UTM
UNIVERSITI TEKNOLOGI MALAYSIA

PROJECT PROPOSAL PRESENTATION

TOPIC-BASED ANALYSIS OF SOCIAL MEDIA POSTS USING RNN AND LSTM

PRESENTED BY   : ZHU QIAN
MATRIC NO   : MCS241017
VENUE   : DISCUSSION ROOM 4, LEVEL 2, BLOCK N28A
DATE   : 17TH JAN 2025
LECTURER   : ASSOC. PROF. DR MOHD SHAHIZAN BIN OTHMAN

Innovating Solutions

INTRODUCTION

BACKGROUND

- The weaknesses of the current working model underscore the necessity for integrating new data sources to enhance operational efficiency.

- A more sophisticated and systematic approach is required to address these challenges effectively.

- By leveraging data science and machine learning techniques to analyze topic-based social media posts, both enterprises and governments can identify valuable insights to mitigate these issues.

PROBLEM STATEMENT

- Social media posts are presented in various data formats such as text and emojis often mixed with irrelevant content and influenced by emotions. Traditional survey methods frequently produce false or unreliable results, which often deviate significantly from the actual situation.

RESEARCH QUESTIONS & OBJECTIVES

Research Question

- **RQ 1**: How can data science and machine learning techniques be leveraged to filter out irrelevant, non-compliant, and emotionally biased content from social media, ensuring the quality and reliability of data for decision-making?

- **RQ2**: What strategies can be developed to distinguish genuine user interactions from fake accounts and internet trolls on social media platforms, thereby improving the accuracy of public sentiment analysis?

- **RQ3** : How can social media insights offer a more cost-effective alternative to traditional web surveys for decision-making?

Research Objectives

- To identify significant relationships between the content of posts and the topic.

- To build and develop analytical models that capture the topic inclination of posts.

- To measure public reactions to the topic by summarizing the analysis results.

- To define the best parameter estimate.

SCOPE OF THE STUDY

This study will focus on assessing the quality of social media data by identifying and mitigating the influence of fake users, internet trolls, and irrelevant content, using advanced data science and machine learning techniques.

The study will explore how social media data can be used to analyze public sentiment and behavior, providing insights into real-world reactions to various topics and issues.

The study will compare the effectiveness and cost-efficiency of social media data versus traditional web surveys, particularly in terms of data authenticity, respondent verification, and decision-making accuracy.

This research will examine how the integration of reliable social media data can enhance decision-making processes for enterprises and governments, particularly in response to dynamic societal and market changes.

LITERATURE REVIEW

LITERATURE REVIEW

Sentiment Analysis in Social Media

- Text Classification
- Sentiment Detection
- Natural Language Processing (NLP)
- Social Media Sentiment

Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) Models

- Deep Learning
- Sequence Modeling
- RNN Architecture
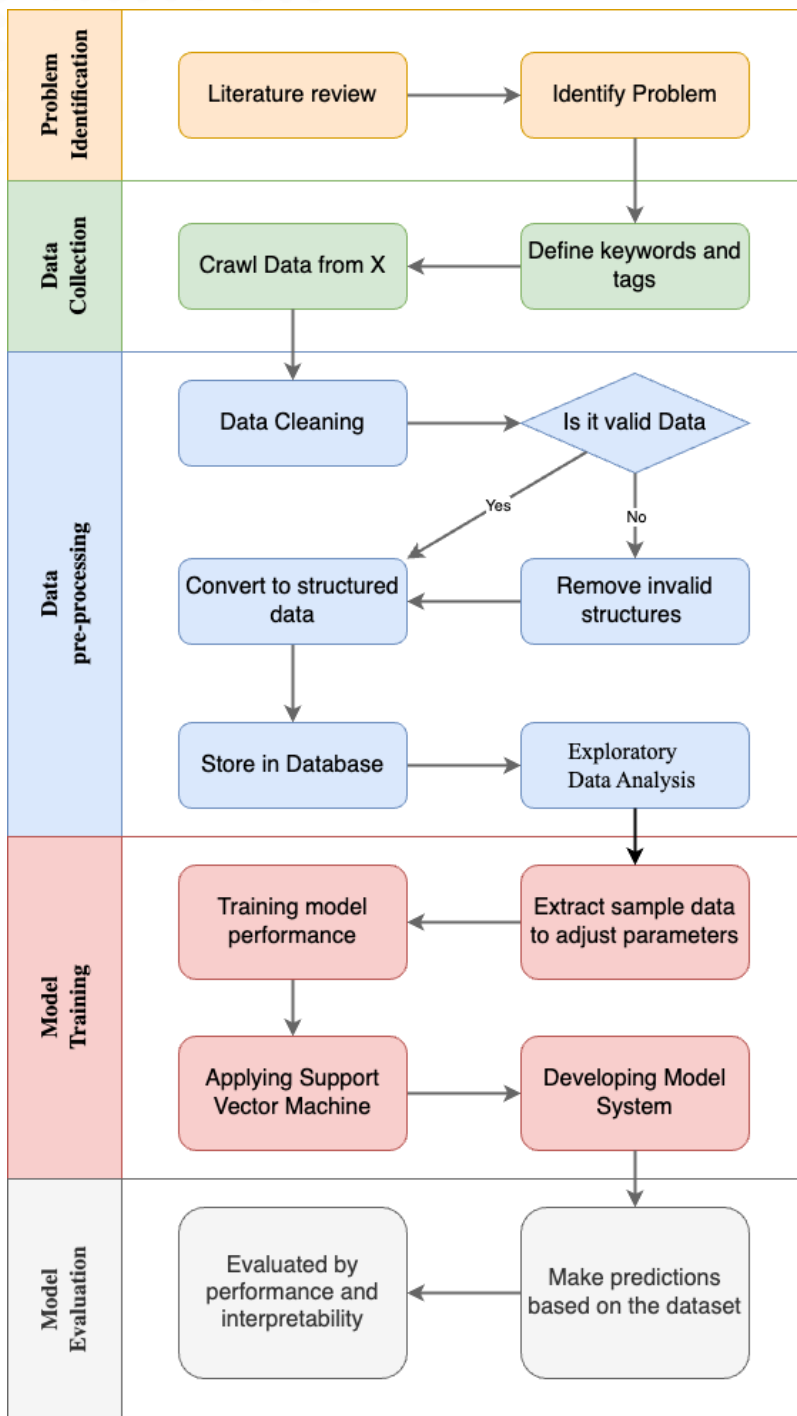- LSTM Networks

Key Components

Topic-Based Analysis of Social Media

- Topic Modeling.
- Latent Dirichlet Allocation (LDA)
- Text Mining
- Topic Detection

4. Challenges in Social Media Data Quality and Fake User Detection

- Data Preprocessing
- Bot Detection
- Content Filtering

www.utm.my

METHODOLOGY

A research framework for analyzing emotions using deep learning and support vector machine algorithm

DATA SOURCES AND COLLECTION METHODS

Data Source	Data Type	Data Collection Method	Data Collection Criteria	Data Collection Period
• X platform serves as the primary source for the social media data utilized in this study.	• The core unit of data collected consists of individual tweets, including their text content, timestamps, user information (if available), and associated metadata (e.g., hashtags, mentions).	• The research employed the Twitter API to programmatically access and retrieve tweets from the platform.	• Specific keywords relevant to the research topics were used to filter tweets. • Relevant hashtags were included in the search criteria to further refine the data collection process.	• This timeframe was chosen to capture the public sentiment surrounding the events during a period of heightened media coverage and public discussion

DATA CLEANING

- Remove duplicate data within the dataset.

- Identify and remove any URLs (web addresses) present in the tweets. These links often contain irrelevant information.

- Remove mentions of other Twitter users (e.g., "@username")..

- Remove hashtags (e.g., "#topic"). While hashtags can provide context, they can also introduce bias or noise.

- Remove punctuation marks (e.g., periods, commas, exclamation marks) and special characters (e.g., $, %, &).

- Convert all text in the tweets to lowercase.

Model Training

- Select the appropriate SVM classifier from the scikit-learn library in Python.

- Define Model Parameters (Hyperparameters) includes Linear, Radial Basis Function and Polynomial.

- Utilize the fit() method of the chosen SVM classifier to train the model on the TF-IDF-transformed training data (X_train_tfidf) and their corresponding labels (y_train).

- save the trained SVM model using pickle.dump() method.

VISUALISATIONS

Sentiment Distribution

- Histogram showing the frequency of different sentiment scores.
- Insights showing whether tweets are predominantly positive, negative, or neutral.

Engagement

- Use scatter plots to visualize the relationship between sentiment scores and engagement metrics.
- Insights show that tweets with high positive sentiment tend to receive more likes and retweets.

Word Clouds

- Create separate word clouds for tweets with positive, negative, and neutral sentiment.
- Gain a deeper understanding of the language and vocabulary used to express different emotions.

www.utm.my

INITIAL INSIGHTS

Significance in Expressing Sentiment

- "Favorites" were the primary method for users to express support or agreement.

Polarized Sentiment

- Republican voters, led by Donald Trump, displayed significantly more positive sentiment with minimal neutral opinions.
- Democratic supporters exhibited a larger proportion of neutral and negative sentiment, reflecting a lack of unified trust in their candidate.

Overall Positive Public Sentiment

- The general sentiment among Twitter users indicated optimism and support for the Republican Party, positioning them as likely leaders in the post-election landscape.

THANK YOU

univteknologimalaysia   utm.my   utmofficial