Acta Scientific COMPUTER SCIENCES

Volume 6 Issue 8 August 2024

Research Article

A Novel Method for Social Media Trend Analysis Using Categorization of Posts

KLVR Saraswathi[1]*, K Bhanu Charan[1], J Yogendra[1] and K Kranthi Kumar[2]

[1]*Students of Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

[2]*Faculty of Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

***Corresponding Author:** KLVR Saraswathi, Students of Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.

Received: May 06, 2024

Published: July 17, 2024

Abstract

This project aims the challenge of navigating social media by categorizing trending topics like cinema, crime, education, business, and sports. Unlike conventional methods, we use machine learning and natural language processing to explore deeper into insights. By precisely categorizing topics and visualizing them based on their content and context, users can effortlessly discover relevant information. Our approach offers a clear pathway to the discussions that capture users' interest the most. Conventional systems often rely solely on sentiment analysis, overlooking context and precision. Our innovative approach presents a more sophisticated solution for social media analysis. Users can swiftly access the latest movie buzz, educational trends, or business insights with assurance. Through advanced algorithms and visualization techniques, we ensure the precise categorization of trending topics. Our project empowers users to make informed decisions and uncover valuable insights from social media.

Keywords: Social Media; Trending Topics; Categorization; Visualisation; Machine Learning; Natural Language Processing; Clustering

Introduction

In today's digital communication world, social media in [1], platforms have become an important communication, information sharing and leadership tool and are of great value for people to benefit from and understand information about their environment. So millions of users will create tons of content every day, keeping up with the content and understanding different topics and trends can be challenging for end users using social media. Covering everything from blockbuster releases to global business trends, these topics shape conversation and influence understanding of the world, making social media a tool designed for vast amounts of data.

However, effectively categorizing and analyzing trending topics in [2], from the social medias can be a challenge. Existing methods often rely on simplistic approaches such as sentiment analysis, which may fail to capture the category and meaning of the context and themes inherent in social media conversations and posts, this can only tell the nature of the post rather than the category of the post. As a result, users are left grappling with fragmented insights and disjointed narratives and can't able to extract meaningful insights and make informed decisions. Recognizing the need for a more robust and comprehensive approach to analyse and know the information by categorizing the posts generated from social media, our project sets out to bridge this gap by proposing a novel framework for categorizing trending topics into distinct domains.

In this paper, we present the conclusion of our efforts a Novel method for Trend analysis categorization of trending topics from multiple social media. By using advanced machine learning techniques and natural language processing algorithms, we aim to unravel the complexities of online trends and offer users a clearer understanding of the topics dominating social media conversations. Through intensive analysis across domains such as cinema, crime, education, business, and sports, we aim to provide stakeholders with actionable insights like visualizations to guide decision-making processes and promote deeper interactions with trending topics and can know what is trending on social media in each category. We also developed a user interface using Web Technologies Stack i.e. HTML, Bootstrap, CSS, JavaScript and flask to make our project more user friendly in real time.

Objective

The main objective of our project is to categorize trending topics on social media platforms into five distinct categories: cinema, crime, education, business, and sports by using unsupervised learning techniques, preprocessing of text, visualizations and extracting

meanings from the sentences. By achieving this objective, we aim to gain insights into the distribution of trending topics across different categories and social media platforms.

Scope of study

Social media users face significant difficulties when trying to find popular subjects related to their interests. Because social media is an open platform anybody can upload stuff, from major news stations to regular people, and users have to filter through a huge amount of material to find significant trends. Users may find this approach time-consuming as they search through different categories and themes. Additionally, it takes a lot of time and work to stay up to date on worldwide trends in a variety of sectors. Thus, resolving this issue effectively will let people analyse material according to their interests.

The main aim is to create a user interface System that helps user to observe and gain knowledge upon each category posts of their interest, So that users can easily stay updated on the latest trending topics in each category and can also gain insights like what topic is trending currently by observing the visualizations. This system also provide a feature to the user to predict what category that it belongs to. This feature enhances user productivity by saving time and ensuring accurate consumption of content, thus providing a more efficient and informed browsing experience.

Related works

In the new era of technology, social media platforms have become a very important tools for communication, sharing information, and gaining social connections. With billions of people are active daily on platforms like Facebook, Twitter, Instagram, and LinkedIn, these virtual place have evolved into very big ecosystems where trends emerge, ideas flow and information spreads. The vast quantity and different nature of content produced on social media lead to arise opportunities as well as obstacles for researchers, marketers, and analysts to derive valuable information from this large set of data flow.

In [3], researchers employed a post classifier methodology utilizing the Bag of Words technique to categorize posts. This process involved preprocessing the text data i.e. by removing stop words and unnecessary emojis and punctuations and employing a Naive Bayes classifier for the categorization task. Building upon this groundwork, the current study introduces a novel model specifically tailored to predict the category of Arabic text. This model was developed using a vast and diverse dataset, allowing for more accurate predictions and enhancing the understanding of categorization patterns in Arabic text.

A framework for categorizing social media post conducted qualitative content analysis involving deductive coding, inductive coding, and validation coding procedures to categorize brand posts This categorization provides guidance for marketers to stimulate daily customer interaction on social media and offers a solid conceptual foundation for researchers to categorize, code, and model brand posts.

By this above works, we go forward on a comprehensive exploration of existing research in the field of trend analysis and classification within social media platforms [4]. Our objective is to categorize trending topics on social media platforms into five distinct categories: cinema, crime, education, business, and sports by using unsupervised learning techniques and preprocessing of text. By achieving this objective, we aim to gain insights into the distribution of trending topics across different categories and social media platform.

Methodology

Methodology includes step by step implementation of our proposed system with detailed explanation for categorizing the social media posts into different categories such as crime, sports, cinema, education, business using advanced machine learning algorithms and natural language processing techniques. We provide a structural design and arrangement of stages in detail in our system architecture.

Figure 1 describes our system architecture, which defines the blueprint or overall structure of our model , illustrating how different components like data collection, importing, preprocessing, feature extraction, dimensionality reduction, clustering, and prediction algorithms are interconnected.

The following is the step by step detailed explanation of our proposed system architecture and the methodology we followed for categorizing the posts.

Data collection

We collected 1000 real time post texts from different social medias such as Instagram, Reddit, X. We also used an online platform named “Social Searcher” for collecting treding post texts.

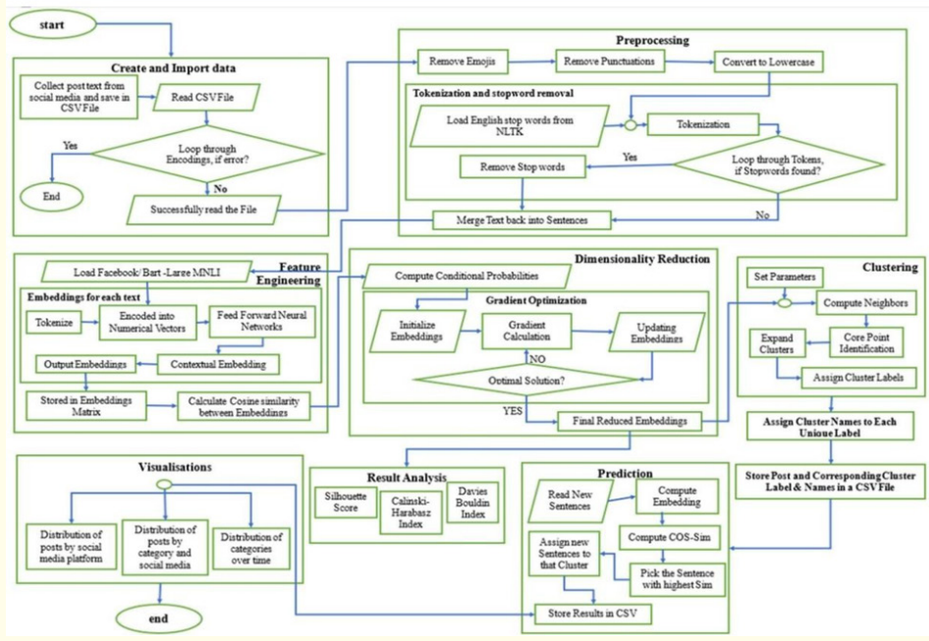

Figure 1: Proposed System Architecture.

We stored the collected data in a csv file for further use. The collected real time social media posts cover different categories. Large data set usage can improve the learning of different patterns in the user driven post texts while extracting features from those texts. As we collected data from social media platforms and stored in csv file, the data is semi structured and may contain noise. To use the data in further stages we first read the data using different encodings, so that we won't get any error while reading the file and store the post texts into a data frame using pandas with the column named posts.

Preprocessing

In this stage we used natural language processing to preprocess the post texts. The texts underwent preprocessing to clean and standardize the text for further analysis. This stage has different sub-stages which include removing emojis, punctuation marks, and special characters, as they don't play much significance while categorizing the text into categories and ensuring only text is remained for analysis. We converted all the texts to lower cases so that it mitigates issues related to case sensitivity and to ensure uniformity during analysis. We also tokenized sentences and removed stopwords such as "the", "is", "and" etc., to reduce noise. Then we merged tokens back into the sentences.

After preprocessing we get the lower cased post texts with no emojis, punctuations, special characters, and stopwords, which are stored in a processed_post column and used for further analysis.

Feature extraction

It is very important stage in our methodology because in feature extraction step we extract the meaningful information from the processed texts. We began by converting the processed posts into dense vector representations (numerical representation) in an high dimensional space. We call those dense vector representations as sentence embeddings, they contain the semantic meaning and context of the processed posts.

In this step we used one of the sentence transformer model named "facebook/bart-large-mnli" for converting the processed posts into sentence embeddings.

Then used cosine similarity calculation between the sentence embeddings to get the degree of relatedness among different posts within the semantic space and stored the values in cosine similarity matrix. If the cosine similarity is high the posts are more similar or if the cosine similarity is low then the posts are less similar to each other. Calculating the cosine similarity helps in clustering similar posts together and identifying trends within the data set.

The two key aspects in feature engineering – generating sentence embeddings and calculating cosine similarity captured the essence of the processed posts.

Dimensionality reduction

As the features extracted are in high dimensions we used t-SNE for dimensionality reduction for better visualisation of relationships and patterns of the cosine similarity matrix. Reducing the dimensions is essential for understanding patterns of the social media posts.

We used (t-SNE)t-distributed Stochastic Neighbour Embedding because it is particularly effective at preserving local structures and patterns while converting into lower-dimensional space. It aims to minimize the Kullback-Leibler(KL) divergence which ensures similar distribution in the low-dimensional space as the high-dimensional space.

Using t-SNE we reduced the dimesions of our embeddings from high to two-dimensional space by preserving the features of processed posts. Now, these reduced embeddings are given as input to the clustering, t-SNE improves the clustering performance, leading to more accurate categorization of posts.

Clustering

The extracted features i.e., reduced embeddings were then fed into a clustering algorithm to group similar posts into clusters. We choosed Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm due to its capability to identify clusters of arbitrary shapes, handle noise effectively and suits well for the large set of data with varying densities and complex structures. This makes the DBSCAN suiTable for categorize the social media posts with different content and context.

The most important parameters while using DBSCAN are eps and min_samples. Their values are user defined and play a significant role in clustering the posts. Eps(epsilon) determines the maximum distance between two data points and min_samples are the number of data points with in the specified eps(radius). We set the eps and min_samples values as 8 and 33 respectively.

After the DBSCAN clustering is completed we labelled each cluster with suiTable labels such as sports, crime, education, cinema, business. And the cluster -1 represents the noise data points.

Prediction

Following clustering, we developed a prediction mechanism to assign appropriate cluster labels to the real-time trending posts. First we computed embeddings of new posts and then calculated cosine similarity between existing reduced embeddings and the new sentence embeddings and focused on the existing processed post which have highest cosine similarity with the new sentences and then assigned the processed post label to the new sentences. In this way we predicted the labels for the real-time trending posts.

The real time trending posts are the data set of 200 trending posts, which we collected from different social media focusing more on the trending topics of each day of the month march 2024.

To demonstrate the use of our proposed system we also developed a web application based on the prediction results of real time data. In our web application we displayed the category cards enabling users to visualise the trends category wise and provided the top trends of each category and also users can predict their own texts by just clicking a button at the bottom of the home page. User interface of all the pages from our web application are shown in Figure 7.

Evaluations

To evaluate the quality and effectiveness of our DBSCAN clustering results, we used metrics such as Silhouette Score, Calinski-Harabaz Index, Davis-Bouldin Index and their scores are visualized using a bar chart in the Figure 3. These metrics provide insights into the separation, cohesion, and compactness of the clusters, enabling us to quantify and qualify the performance of our clustering algorithm and we compared the output values of those metrics with ideal clustering outcomes in the Table 1. We have also used visualisations like scatter plot to visualise the clusters and how they are arranged in a two dimensional space as shown in Figure 2. By integrating data collection, preprocessing, feature extraction, dimentioanlity reduction, clustering, prediction, and evaluation stages, our methodology offers a novel and comprehensive framework for analyzing social media trending posts and extracting meaningful insights from those trending posts.

Results

Cluster Label	Cluster Name	Count
0	Business	200
1	Education	202
2	Cinema	197
3	Crime	200
4	Sports	201

Table 1: Clustering Performance Assessment Table.

Discussions

The results obtained from our proposed system are very accurate. As we used advanced techniquessuch as facebook/bart-large-mnli for senetence embeddings and DBSCAN for clustering, the combination of both achiebed high-quality categorization results.

Our clustering algorithm upon evalution with metrics gave favourable outcomes, including Silhouette Score of 0.73, indicating

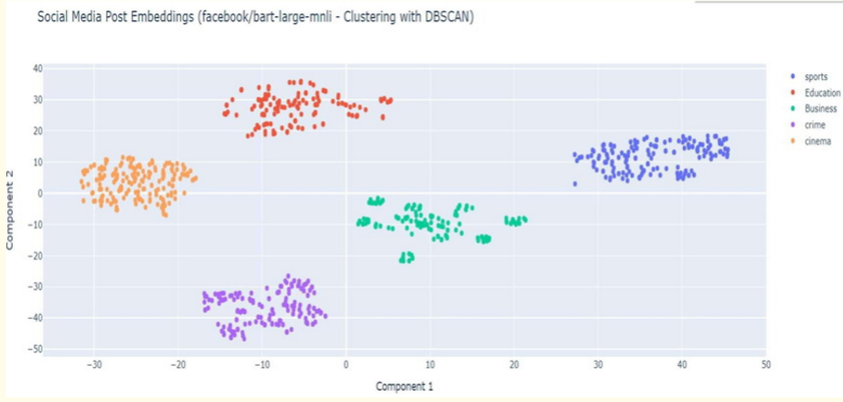

Figure 2: DBSCAN Cluster Results.

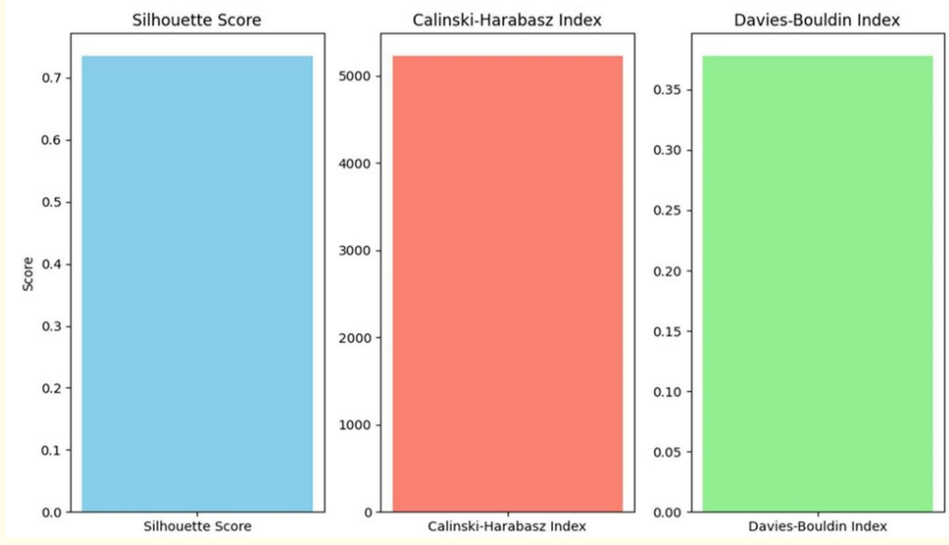

Figure 3: Clustering evaluation scores of different metrics.

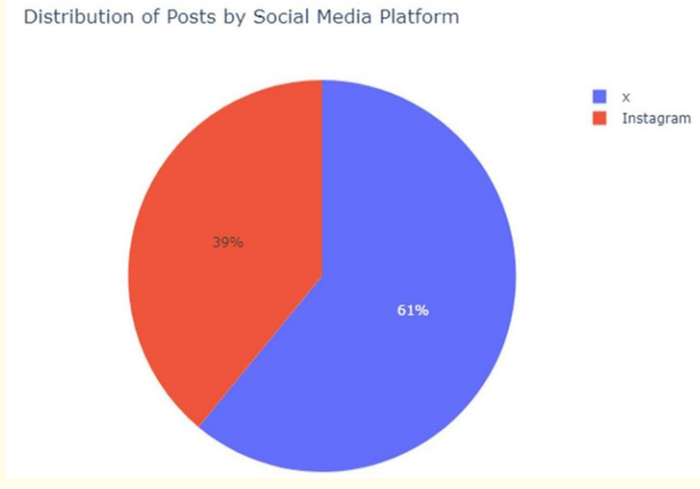

Figure 4: Pie chart illustrating the distribution of posts by social media platform.

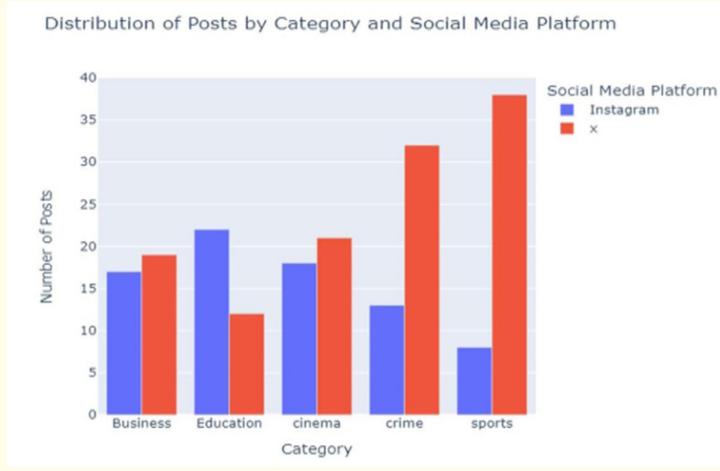

Figure 5: Grouped bar chart showing distribution of posts by category and social media platform.

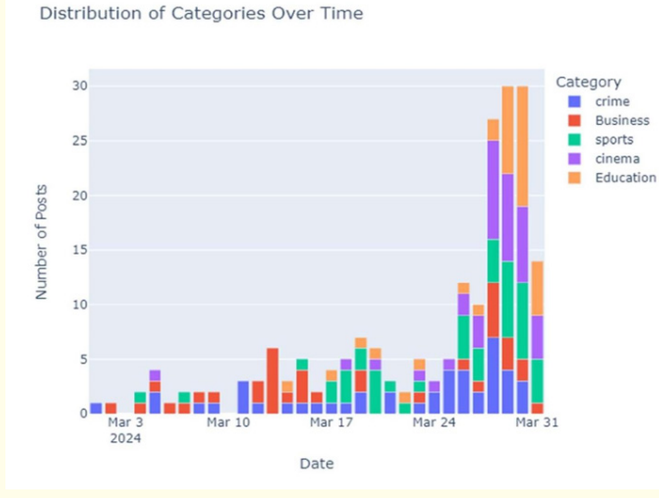

Figure 6: Stacked bar chart illustrating distribution of categories over time.

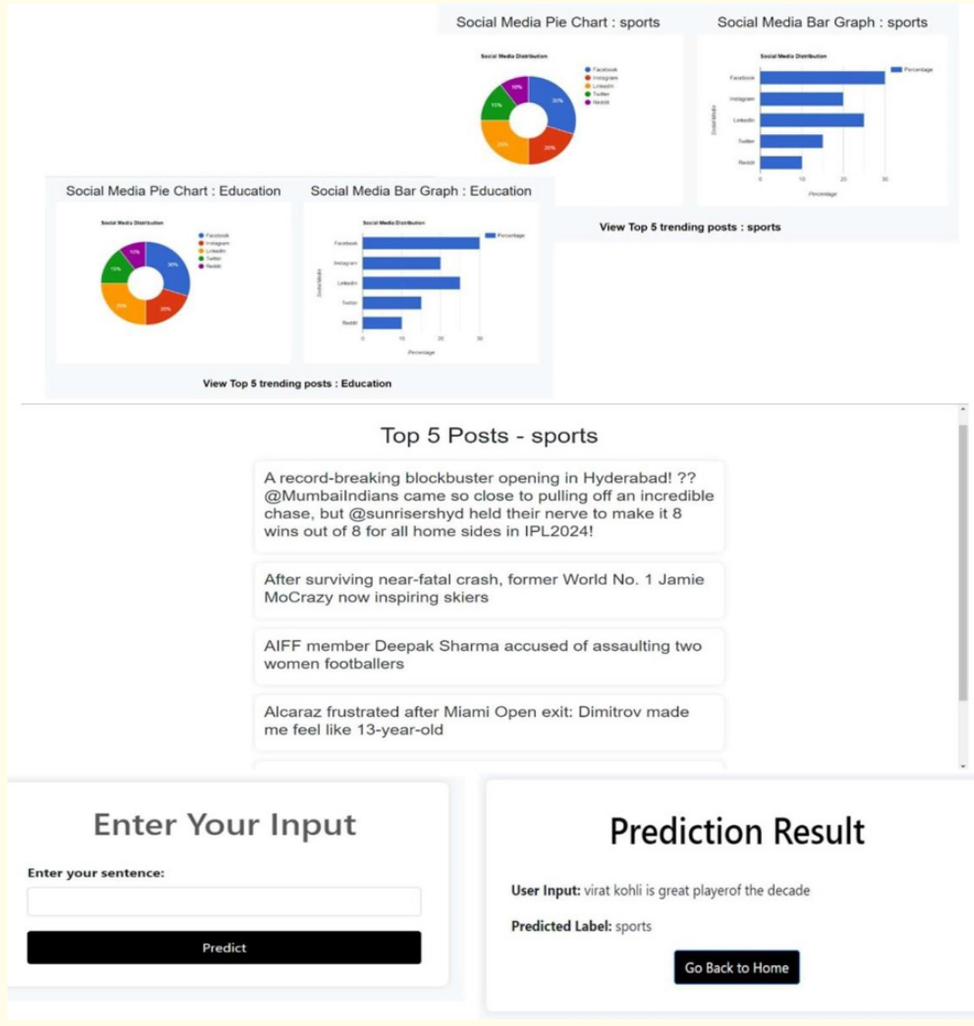

Figure 7: User Interface of all pages of our Web application.

well- defined clusters, the Calinski-Harabasz Index is of 5226.22, signifying good separation between clusters, and the Davies-Bouldin Index is of 0.38 indicating compact and well-separated clusters as described in Table 1. These metric's scores showthe effectiveness of our novel method for accurately categorizing social media posts, which helpsthe users like sports professionals, business professionals, police, students, movie makers get insightful information, know what's trending in their specified fields like sports, business, crime, education, cinema respectively.

In Figure 4,5,6, Visualizations of the categorized posts across different social media platforms provided clear insights into the distribution of trends within each category. For example, the stacked bar chart illustrated the temporal evolution of trends, while the grouped bar chart showcased the distribution of posts by category and platform, helping in the identification of emerging patterns and trends.

Our proposed method have many advantages over existing systems on social media post categorization. Unlike simple bag-of-words technique, our proposed methodology uses semantic embeddings to capture the contextual information of social media posts effectively, which resultsin meaningful categorization results. The use of advanced clustering algorithms such as DBSCANenhances the robustness and scalability of our proposed system, allowing it to adapt dynamically to varying data densities and distribution patterns. Additionally, our method addresses the challenge of informal language commonly found in social media posts by incorporating semantic embeddings pretrained on diverse linguistic data, improving the accuracy and robustness of the categorization process.

Despite the promising results achieved, our proposed system has certain limitations, our systemworks robust and accurate to large data sets which has dense distributions, it may not work ac-

curately when we use small datasets or the dataset is scattered with less densities.

Conclusion

In conclusion, our novel method effectively categorizes social media posts into different categorieslike sports, crime, cinema, education, business offering valuable insights for sports professionals, police, movie makers, students, and business professionals. Through advanced algorithms along with sentence embeddings together worked as a new system that can categorize posts accurately and advanced visualizations helped in easy data interpretations and helped in gaining meaningful insights. We provided a robust and novel framework for social media trend analysis. The web application part enhances accessibility, enabling real-time exploration of trends.

Acknowledgements

We worked under the supervision of our professor Dr. K. Kranthi Kumar, Associate Professor,Department of Science and Technology, SNIST.

Bibliography

1. Bitner M J and Zeithaml V A. "Technology's Impact on the Gaps Model of Service Quality". (2010).

2. Kapoor KK and Tamilmani K. "Advances in Social Media Research: Past, Present, and Future". Springer (2017).

3. Baatarjav EA and Dantu R. "Unveiling Hidden Patterns to Find Social Relevance". (2011).

4. Chen PL., *et al*. "Analysis of Social Media Data: An Introduction to the Characteristics and Chronological Process". Springer (2018).