PRESIDENCY UNIVERSITY

Private University Estd. in Karnataka State by Act No. 41 of 2013

PROJECT REPORT ON

PSCS_35: 360-DEGREE FEEDBACK SOFTWARE FOR NEWS STORIES IN REGIONAL MEDIA (USING AI/ML)

Department: Computer Science Engineering

Submitted by:

- **YASHAS RAJ : 20231CSE3019**
- **SACHIN KUMAR SAROJ : 20231CSE3018**
- **RAKSHITHA REDDY KS : 20231CSE3089**

Abstract – Revolutionising Regional News Analysis

The **PSCS_35** project directly addresses a critical gap in regional media analytics by introducing an advanced, AI/ML-driven 360-degree feedback system for news stories. Regional media plays a vital role in shaping local narratives, influencing public sentiment, and driving policy discussions. However, traditional methods of sentiment analysis and thematic categorisation are often **manual, labour-intensive, and inconsistent**, making it difficult for stakeholders to quickly understand and act upon public opinion.

The proposed system leverages cutting-edge **Artificial Intelligence (AI)** and **Machine Learning (ML)** algorithms to automate the extraction and interpretation of public sentiment—categorising articles into positive, negative, or neutral tones—and classifying them into relevant departments such as **politics, economy, health, education, and social affairs**. This eliminates the subjectivity and delays inherent in manual processes while enabling scalability across multiple sources.

The platform operates through a structured pipeline:

1. **Data Collection** – Automated scraping of news articles from selected regional media outlets.
2. **Data Preprocessing** – Text cleaning, tokenisation, and preparation for analysis.
3. **Sentiment Analysis** – Lexicon-based techniques to determine emotional tone.
4. **Departmental Classification** – Rule-based and keyword-driven classification of news content.
5. **Visualisation** – Results presented on a dynamic, user-friendly dashboard, offering clear insights through charts, graphs, and trend analysis.

This comprehensive system allows **media houses** to refine content strategies, **policymakers** to respond to emerging issues, and **researchers** to track discourse trends with unprecedented speed and accuracy. By integrating automation, PSCS_35 not only improves **efficiency** and **reliability** but also fosters a **data-driven approach** to understanding the pulse of regional communities.

Ultimately, the PSCS_35 project stands as a transformative step toward **real-time, bias-free, and actionable media intelligence**, with the potential to evolve into a multilingual, deep-learning-powered platform capable of handling diverse news ecosystems at scale.

INDEX :

Problem Statement

Regional media plays a pivotal role in shaping community narratives, amplifying local voices, and influencing public perception. Its impact is particularly significant in regions where local issues often outweigh national ones. However, the process of analysing sentiment and categorising news stories in this domain is fraught with operational and methodological challenges, many of which limit both the speed and accuracy of insights derived.

1. **Lack of Scalability** – The sheer volume of news generated across multiple regional outlets each day makes manual review impractical. Human analysts are unable to process this data at the required scale, resulting in incomplete coverage and missed stories of importance.
2. **Inconsistency and Bias** – Manual sentiment analysis relies heavily on individual interpretation, which can introduce significant variability in results. Factors such as personal bias, mood, or limited subject-matter understanding can distort sentiment assessments and classification accuracy.
3. **Delayed Feedback** – Traditional analysis methods are slow, often producing insights long after the news has lost its relevance. This delay reduces the usefulness of findings for media houses, policymakers, and other stakeholders who require timely information to act effectively.
4. **Limited Depth of Analysis** – Existing manual approaches frequently produce only surface-level insights. They struggle to capture subtle shifts in public mood, detect sarcasm or irony, or identify nuanced departmental relevance—factors critical for deeper understanding and precise policy or editorial response.

These limitations highlight the urgent need for an automated, intelligent, and scalable solution capable of delivering **real-time, consistent, and context-aware sentiment analysis** tailored for regional news ecosystems.

Objectives – Project Goals

The **PSCS_35** project is designed with clearly defined objectives that directly address the challenges faced in regional media sentiment analysis. Each goal focuses on building an intelligent, efficient, and user-friendly system capable of delivering actionable insights at scale.

1. **Automate Data Collection**
   Develop a robust, automated news scraper capable of continuously gathering articles from multiple regional online media sources. This will ensure comprehensive coverage, reduce manual workload, and enable a steady inflow of fresh data for analysis.
2. **Accurate Sentiment Analysis**
   Integrate advanced AI/ML models to determine the sentiment polarity—positive, negative, or neutral—of each news article. The system will aim for high accuracy by minimising human bias and ensuring consistent interpretation across diverse topics.
3. **Intelligent Content Classification**
   Implement automated classification of news articles into predefined departmental categories such as health, education, infrastructure, politics, and economy. This will allow stakeholders to focus on specific areas of interest and retrieve targeted information quickly.
4. **Interactive and Intuitive Dashboard**
   Design and deploy a visually appealing, user-friendly dashboard that presents sentiment trends, departmental distributions, and key performance metrics in real time. The dashboard will offer interactive charts, graphs, and filtering options to facilitate quick and informed decision-making.

By achieving these objectives, **PSCS_35** will establish itself as a powerful tool for enhancing media strategy, monitoring public discourse, and supporting data-driven policy formulation.

Scope – Defining the System's Boundaries

The scope of the **PSCS_35** project defines the capabilities and limitations of the system in its current phase. This ensures clarity for stakeholders and sets realistic expectations for deliverables. The system is focused on providing accurate, automated sentiment analysis and departmental classification of regional news content, while leaving certain advanced features for future development.

In-Scope Features

- **Automated News Scraping** – Continuous extraction of articles from selected regional online media platforms, ensuring timely and diverse data collection.
- **Text Preprocessing** – Cleaning and preparing English-language content through tokenisation, stop-word removal, and standardisation to improve analysis accuracy.
- **Sentiment Analysis** – Implementation of lexicon-based techniques to categorise news articles as positive, negative, or neutral.
- **Keyword-Based Department Classification** – Categorising articles into predefined sectors such as politics, economy, health, education, and infrastructure.
- **Interactive Web Dashboard** – A browser-accessible dashboard offering visualisations of sentiment distribution, departmental breakdown, and trend analytics.
- **Historical Data Storage** – Maintaining processed data for long-term trend monitoring and comparative analysis.

Out-of-Scope Features (Initial Phase)

- Multi-language support for regional dialects and languages beyond English.
- Deep learning-based sentiment analysis models for higher contextual accuracy.
- Highly granular sub-category classification within main departments.
- Real-time news feed integration and streaming analytics.
- Analysis of user-generated content such as social media comments or forum posts.
- User authentication and role-based access controls.

Methodology – The Project Workflow

1. Data Collection

- Automated web scrapers periodically extract articles from a curated list of regional news websites.
- The sources are selected to ensure diversity, credibility, and representation of different sectors and viewpoints.
- Extracted data includes headlines, article text, publication date, source name, and category (if available).

2. Preprocessing

- **Cleaning** – Removal of HTML tags, advertisements, and non-textual elements to ensure noise-free input.
- **Tokenisation** – Breaking down sentences into words or tokens for better analysis.
- **Stop-word Removal** – Eliminating common, non-informative words (e.g., "is," "and," "the") to focus on meaningful content.
- **Lowercasing** – Standardising all text to lowercase for consistent processing.
- **Data Structuring** – Organising cleaned text into a standardised format suitable for sentiment and classification models.

3. Sentiment Analysis

- Applying **lexicon-based NLP techniques** to categorise news articles as **Positive**, **Negative**, or **Neutral**.
- Sentiment scoring considers the frequency and weight of positive/negative terms, with adjustments for negations and intensifiers.
- Output is stored alongside article metadata for visualisation and trend analysis.

4. Department Classification

- Using **keyword-based rule matching** to classify each article into predefined departments such as **Politics**, **Economy**, **Health**, **Education**, or **Infrastructure**.
- Classification rules are based on curated keyword dictionaries relevant to each department.
- Articles that match multiple departments are tagged with all relevant categories to maintain contextual relevance.

5. Dashboard Visualisation

- An **interactive, web-based dashboard** built with **Flask (backend)** and **Chart.js (frontend)**.
- Displays sentiment distribution (pie/bar charts), departmental breakdown, and sentiment trends over time.
- Enables filtering by date range, department, and sentiment category.
- Allows stakeholders to quickly interpret public opinion and assess news impact.

Tools & Technologies

The **PSCS_35** system is built using a carefully selected set of tools, frameworks, and programming languages to ensure **efficiency**, **scalability**, and **ease of maintenance**. Each technology was chosen for its suitability in handling the project's specific requirements, ranging from **data extraction** and **text processing** to **machine learning** and **interactive visualisation**.

1. Programming Language

- **Python** – The backbone of the project due to its simplicity, versatility, and rich ecosystem of AI/ML and data processing libraries.
- Supports quick prototyping and seamless integration of web frameworks, NLP models, and data manipulation tools.

2. Web Framework

- **Flask** – A lightweight yet powerful Python web framework used for:
  - Building backend APIs for sentiment and classification results.
  - Serving the interactive dashboard to end users.
  - Ensuring rapid development and flexible deployment.

3. Natural Language Processing (NLP) Libraries

- **TextBlob** – For quick sentiment analysis using lexicon-based methods.
- Provides built-in functions for tokenisation, part-of-speech tagging, and polarity/subjectivity scoring.

4. Data Handling & Analysis

- **Pandas** – For efficient data cleaning, manipulation, and transformation.
- Allows handling large datasets of news articles while maintaining high performance and readability.

5. Front-End Technologies

- **HTML5 & CSS3** – For structuring and styling the dashboard interface.
- Ensures a clean, intuitive, and responsive user experience.
- **JavaScript** – For adding interactivity to the dashboard and enabling dynamic data updates.

6. Data Visualisation

- **Chart.js** – A JavaScript charting library used to create interactive, aesthetically appealing graphs and charts.
- Visualises sentiment trends, departmental breakdowns, and other key performance indicators (KPIs).

7. Data Storage

- **SQLite / MySQL** (configurable) – For structured storage of processed news data and historical records.
- Enables retrieval of past sentiment trends for longitudinal analysis.

8. Supporting Tools

- **BeautifulSoup & Requests** – For web scraping of regional news websites.
- **Git** – For version control and collaborative development.
- **VS Code / PyCharm** – For efficient code development and debugging.

By leveraging this technology stack, **PSCS_35** achieves a balance of **performance**, **flexibility**, and **ease of deployment**, ensuring the system can adapt to evolving project needs and scale to handle increasing volumes of news data.

Advantages – Key Benefits of PSCS_35

The **PSCS_35** system offers a range of advantages that directly address the shortcomings of traditional manual sentiment analysis and news classification. By leveraging AI/ML automation, the system transforms the way regional news content is processed, evaluated, and utilised.

1. Automation of Analysis

- Eliminates the need for time-consuming manual review by automatically processing large volumes of news articles.
- Reduces human intervention, freeing analysts to focus on higher-level strategic decisions.

2. High Processing Speed

- Capable of analysing thousands of articles within minutes.
- Provides near real-time insights, enabling media houses and policymakers to respond faster to emerging trends.

3. Scalability

- Easily expandable to include new data sources, more departments, and additional sentiment categories.
- Can scale horizontally to handle growing volumes of regional news without performance loss.

4. Reduced Human Bias

- AI-driven sentiment analysis minimises subjectivity, ensuring more consistent and objective evaluations of news tone and content.

5. Cost-Effectiveness

- Reduces the need for large manual analysis teams, cutting operational costs.
- Offers a high return on investment (ROI) by delivering valuable, actionable insights at lower expense.

6. Data-Driven Decision Making

- Empowers media organisations, researchers, and policymakers with factual, quantifiable data to guide editorial choices, policy strategies, and public engagement efforts.

Limitations

While **PSCS_35** introduces significant improvements in regional news analysis, it is important to recognise its current limitations. Understanding these constraints helps in planning realistic expectations and prioritising future upgrades.

1. Limited Language Support

- Currently designed to process and analyse only **English-language** news articles.
- Lacks the ability to handle multiple regional languages, which restricts its reach in linguistically diverse regions.

2. Dependency on Data Quality

- The accuracy of sentiment analysis and classification heavily depends on the **quality, structure, and formatting** of the scraped news data.
- Poorly written or incomplete articles can reduce the reliability of the results.

3. Contextual Understanding Challenges

- The current lexicon-based sentiment model may **struggle with sarcasm, irony, idiomatic expressions**, or highly context-dependent statements.
- This can lead to incorrect sentiment classification in certain nuanced cases.

4. Classification Accuracy Limitations

- Uses **keyword-based** departmental classification in the initial version, which may overlook subtle thematic relationships or misclassify ambiguous content.

5. Data Source Restrictions

- Works only with publicly available news content.
- Cannot directly process articles behind **paywalls** or from sources with restricted access, limiting data coverage.

6. Lack of Real-Time Processing (Initial Phase)

- The current system does not offer **instantaneous** analysis of breaking news.
- Insights are generated based on batch processing rather than live streaming data.

7. Potential Overfitting in Future ML Models

- As more advanced AI models are introduced, there is a risk of **overfitting** if not properly validated against diverse and unseen datasets.

Future Enhancements

1. Multi-Language Support

Currently, the system operates only in English, which limits its ability to analyse the rich diversity of regional news in India. A key enhancement is the integration of Natural Language Processing (NLP) models capable of handling multiple Indian languages, such as Hindi, Kannada, Tamil, Telugu, Marathi, and Bengali. This will require:

- Implementation of multilingual transformer-based models such as mBERT or XLM-R.
- Language detection algorithms to automatically identify the language of an article before processing.
- Regional lexicons and sentiment dictionaries tailored to cultural and linguistic nuances.
  Impact: This will vastly expand coverage, making PSCS_35 more inclusive and valuable for diverse media houses.

2. Advanced AI-Based Classification

The current keyword-based classification approach, while functional, can misclassify articles that have implicit or indirect departmental relevance. Future upgrades will focus on:

- Training deep learning models like BERT, RoBERTa, or DistilBERT for context-aware classification.
- Using hierarchical classification models to allow for main categories (e.g., Politics) and subcategories (e.g., Election Campaigns).
- Continuous model retraining using newly scraped datasets to adapt to evolving news trends.
  Impact: Greater accuracy, improved contextual understanding, and the ability to adapt to new and emerging categories.

3. Real-Time News Integration

Currently, data processing is batch-based, which can cause delays in insight delivery. The upgraded system will integrate:

- Live news APIs from major aggregators.
- Real-time data pipelines using Apache Kafka or Spark Streaming.
- On-the-fly sentiment computation for breaking news events.
  Impact: Media houses and policymakers will have access to immediate feedback, enabling rapid decision-making in critical situations.

4. Enhanced Sentiment Analysis with Deep Learning

Lexicon-based sentiment analysis may fail in detecting sarcasm, irony, or complex emotional tones. Future enhancements will involve:

- Transformer-based sentiment models like BERT, ALBERT, or ELECTRA.

- Emotion classification (e.g., anger, joy, fear, trust) beyond simple polarity (positive/negative/neutral).
- Contextual embeddings to capture meaning at sentence and paragraph levels.
  Impact: More accurate sentiment detection, especially for nuanced and context-heavy news narratives.

5. User Authentication & Personalised Dashboards

To make the system more secure and user-oriented, future releases will include:

- Role-based access control for different types of users (journalists, analysts, policymakers).
- Personalisation features such as configurable sentiment thresholds, departmental filters, and report scheduling.
- Cloud-based multi-user dashboards for collaborative analysis.
  Impact: Greater usability, security, and adaptability to specific user needs.

6. Social Media and Reader Comment Analysis

Public opinion is not limited to news articles; it thrives on interactive platforms. Future iterations will:

- Scrape reader comments from news websites.
- Analyse sentiment trends across social media platforms like Twitter (X), Facebook, and Instagram.
- Cross-reference public reaction data with the news article sentiment for a more complete picture.
  Impact: Enables a true “360-degree” view of public discourse by combining editorial content with audience responses.

7. Predictive Analytics and Trend Forecasting

To make the system more proactive, planned enhancements include:

- Machine learning models to predict emerging news topics and sentiment trends.
- Seasonal and event-based forecasting to help media houses prepare content in advance.
- Automated alerts for sudden sentiment shifts in specific topics or regions.
  Impact: Empowers media organisations to anticipate public mood and act before stories escalate.

8. Integration with Policy and Decision-Making Tools

Beyond media houses, policymakers can benefit from sentiment insights. Future versions may:

- Integrate with government or NGO decision-support systems.
- Provide customised policy impact analysis reports.

Conclusion – Summary of Outcomes

The **PSCS_35** project successfully addresses the growing need for an automated, efficient, and unbiased approach to sentiment analysis and classification of regional news. By integrating **Artificial Intelligence (AI)**, **Machine Learning (ML)**, and **Natural Language Processing (NLP)** techniques, the system streamlines the process of collecting, processing, and interpreting large volumes of news data from regional sources.

Through the implementation of automated news scraping, lexicon-based sentiment detection, and keyword-driven departmental classification, the project delivers a reliable framework that reduces manual effort, minimises subjectivity, and enhances the speed of analysis. The inclusion of an interactive dashboard allows stakeholders such as media houses, policymakers, and researchers to visualise sentiment trends, departmental distributions, and public opinion patterns in a clear and actionable format.

While the current version of PSCS_35 demonstrates strong potential, it also reveals opportunities for significant improvements. Limitations such as single-language processing, reliance on lexicon-based models, and batch-mode analysis provide a clear roadmap for future enhancements, including multi-language support, deep learning-based classification, real-time analytics, and predictive sentiment forecasting.

Overall, **PSCS_35** represents a critical step toward transforming how regional media sentiment is analysed and interpreted. By combining automation, data analytics, and AI-driven insights, the system lays the groundwork for a more informed, responsive, and inclusive media landscape—one that empowers decision-makers with timely and accurate sentiment intelligence.