CS3263 Project Proposal

Chen Jiarui - A0245205N - e0893429@u.nus.edu
Gu Haiwei - A0245155H - e0890717@u.nus.edu
Xiao Yan - A0245626B - e0902032@u.nus.edu

Topic: AI Assistant for News Sentiment Classification

Type: Application Project

Background

With the vast amount of news published daily, understanding the overall sentiment of a given time period is crucial for analyzing public sentiment, market trends, and social dynamics. News sentiment analysis applies NLP methods to categorize news articles as positive, negative, or neutral.

Problem statement

Traditional news consumption often lacks a structured way to assess the overall sentiment of a given period, making it difficult to track shifts in public mood and media narratives. Current sentiment analysis models often focus on individual articles rather than summarizing trends over time. This project aims to bridge this gap by developing a sentiment analysis model that aggregates news sentiment from certain timeline aspects. It utilizes a Kaggle news dataset to analyze sentiment trends (Useful data include Source, e.g., BBC, Title, Type, e.g. Business, Description, Sentiment, e.g. Positive, Neutral, Negative). Our goal is to train a sentiment analysis model that helps users to accurately summarize the sentiment of news over a selected timeframe (such as concluding the sentiment as 80% positive, 10% Neutral, and 10% Negative, giving a week's news). By summarizing sentiment over time, people can identify key patterns and shifts in public opinion.

Approach:

Our approach involves preprocessing news texts through tokenization, stemming, stopword removal, and TF-IDF or word embeddings for feature extraction. We will experiment with Naïve Bayes, Logistic Regression, and a simple Neural Network to classify sentiment into positive, negative, or neutral categories. Hyperparameter tuning and cross-validation will optimize model performance, and ablation studies will assess the impact of different preprocessing and feature engineering techniques. Performance will be evaluated using accuracy, F1-score, and confusion matrices to ensure robustness.

Team role

Chen Jiarui: Pre-process the Kaggle dataset with tools such as pandas, classify the types of news for more efficient training.
Gu Haiwei: Build environment on SoC Computing Cluster, deploy training code, and monitor training process.
Xiao Yan: Experiment with different models, features, and pre-processing methods. Analyse how each section contributes to the final result and where improvement can be applied.