working of SVM with CNN 14
Figure 12: Working of LDA 14
Figure 13: Blueprint of LDA Model 14
Figure 14: Working of Porter Stemming Algorithm 15

LIST OF TABLES

Table 1: Gantt Chart Table for the Sentiment-centric News Portal 9

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Algorithm
SVM	Support Vector Machine
TF	Term Frequency

1. INTRODUCTION

The "Sentiment-centric News Portal" project aims to revolutionize how we consume and analyze news by leveraging advanced natural language processing and machine learning techniques. This innovative system will crawl multiple Nepali news portals, automatically translating the content into English to ensure accessibility for a broader audience. The core of the sentiment analysis will be powered by algorithms combining Support Vector Machines (SVM) with Convolutional Neural Networks (CNN), offering high accuracy in discerning the emotional tone of news articles. By performing this sentiment analysis on the aggregated news, the platform will provide users with a deep understanding of the emotional context surrounding various events and topics in Nepal and beyond.

The system will also categorize news articles by genre, such as sports, politics, international affairs, and finance, allowing for more targeted information consumption. A key feature of this project is its time-series analysis capability, which employs Latent Dirichlet Allocation (LDA) to track and visualize sentiment trends over time. This approach will offer valuable insights into the evolution of public opinion on specific topics, entities, or general news trends within the Nepali-speaking world.

By combining these diverse functionalities and focusing on Nepali news sources, the Sentiment-centric News Portal will empower users to navigate the complex modern information landscape of Nepal with greater clarity. This tool will be particularly valuable for researchers, policymakers, and anyone interested in gaining insights into Nepali public opinion and news trends, ultimately fostering more informed decision-making and public discourse both within Nepal and internationally.

2. PROBLEM STATEMENT

The digital age has escorted in an era of information overload, particularly challenging for Nepali news content. Readers, researchers, and decision-makers face significant hurdles in efficiently processing and understanding the vast volume of news articles published daily. These challenges include language barriers for non-Nepali speakers, difficulty in discerning the emotional context of news, and inefficient classification of news genres. Moreover, a growing debate has emerged regarding the impact of constant news consumption on individuals' well-being. While news can illuminate truth and keep the public informed about global events, it often presents a one-sided view of reality. Health experts argue that frequent news exposure can be toxic, constantly triggering the limbic system and potentially causing stress and anxiety. Current news systems fail to address these concerns, lacking the ability to provide balanced perspectives or temporal analysis of sentiment trends in Nepali news. Additionally, existing sentiment analysis tools struggle with the nuances of the Nepali language.

The Sentiment-centric News Portal project aims to address these complex issues by developing a comprehensive system that leverages advanced machine learning techniques, including CNN with SVM, to provide a more accessible, and insightful view of Nepali news content. This approach will not only bridge the gap between local news sources and global understanding but also offer users tools to manage their news consumption more mindfully, balancing the need for information with personal well-being.

3. OBJECTIVES

- To create an advanced sentiment analysis model using CNN with SVM that accurately interprets the emotional tone of news articles.
- To develop a time-series analysis feature using LDA to track and visualize sentiment trends over time.

4. METHODOLOGY

4.1. REQUIREMENT IDENTIFICATION

4.1.1. STUDY OF EXISTING SYSTEM

In analyzing the current landscape of Nepali news portals, two prominent systems were examined: The Rising Nepal and Online Khabar. This study reveals significant gaps in the existing news delivery systems, highlighting the need for our proposed Sentiment-centric News Portal.

a. The Rising Nepal

The Rising Nepal positions itself as the nation's first English broadcast news portal. While this claim underscores its pioneering role in Nepal's digital news landscape, the system presents several limitations: -

- Language Limitation: Despite being a Nepali news broadcaster, all content is exclusively in English. This approach potentially alienates a significant portion of the Nepali-speaking audience who prefer news in their native language.
- Technical Issues: The portal suffers from navigational problems, with some buttons leading to faulty pages. This negatively impacts user experience and accessibility of information.
- Lack of Sentiment Analysis: Although the portal asks readers how they feel after reading news articles, it does not implement any functional sentiment analysis. This represents a missed opportunity to gauge and analyze public opinion.

Figure 1: Home Page of The Rising Nepal [1]

b. Online Khabar

Online Khabar claims to be Nepal's number one news portal in the Nepali language. While it caters well to the Nepali-speaking audience, it also has notable drawbacks:

- Language Barrier: The absence of English translations makes the content inaccessible to non-Nepali speakers, limiting its reach and utility for a global or diverse local audience.
- Lack of Sentiment Analysis: Like the Rising Nepal, Online Khabar does not incorporate sentiment analysis in its news delivery system, missing out on valuable insights into public sentiment and trends.

Figure 2: Home Page of Online Khabar [2]

These findings underscore the need for a more comprehensive and technologically advanced news portal that addresses the limitations of existing systems. Our proposed Sentiment-centric News Portal aims to bridge these gaps by offering multilingual support, implementing robust sentiment analysis, and providing a more inclusive and insightful news consumption experience

4.1.2. LITERATURE REVIEW

Emotional Impact of News on Social Judgments (Baum, 2021) Baum's study in Social Cognitive and Affective Neuroscience explores how emotional news content affects social judgments, independent of perceived media credibility. This research underscores the importance of considering emotional content in news analysis, as it can significantly influence public opinion and decision-making processes [3].

Social Media News Headlines and Well-Being (Mousoulidou et al., 2024) Mousoulidou and colleagues investigated the influence of social media news headlines on well-being, examining emotional states, emotion regulation, and resilience. Their work in the European Journal of Investigation in Health, Psychology and Education highlights the potential psychological impacts of news consumption, aligning with our project's aim to provide more balanced and mindful news delivery [4].

Ensemble Machine Learning for Twitter Sentiment Analysis (Radiuk et al., 2022) Presented at the 6th International Conference on Computational Linguistics and Intelligent Systems, Radiuk et al.'s research proposes an ensemble machine learning approach for Twitter sentiment analysis. It was discovered that the CNN model with the SVM classifier demonstrated the best performance than any other techniques [5].

N-Gram, TF-IDF, and Ensemble Methods in Sentiment Classification (Rahman et al., 2020) Rahman and colleagues compared various methods for sentiment classification, including N-Gram, TF-IDF, and ensemble techniques. Their findings, presented at the International Conference on Cyber Security and Computer Science, In this investigation, the authors explored different n-gram models (including unigrams and bigrams) along with term frequency-inverse document frequency (TF-IDF) features, informing our choice of using CNN with SVM for sentiment analysis [6].

Multilingual Sentiment Analysis (Das et al., 2023) Das et al.'s study in Heliyon emphasized the critical role of text preprocessing techniques, particularly highlighting the Porter Stemming Algorithm. This preprocessing step proved essential in enhancing the performance and accuracy of machine learning models [7].

LDA Topic-Based Sentiment Analysis in Tourism (Ali et al., 2022) Ali and colleagues' work, published in MethodsX, demonstrates the application of Latent Dirichlet Allocation (LDA) for topic-based sentiment analysis in the tourism sector. Their approach to combining topic modeling with sentiment analysis aligns with our project's goal of implementing time-series analysis using LDA to track sentiment trends in news content [8].

4.1.3. REQUIREMENT ANALYSIS

a. Functional requirements

- The system shall automatically translate the extracted Nepali content into English.
- The system shall categorize news articles into respective genres such as sports, politics, international affairs, and finance.
- The system shall classify the sentiment of news articles as positive, negative, or neutral.
- The system shall visualize sentiment trends over time for specific topics and entities.
- The user shall be able to filter news articles by sentiment, genre, and date.
- The user shall be able to login to the website to view their saved news articles.

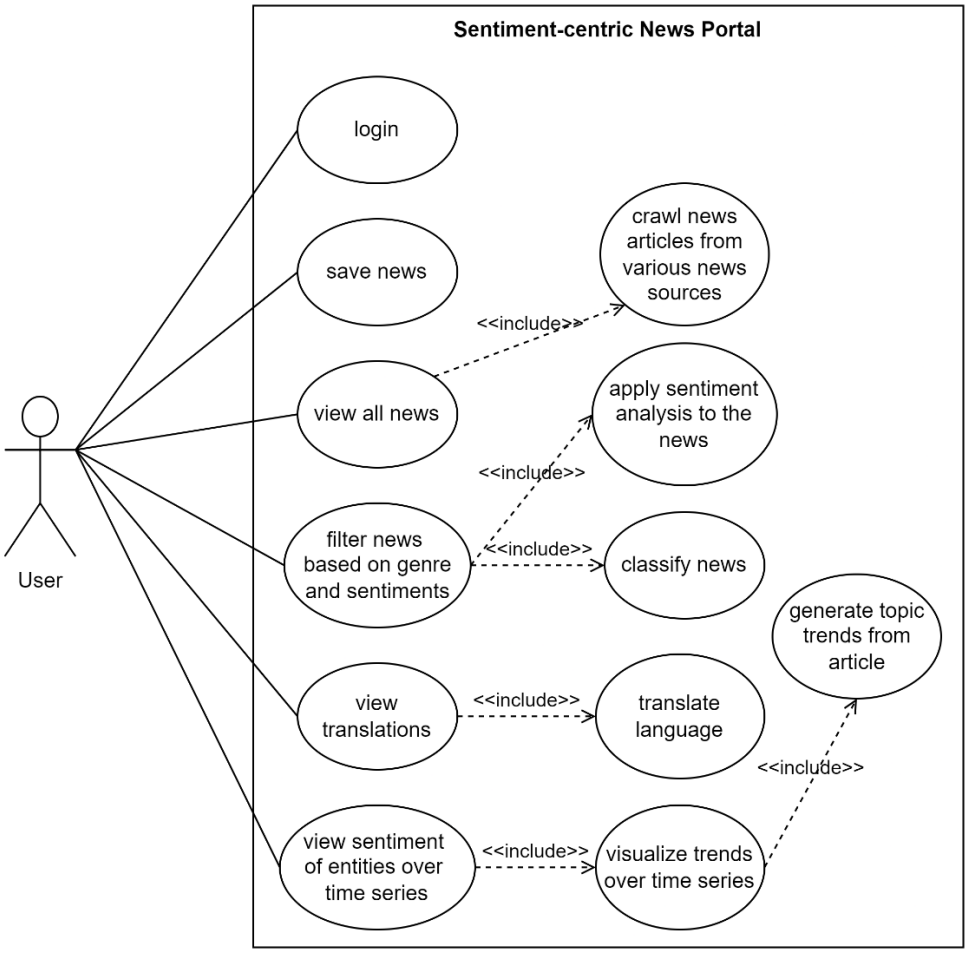

Figure 3: Use Case Diagram of the Sentiment-centric News Portal

In this diagram, we have an actor who interacts with the news portal system. The actor could be any individual engaging with the platform. This diagram outlines the functionalities available to users, with an emphasis on sentiment analysis for filtering and viewing news content.

b. Non-functional requirements

- The system must process and analyze news articles in real-time or near real-time.
- The sentiment analysis and translation processes must achieve high accuracy rates.
- The categorization and time-series analysis must be precise and reliable.
- The translation process must ensure the translated text retains the original meaning and context.

4.2. FEASIBILITY STUDY

4.2.1. TECHNICAL FEASIBILITY

The proposed application Sentiment-centric News Portal is a web-based application that uses python-based frameworks. JavaScript, CSS and HTML pages for development of front end while Python is used in backend. Internet connection is required to function properly. For the operation, the application supports Windows and MAC Operating System. The application can be declared technically feasible as all the technical resources are easily available and accessible.

4.2.2. OPERATIONAL FEASIBILITY

The proposed system will be designed in a way such that a layman can operate it without having to care about technical knowledge. It will enable a person to read news from a single Sentiment Analysis of News 11 website and gives users a clear understanding of the nature of the news beforehand. The simple design makes it an easy-to-use application. This application can be accessed from any place that has an internet connection. Hence, it is understood that Sentiment-centric News Portal is an operationally feasible application.

4.2.3. ECONOMIC FEASIBILITY

For the sentiment analysis of Nepali news headlines, there was no monetary expenses and was built by using the existing resources. Therefore, this project can be declared too economically feasible.

4.2.4. SCHEDULE FEASIBILITY

The schedule feasibility of the project was analyzed using Gantt-chart representation for different tasks involved in the development of the system. The project activities are planned to be completed within following time frames: -

Table 1: Gantt Chart Table for the Sentiment-centric News Portal

S.N.	Tasks	Duration
1	Research and Requirement Analysis	7 days
2	Dataset Collection	6 days
3	Create Web Scraping Algorithm	6 days
4	Create Sentiment Analysis Model	15 days
5	Create Article Classification Model	10 days
6	Create Web UI for Output	9 days
7	Testing	4 days
8	Documentation	5 days

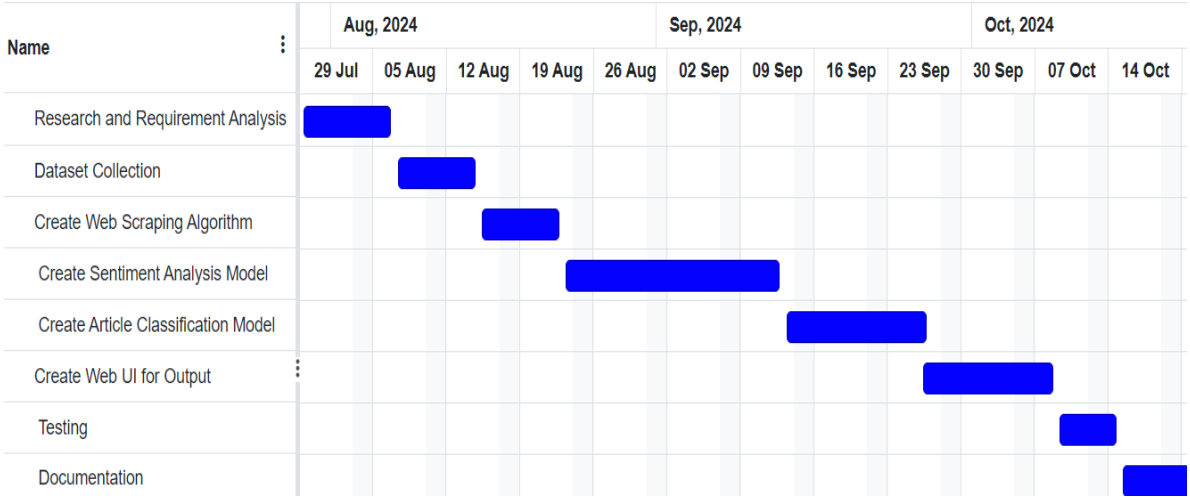

Figure 4: Gantt Chart for the Sentiment-centric News Portal

4.3. HIGH LEVEL DESIGN OF SYSTEM

4.3.1. SYSTEM DEVELOPMENT MODEL

The waterfall model was selected for the development of the Sentiment-centric News Portal due to its systematic progression through development phases. This approach ensures a clear sequence of steps, from requirements analysis to deployment, aligning with the project's structured goals. The emphasis on thorough planning, well-defined stages, and minimal changes post-development suits the project's characteristics, where a stable set of requirements and a systematic approach are prioritized.

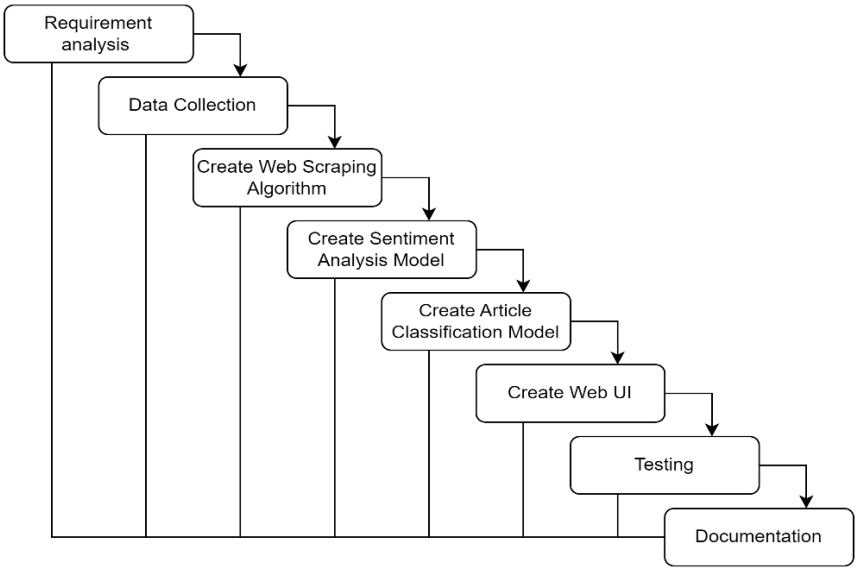

Figure 5: Waterfall Model

4.3.2. FLOWCHART

The following diagram depicts the flow of activities in the Sentiment-centric News Portal:

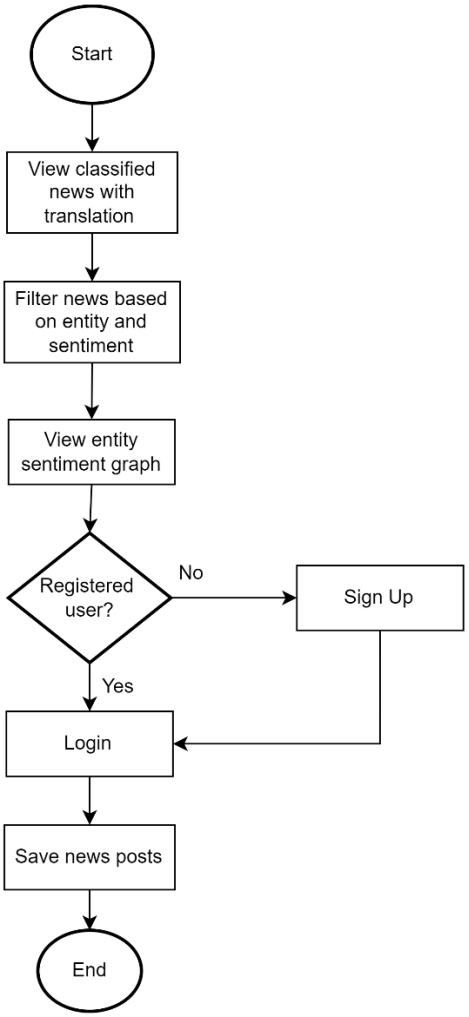

Figure 6: Flow Chart for Users of the Sentiment-centric News Portal

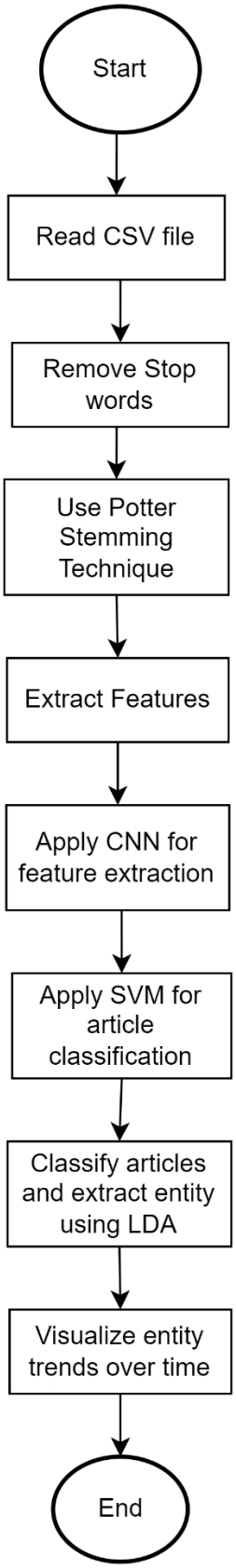

Figure 7: Flowchart for Implementation of Algorithms

4.3.3. WORKING MECHANISM OF PROPOSED SYSTEM

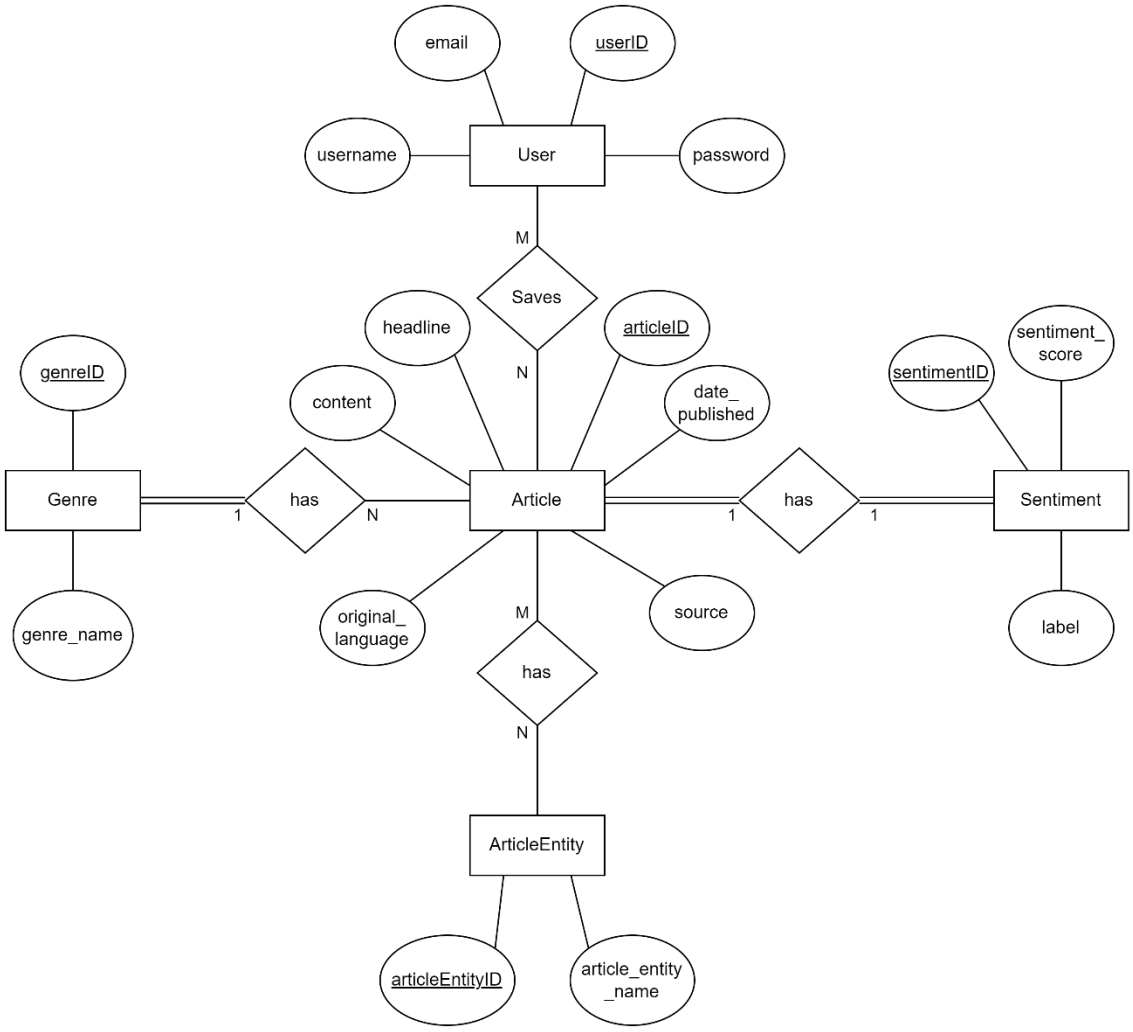

Figure 8: ER Diagram for Sentiment-centric News Portal

The above diagram illustrates the relationships between various entities: Users, Articles, Genres, Sentiments, and ArticleEntities. Users (with attributes userID, email, username, and password) can save multiple Articles. Each Article, identified by attributes like articleID, headline, content, date_published, original_language, and source, can belong to one Genre and is associated with one Sentiment. The Sentiment entity, characterized by sentimentID, sentiment_score, and label, is linked to a single Article. Articles can also be related to multiple ArticleEntities, which include attributes like articleEntityID and article_entity_name. The diagram captures these relationships with cardinalities and highlights the interconnections between the system's key components.

4.3.4. ALGORITHMS

a. SVM

SVM will help in classifying news articles based on their sentiment, especially when dealing with complex, non-linearly separable data. The decision function of SVM can be represented as:

$$f(x) = \omega^T x + b$$

Where:

- ω is the weight vector
- x is the input vector
- b is the bias term

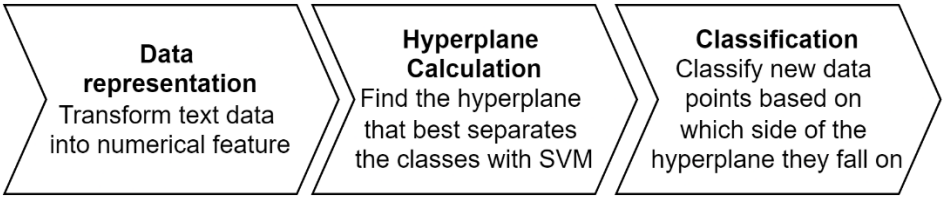

Figure 9: Working of SVM

b. CNN

CNN will be used to extract meaningful features from the text of news articles, capturing nuances in language that are important for sentiment analysis. The output of a convolutional layer can be represented as:

$$y = f(W * x + b)$$

Where:

- y is the output feature map
- W is the filter weight matrix
- x is the input feature map
- b is the bias term
- f is the activation function

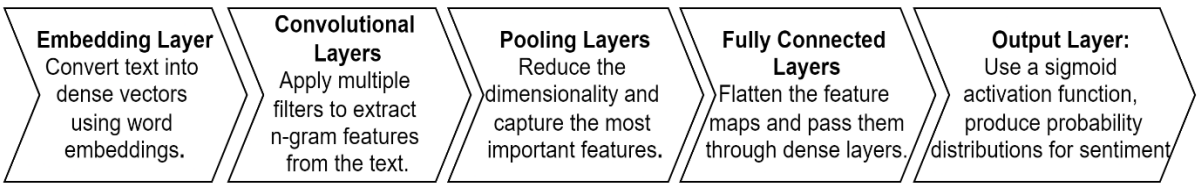

Figure 10: Working of CNN

c. CNN with SVM

This hybrid model will be used for sentiment analysis, potentially offering improved accuracy over method alone. The combination of SVM and CNN leverages the strengths of both algorithms. The combined model can be described as:

$$f(x) = \omega^T \varphi\big(CNN(x)\big) + b$$

Where:

- φ represents the feature extraction process performed by the CNN

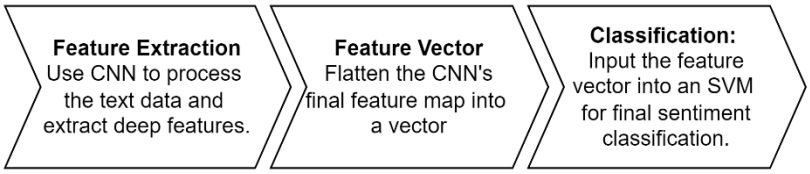

Figure 11: Working of SVM with CNN

d. LDA

LDA will be used to track and visualize how sentiment evolves for specific topics or general news trends. The joint probability distribution of LDA can be represented as:

$$P(\omega, z, \theta, \varphi, \beta) = P(\theta)\ \Pi_{k=1}^{K}\ P(\ \varphi_k)\ \Pi_{n=1}^{N}\ P(z_n|\theta)\ P(\omega_n|\varphi_{z_n})$$

Where:

- ω_n are the words in the documents.
- z are the topic assignments.
- θ is the topic distribution for a document.
- φ is the word distribution for a topic.
- K is the number of topics.
- N is the number of words in a document.

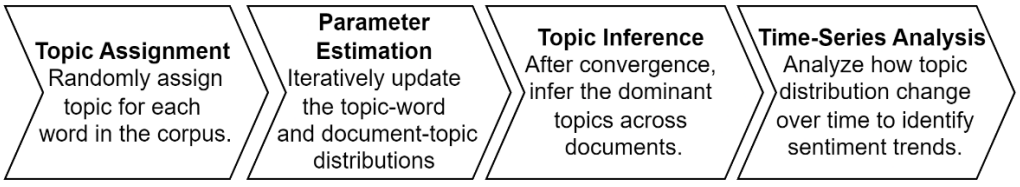

Figure 12: Working of LDA

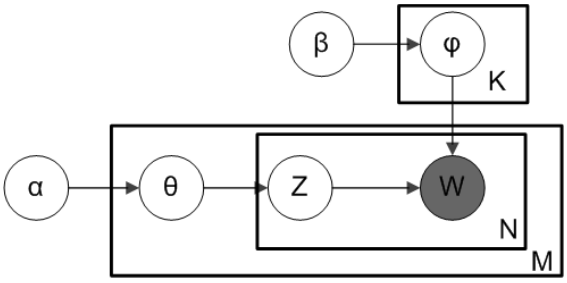

Figure 13: Blueprint of LDA Model [9]

e. Porter Stemming Algorithm

The Porter Stemming technique will be used in the preprocessing stage to standardize words before sentiment analysis. The stemming rules are a series of conditions and transformations applied to words.

For example:

- If a word ends with "sses", it is replaced by "ss".
- If a word ends with "ies", it is replaced by "i".

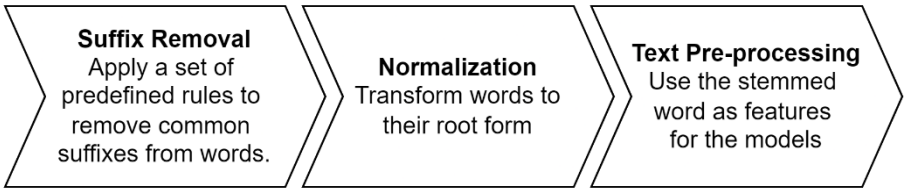

Figure 14: Working of Porter Stemming Algorithm

This algorithmic pipeline will enable the Sentiment-centric Intelligent News Portal to conduct thorough sentiment analysis, topic modeling, and trend analysis on Nepali news content, providing users with comprehensive insights into news sentiment and themes.

5. EXPECTED OUTCOME

The Sentiment-centric News Portal is set to transform how we consume and understand news by making Nepali articles accessible to a wider audience through automatic translation into English. Leveraging sentiment analysis powered by SVM and CNN, the portal will accurately gauge the emotional tone of news stories. It will use LDA to categorize news into genres like sports, politics, international affairs, and visualize sentiment trends over time. With an intuitive user interface, secure data storage, and thorough documentation, the portal will provide researchers, policymakers, and the public with valuable insights into the sentiment trends in Nepali news. This will enable more informed decision-making and foster richer public discourse. The Sentiment-centric News Portal promises to be a vital tool for navigating the modern news landscape.

REFERENCES

[1] "The Rising Nepal | Nepal's First English Broadsheet Daily.," [Online]. Available: https://risingnepaldaily.com/. [Accessed 4 August 2024].

[2] "Online Khabar," [Online]. Available: https://www.onlinekhabar.com/. [Accessed 4 August 2024].

[3] R. A. R. Julia Baum, "Emotional news affects social judgments independent of perceived media credibility," *Social Cognitive and Affective Neuroscience,* vol. 16, no. 3, 2021.

[4] L. T. A. C. Marilena Mousoulidou, "Social Media News Headlines and Their Influence on Well-Being: Emotional States, Emotion Regulation, and Resilience," *European Journal of Investigation in Health, Psychology and Education,* vol. 14, no. 6, 2024.

[5] O. P. N. H. Pavlo Radiuk, "An Ensemble Machine Learning Approach for Twitter Sentiment Analysis," in *6th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2022)*, Gliwice, Poland, 2022.

[6] K. B. M. B. B. M. H. R. K. S. &. T. I. Sheikh Shah Mohammad Motiur Rahman, "An Investigation and Evaluation of N-Gram, TF-IDF and Ensemble Methods in Sentiment Classification," in *International Conference on Cyber Security and Computer Science*, Springer, Cham, 2020.

[7] M. I. M. M. H. S. R. M. H. S. A. K. Rajesh Kumar Das, "Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models," *Heliyon,* vol. 9, no. 9, 2023.

[8] B. O. K. S. Twil Ali, "Analyzing tourism reviews using an LDA topic-based sentiment," *MethodsX,* vol. 9, 2022.

[9] L. G. Serrano, Grokking Machine Learning, 2021.