Project Goals and Plan

May 22, 2024

Contents

1 Problem Statement

Sentiment Analysis of Western Media Coverage on India

In today's globalized world, media coverage plays a crucial role in shaping public perception and opinion. The portrayal of countries and their events in foreign media can significantly influence international relations, tourism, investment, and public sentiment. India, as one of the world's largest democracies and rapidly growing economies, frequently features in Western media. However, the sentiment of this coverage—whether positive, neutral, or negative—can vary widely.

Despite the importance of understanding these portrayals, there is a lack of comprehensive analysis focused on the sentiment of Western media coverage about India. This project aims to fill this gap by conducting a detailed sentiment analysis on a large dataset of news articles from prominent Western media outlets. By leveraging natural language processing (NLP) techniques and machine learning algorithms, the project will identify trends, patterns, and biases in the sentiment of this coverage.

The main objectives of the project are:

- To collect and preprocess a substantial dataset of news articles about India from Western media sources.
- To develop a robust sentiment analysis model capable of accurately classifying the sentiment of these articles.
- To analyze the sentiment data to uncover trends over time, differences between media outlets, and potential biases in coverage.
- To provide visualizations and reports summarizing the findings, which can be valuable for policymakers, researchers, and the general public.
- The insights gained from this project will contribute to a better understanding of how India is portrayed in Western media and can help in addressing any identified biases or misinformation. It will also provide a foundation for further research in the field of media studies and sentiment analysis.

2 System Model / Architecture

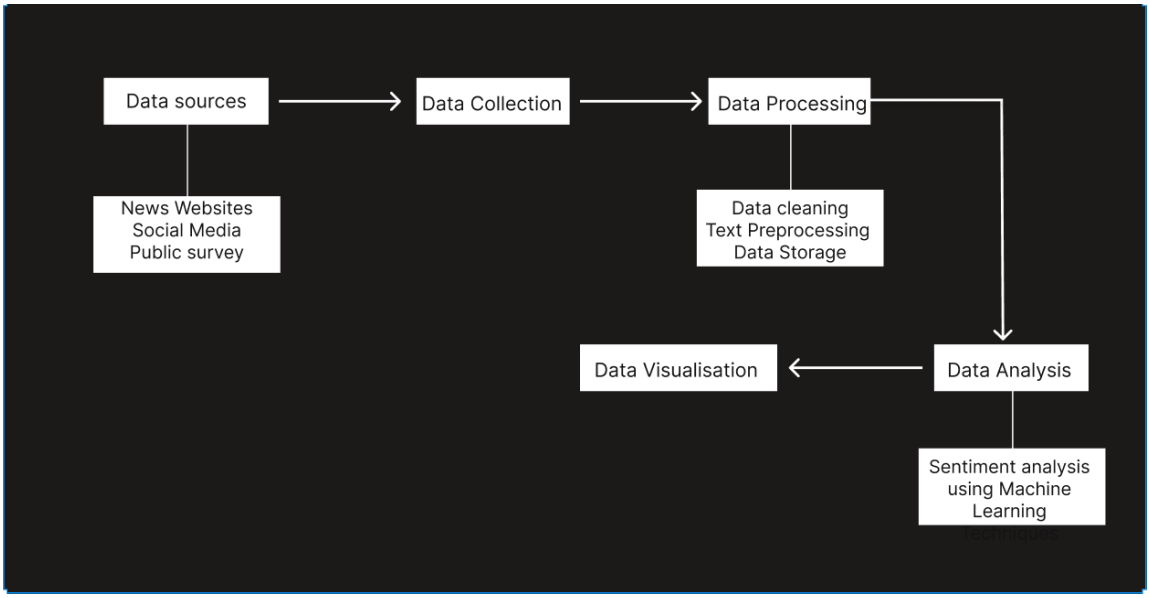

Figure 1: System Model

3 Description of Tools

1. Data Collection Tools

a. Web Scrapers:

- *BeautifulSoup*: A Python library used for parsing HTML and XML documents. It creates parse trees from page source code that can be used to extract data easily.
- *Scrapy*: An open-source and collaborative web crawling framework for Python. It's used to extract data from websites and can handle large-scale scraping operations.

b. API Integrations:

- *Tweepy*: A Python library for accessing the Twitter API. It allows for easy retrieval of tweets, user data, and other information from Twitter using their API.
- *Facebook Graph API*: Allows access to data on Facebook. It enables you to retrieve posts, comments, and user interactions.
- *PRAW (Python Reddit API Wrapper)*: A Python package that allows for easy access to Reddit's API to extract posts, comments, and other data.

2. Data Processing Tools

a. Data Cleaning:

- *Pandas*: A powerful Python data analysis toolkit that provides data structures and functions needed to clean and manipulate data.

- *OpenRefine*: A powerful tool for working with messy data. It allows for cleaning, transforming, and extending data with web services.

b. Text Preprocessing:

- *NLTK (Natural Language Toolkit)*: A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in Python. It includes libraries for tokenization, stemming, and more.
- *SpaCy*: An open-source software library for advanced NLP in Python. It features fast and accurate tokenization, part-of-speech tagging, dependency parsing, and named entity recognition.

3. Analysis Tools

a. Sentiment Analysis:

- *VADER (Valence Aware Dictionary and sEntiment Reasoner)*: A lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media.
- *TextBlob*: A Python library for processing textual data. It provides a simple API for diving into common NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and translation.
- *BERT (Bidirectional Encoder Representations from Transformers)*: A transformer-based machine learning technique for NLP pre-training. It's used for tasks like sentiment analysis and can be fine-tuned with specific data.

d. Comparative Analysis:

- *Pandas & NumPy*: For data manipulation and numerical operations in Python.
- *SciPy*: A Python library used for scientific and technical computing. It contains modules for statistics, optimization, integration, and more.

4. Visualization & Reporting Tools

a. Dashboards:

- *Tableau*: A powerful data visualization tool used for converting raw data into an understandable format. It is used to create interactive and shareable dashboards.
- *Power BI*: A business analytics service by Microsoft that provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards.

b. Reports:

- *Jupyter Notebooks*: An open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. It's used for data cleaning, transformation, analysis, and visualization.
- *LaTeX*: A typesetting system commonly used for technical and scientific documentation.

c. Custom Visuals:

- *Matplotlib*: A plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications.
- *Seaborn*: A Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

4 Dataset

The dataset for this project will consist of multiple types of data sources, each contributing unique insights into how India is covered and perceived in the Western media, specifically focusing on the UK and USA. Below is a detailed description of the dataset components:

4.1 News Articles

Sources:

- **UK Media**: BBC, The Guardian, The Times, Financial Times, The Telegraph.
- **USA Media**: CNN, The New York Times, Washington Post, The Wall Street Journal, NBC News.

Data Points:

- **Article Title**: The headline of the news article.
- **Publication Date**: Date when the article was published.
- **Author**: The author(s) of the article.
- **Source**: The name of the news outlet.
- **Full Text**: The complete text of the article.
- **URL**: The link to the original article.
- **Metadata**: Tags, categories, and section (e.g., World News, Politics, Business).

Collection Method:

- **Web Scraping**: Use Scrapy to scrape articles from the websites of the specified news sources.
- **APIs**: Where available, use news APIs to collect articles.

4.2 Social Media Posts

Sources:

- **Twitter**: Tweets mentioning India.
- **Facebook**: Public posts and comments mentioning India.
- **Reddit**: Posts and comments from relevant subreddits (e.g., r/worldnews, r/India, r/geopolitics).

Data Points:

- **Post Content**: The text content of the post or comment.
- **Username**: The username of the person who posted.
- **Timestamp**: Date and time when the post was made.
- **Platform**: Twitter, Facebook, or Reddit.
- **Engagement Metrics**: Likes, shares, retweets, comments.
- **Location**: Geolocation data if available (to distinguish between UK and USA posts).
- **URL**: The link to the original post.

Collection Method:

- **APIs**: Use Tweepy for Twitter, Facebook Graph API for Facebook, and PRAW for Reddit.

5 Analysis Approach and Expected Outcome

5.1 Type of Analysis

The analysis of Western news coverage about India will primarily focus on sentiment analysis, aiming to gauge the overall sentiment expressed in news articles and social media posts.Assessing the sentiment polarity (positive, negative, or neutral) of the text data. The analysis will involve:

1. **Comparative Analysis**: Comparing sentiment across different media sources, platforms, and time periods.

2. **Trend Analysis**: Identifying trends and patterns in sentiment over time or in response to specific events or topics.

3. **Geographical Analysis**: Analyzing sentiment variations between the UK and USA media.

5.2 Comparative Analysis

- **Statistical Tests**: Conduct statistical tests (e.g., t-tests) to compare sentiment distributions between different media sources and platforms.

- **Visualization**: Use visualizations such as box plots, histograms, or bar charts to illustrate sentiment differences.

5.3 Trend Analysis

- **Time-Series Analysis**: Analyze sentiment trends over time using time-series analysis techniques.

- **Topic Modeling**: Apply topic modeling techniques to identify dominant topics and their associated sentiment over time.

5.4 Geographical Analysis

- **Geospatial Visualization**: Map sentiment scores geospatially to visualize regional variations in sentiment.

- **Statistical Analysis**: Perform statistical tests to compare sentiment between regions (UK vs. USA).