Social Network Analysis

Identifying Communities of Interest

Aim:-

Analyze a social network dataset (e.g., online forum discussions, social media groups) to identify communities with shared interests or opinions. Explore the characteristics and potential applications of these communities.

Introduction:-

In the expansive realm of online social networks, communities serve as integral components that shape interactions, drive discussions, and mold opinions. Understanding the dynamics within these communities holds paramount importance across diverse applications, spanning from targeted marketing endeavors to the formulation of effective social policy interventions. At the heart of community analysis lies the identification of "Communities of Interest" – groups of individuals united by shared interests or perspectives.

In this analysis, we delve into a social network dataset, encompassing platforms like online forums or social media groups, where users engage in dialogue and exchange ideas. Employing data-driven techniques, our objective is to discern cohesive communities within the network, delineating groups characterized by similar interests or opinions. By exploring the inherent characteristics of these communities and delving into prevalent topics and sentiments, we aim to unravel the underlying structure of the social network. This understanding not only facilitates targeted marketing strategies for businesses but also equips social scientists and policymakers with invaluable insights to gauge public sentiment, track trends, and tailor interventions to address specific community needs effectively.

Purpose:-

- Understand Social Network Dynamics: Analyze social network data to comprehend how individuals connect and exchange opinions within digital communities.
- Detect and Characterize Communities: Utilize data-driven techniques to identify cohesive groups within the network, discerning patterns of interaction and shared interests.
- Explore Community Characteristics: Examine attributes like size, density, and connectivity to understand their structural significance.
- Analyze Topics and Sentiments: Delve into prevalent topics and sentiments to elucidate community dynamics and attitudes.
- Explore Potential Applications: Investigate applications in targeted marketing, social policy interventions, trend monitoring, and opinion analysis.
- Facilitate Knowledge Discovery: Generate actionable insights to inform decision-making and foster innovation in digital community engagement.

Twitter Sentiment Analysis:-

I LOVEEEE dogs
@beautygirl5 I love you <3
I enjoyed the food.
The game yesterday was intense!
@LOLTrish hey long time no see!
You put smiles on my face.
Today was a good day.
I love this notebook!
Positive
Negative
@bigdennis4 nobody asked you!
This week is not going as I had hoped
life has been like hell...
Don't force a joke if it ain't funny
I'm learning R programming.
So many homeworks !!!
Ugh. Can't sleep. Its 1:30am.
My Nokia 1110 died..

Twitter serves as a vast repository of human emotions, making sentiment analysis pivotal in discerning sentiments like happiness or anger. This analysis classifies opinions into positive or negative, aiding industries in gauging customer satisfaction and enhancing services. Leveraging social media data for semantic analysis has become commonplace, offering insights into public sentiment.

With over 200 million users, Twitter is a prominent microblogging platform globally. In 2017 alone, users shared approximately 8.3 million tweets per hour. Natural language processing techniques are instrumental in preprocessing tweets, involving tokenization and removing stop words and special characters. This preprocessing is vital for extracting meaningful insights and trends in people's emotions on Twitter.

Dataset:-

The sentiment140 dataset, sourced from Kaggle, comprises 1,600,000 tweets obtained via the Twitter API. Annotated with sentiment polarity (0 for negative, 4 for positive), it serves as a valuable resource for sentiment analysis tasks. The dataset encompasses various fields, including tweet ID, date, user, and tweet text, facilitating comprehensive analysis. Preprocessing techniques, such as tokenization and stop word removal, are employed to prepare the text data for sentiment classification.

Source of Data**:** https://www.kaggle.com/kazanova/sentiment140

Problem description:-

To develop a sentiment analyzer aimed at addressing the complexities inherent in discerning sentiment from Twitter tweets, particularly focusing on categorizing them as positive or negative sentiments. This endeavor involves the utilization of neural network methodologies, leveraging TensorFlow for robust implementation and analysis.

Evolution measures:-

After model training, a series of evaluation measures are applied to assess the predictive efficacy. The evaluation process entails the utilization of a set of predefined metrics to gauge the performance and accuracy of the models:

- Accuracy
- Confusion matrix with plot
- ROC Curve

Technical Approach:-

I am utilizing the Python programming language for implementation, with Jupyter Notebook as my chosen development environment. For this project, I have selected the Sentiment 140 dataset, which I will divide into a 70% training set and a 30% testing set. Once the model training is complete, I will proceed to evaluate the performance of the trained model thoroughly.

Data Preparation:-

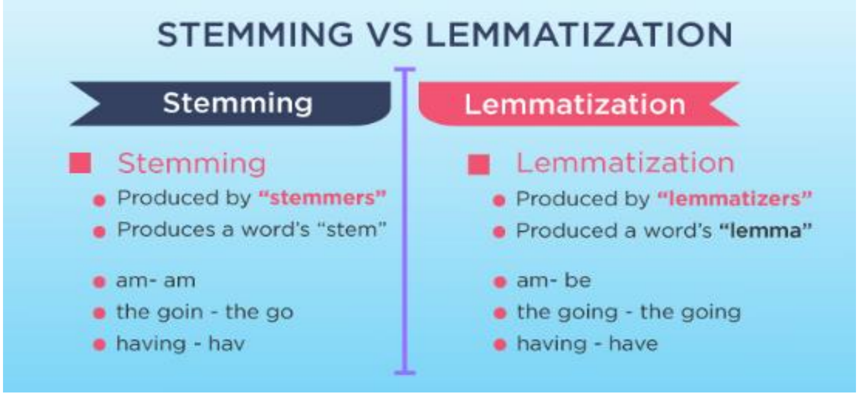
I began by selecting relevant columns and assigning class labels. To ensure smooth processing on my machine, I downsized the dataset to a manageable size.

Next, I cleaned the text data by converting it to lowercase and removing stop words, punctuation, emails, URLs, and numbers. I then applied tokenization to break down the text into individual words.

- **Applying stemming and lemmatization:-**

Further refinement involved applying stemming and lemmatization techniques to standardize words to their base forms, enhancing the accuracy of my analysis.

Here is the concept of stemming and lemmatization:-



By applying stemming and lemmatization, I effectively standardized the text data, enabling more accurate analysis and classification of tweet sentiments.

- **Labels and Inputs:-**

The sentiment labels were assigned to the tweets to indicate whether they were positive or negative. Inputs, comprising the tweet text, were extracted to serve as the data fed into the machine learning model.

- **Training, Validation, and Testing Data:-**

Finally, I partitioned the dataset into training and testing sets, allocating 70% of the data for training the model and reserving 30% for evaluating its performance. This split ensured that the model could be trained on a sufficiently large dataset while also allowing for robust testing to assess its effectiveness in classifying tweet sentiments. The validation data were employed to assess the model's performance during training.

Data Featurization:-

The tweet text was converted into numerical feature vectors using tokenization. A maximum of 500 features, representing important words distinguishing between positive and negative tweets, were selected for training the model.

Model Implementation:-

A TensorFlow-based neural network model was constructed for sentiment analysis. The model architecture included layers for embedding, LSTM (Long Short-Term Memory), dense, activation, and dropout, designed to effectively learn from and process the input data.

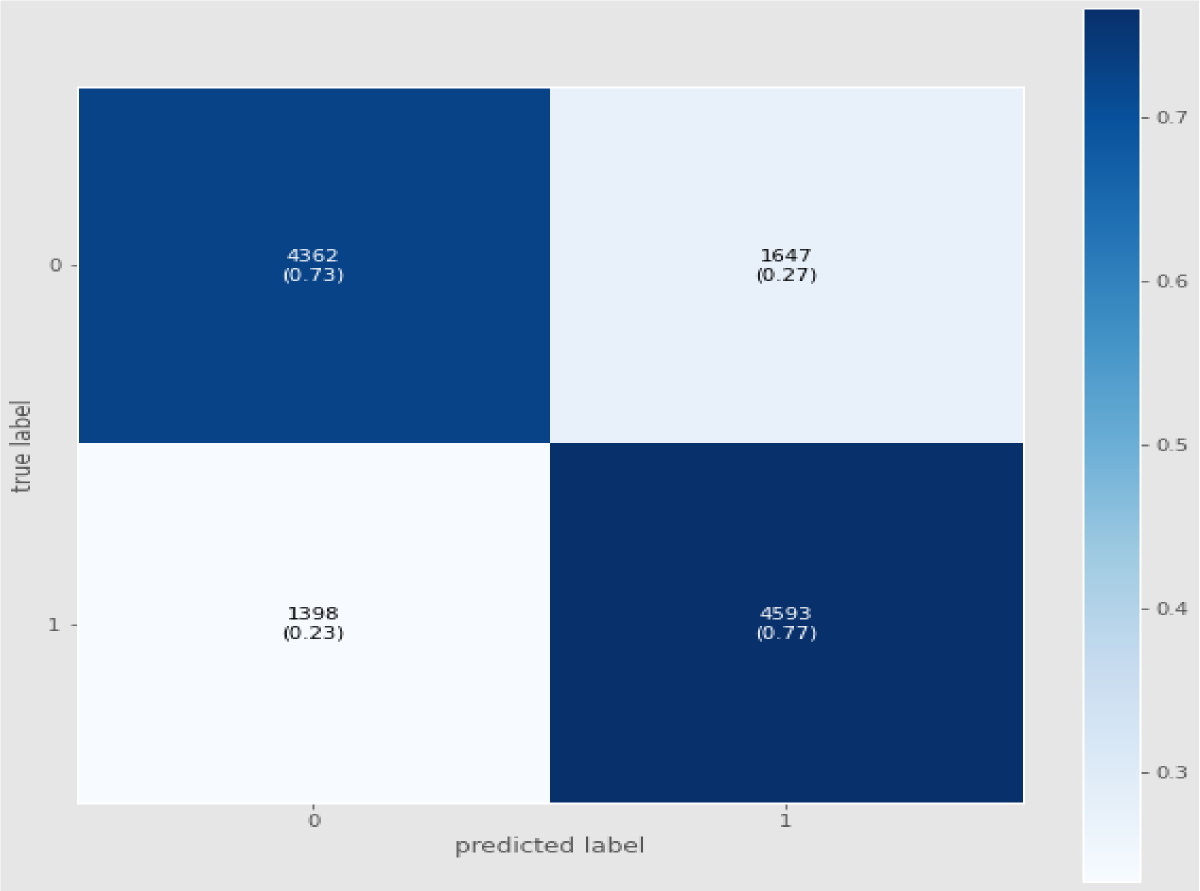
Model Compilation and Training:-

The model was compiled with appropriate loss, optimizer, and metrics settings. Training involved feeding the training data into the model over multiple epochs, adjusting the model's internal parameters to minimize the loss function and improve accuracy.

Model Evaluation:-

After training, the model's performance was evaluated using the testing data. Accuracy, representing the proportion of correctly classified tweets, was calculated to assess the model's effectiveness in sentiment analysis.

Confusion matrix:-

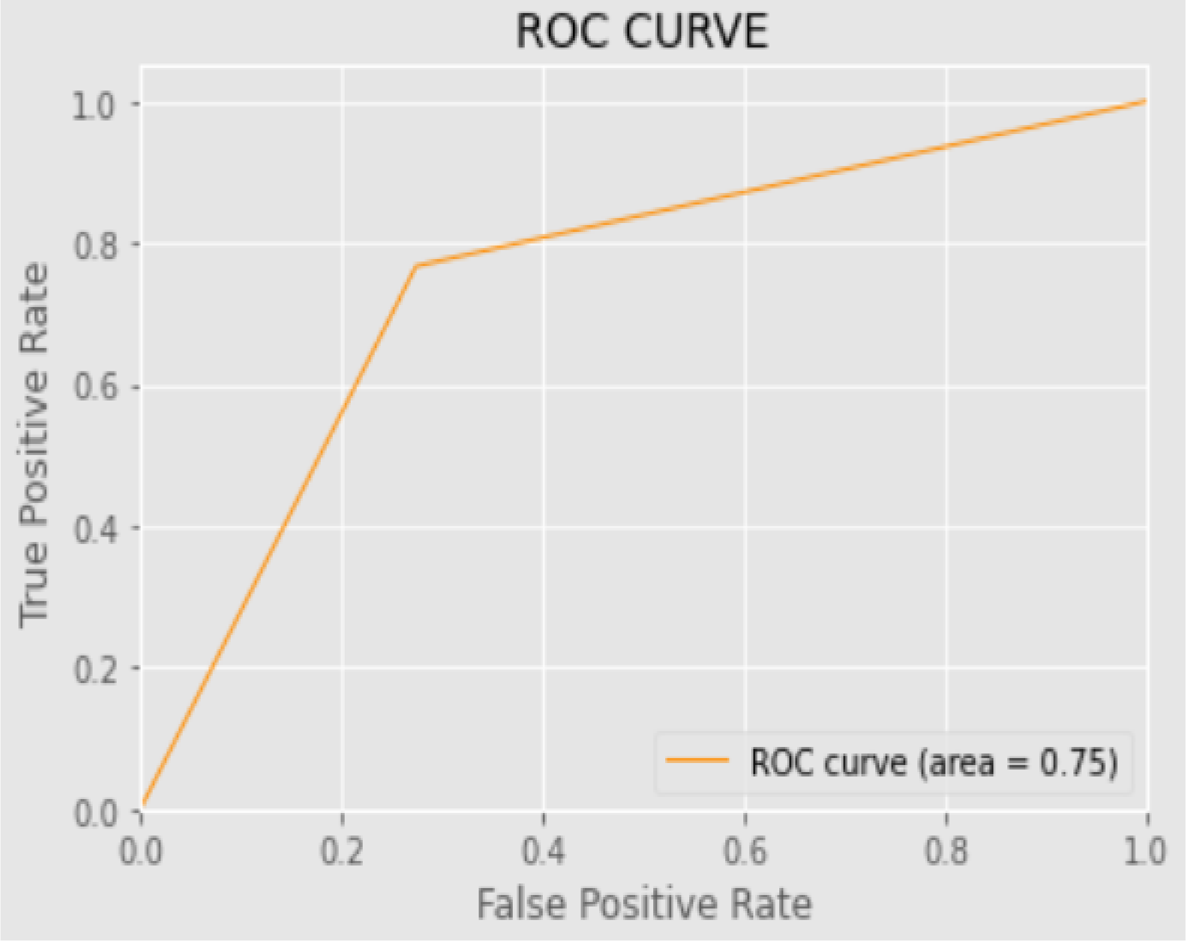


The evaluation metrics employed to assess the model's performance provide critical insights into its effectiveness. Correct predictions are represented by dark blue boxes, while incorrect predictions are denoted by sky blue boxes.

Among the predictions, 4610 tweets were accurately classified as negative sentiments, while 1399 tweets were incorrectly classified as positive sentiments when they were actually negative. Similarly, 4247 tweets were correctly classified as positive sentiments, whereas 1744 tweets were erroneously categorized as negative sentiments when they were, in fact, positive.

These metrics offer a comprehensive understanding of the model's classification accuracy and error rates, guiding further refinement and optimization efforts.

ROC Curve:-



The ROC curve serves as a pivotal tool in assessing the model's performance comprehensively. It visualizes the model's progression from 0 percent predictions, gradually advancing towards true positive predictions, which denote correct classifications.

The ROC curve, short for the receiver operating characteristic curve, offers insights into the classification model's performance across various thresholds. It plots two crucial parameters: the True Positive Rate, representing correct predictions or classifications, and the False Positive Rate, indicating erroneous predictions or classifications. This graphical representation facilitates a nuanced understanding of the model's efficacy across different thresholds, aiding in informed decision-making and optimization strategies.

Conclusion:-

In this study, we employed the Twitter sentiment analysis dataset to conduct a comprehensive analysis of sentiment trends. By leveraging various data preprocessing techniques, we meticulously prepared the tweet text, eliminating extraneous elements to streamline the analysis process.

Utilizing TensorFlow, we trained a sophisticated model, meticulously configuring all relevant settings to ensure optimal performance. Subsequently, we rigorously evaluated the model's efficacy through diverse evaluation measures, providing a robust assessment of its capabilities.

Our methodology and findings offer valuable insights for practitioners interested in text-based projects. While our focus was on binary classification, our approach is adaptable to a wide range of text analysis tasks, albeit with minor adjustments tailored to specific project requirements.

In conclusion, this study contributes to the broader field of sentiment analysis, demonstrating the effectiveness of data-driven approaches in extracting meaningful insights from social media data. As the landscape of text analysis continues to evolve, our methodology provides a solid foundation for future research endeavors and practical applications.

Link of project:-

For access to the full project and code implementation, please visit:-

https://github.com/PratikshaPandaPKP/Web-And-Social-Network-Analysis-Assignment