Opinion Analysis of Media Basis

Project Proposal By Team 42: Wu Tong, Yang Haohong

### 1. Target Task

The primary objective of this project is to analyze the sentiment and opinions expressed by different media outlets towards specific Singapore topics and entities. The goal is to identify potential biases or consistent perspectives among various global media sources. By understanding these sentiments, we aim to uncover patterns in how Singapore-related news is portrayed internationally.

As an optional downstream task, we plan to utilize the developed sentiment analysis model to predict stock price movements based on financial and technology news. This could provide valuable insights for investors and stakeholders interested in the Singaporean market.

### 2. Dataset

We leverage the Global Geographic Graph Singapore (GGGSG) dataset from the GDELT Project. The dataset has been filtered to include English news articles mentioning Singapore from April 4th, 2017, up to July 19th, 2024 (CountryCode == "SN"). This filtered dataset contains **9,185,305** rows, each representing a mention of a Singapore-related entity.

**Key features of the dataset:**

* **Contextual Text**: Up to 600 characters of surrounding text where Singapore is mentioned, converted to lowercase and with punctuation removed.
* **DocTone**: Sentiment score of the article.
* **Metadata**: Includes date and time, URL, title, language code, and geographic information.

To support our analysis, we will utilize a variety of datasets, ranging from general news articles to financial and product review data. The diversity of datasets ensures broad coverage and robust results. These datasets include:

* **Singapore News Articles**: A machine-annotated dataset focusing on Singapore-related news.
* **Financial Phrase Bank**: A collection of financial news headlines with sentiment annotations from the perspective of retail investors.
* **Million News Headlines**: News headlines sourced from Australian news over 19 years.
* **Amazon Review Data**: A comprehensive dataset with product reviews and associated metadata.
* **Stanford Sentiment Treebank**: A corpus with labeled parse trees for sentiment analysis.
* **WordStat Sentiment Dictionary**: A sentiment classification dictionary based on multiple sources.
* **Social Media Sentiment**: Data capturing emotions and interactions from social media.
* **Hotel Reviews**: Reviews of 1,000 hotels, suitable for aspect-based sentiment analysis.
* **NewsMTSC**: A dataset for target-dependent sentiment classification on policy issues.
* **FNSPID**: A dataset that integrates financial news with stock price data for market prediction.
* **Twitter Financial News Sentiment**: Annotated tweets related to financial markets.
* **IMDB**: A dataset of movie reviews for binary sentiment classification.
* **Yelp Reviews**: Reviews used for sentiment analysis from the Yelp Dataset Challenge.

### 3. Comparison Baseline

We will evaluate our models against several baselines for sentiment classification on a subset of the Singapore News Articles dataset:

* **Baseline 1**: TF-IDF + Logistic Regression
  + (Negative): Precision: 0.70, Recall: 0.49, F1-Score: 0.58
  + (Neutral): Precision: 0.63, Recall: 0.72, F1-Score: 0.67
  + (Positive): Precision: 0.61, Recall: 0.63, F1-Score: 0.62
* **Baseline 2**: BERT Embedding + Logistic Regression
  + (Negative): Precision: 0.65, Recall: 0.61, F1-Score: 0.63
  + (Neutral): Precision: 0.64, Recall: 0.69, F1-Score: 0.66
  + (Positive): Precision: 0.64, Recall: 0.58, F1-Score: 0.61
* **Baseline 3**: BERT Embedding + Random Forest
  + (Negative): Precision: 0.72, Recall: 0.41, F1-Score: 0.52
  + (Neutral): Precision: 0.60, Recall: 0.84, F1-Score: 0.70
  + (Positive): Precision: 0.72, Recall: 0.46, F1-Score: 0.56

### 4. Intended Approach

To improve upon the baseline, we intend to explore advanced natural language processing techniques, including:

* **Attention Mechanisms**: Implement models that utilize attention mechanisms to better capture contextual information and the importance of different words in a sentence.
* **Transformer-Based Models**: Utilize pre-trained models like RoBERTa or BERT to obtain contextualized word embeddings that can enhance sentiment classification performance.
* **Word2Vec Embeddings**: Incorporate word embeddings to capture semantic relationships between words.
* **Ensemble Methods**: Combine predictions from multiple models to improve overall performance.
* **Advanced Classifiers**: Experiment with classifiers like Random Forests, Gradient Boosting Machines, or neural networks to capture nonlinear relationships.

We aim to integrate these methods into our pipeline and assess their impact on the model's ability to accurately classify sentiments and detect biases across media outlets.

### 5. Challenges

Several challenges may arise during the course of this project:

* **Large Dataset Processing**: Handling and processing a dataset of over 9 million records (8 GB in size) requires efficient computational strategies, such as distributed computing with PySpark.
* **Imbalanced Data**: There may be an imbalance in the distribution of sentiment classes, which can affect model training and performance.
* **Annotation Quality**: The dataset is machine-annotated using traditional NLP methods, which may introduce errors or inconsistencies in sentiment labels.
* **Domain Adaptation**: The diversity of our datasets poses a challenge for domain adaptation, as models trained on one dataset may not perform well on others.
* **Label Accuracy**: Many datasets are machine-annotated, and their labels may contain errors. We may need to use large language models (LLMs) to generate higher-quality labels.

### 6. Conclusion

Through this project, we aim to improve sentiment classification models by leveraging multiple datasets and advanced techniques like model fusion. By exploring the relationship between media sentiment and stock prices, we hope to contribute valuable insights into financial forecasting based on media trends.