Sentiment Analysis of News Headlines and Stock Price Movements

Methodology

Data Preprocessing

1. Loading Data:
The dataset contains news headlines and corresponding stock price movements. It was loaded into a DataFrame for further processing.

2. Text Preprocessing Techniques Applied:

- **Tokenization**: The text data was split into individual words (tokens) using the nltk library. This helps in processing the text data at the word level.
- **Lowercasing**: All text was converted to lowercase to ensure uniformity, as 'Good' and 'good' should be treated the same.
- **Removal of Stop Words**: Commonly used words (e.g., 'the', 'and', 'is') that do not contribute much to the sentiment were removed using the nltk stopwords list.
- **Handling Punctuation**: Punctuation marks were removed since they do not contribute to the sentiment.
- **Stemming**: Words were reduced to their root forms using the PorterStemmer from the nltk library. For example, 'running' was reduced to 'run'.
- **TF-IDF Vectorization**: The text data was converted into numerical form using the TF-IDF Vectorizer. This method assigns a weight to each word based on its frequency in the document and the corpus.
- **Bag of Words Vectorization**:In this method each unique word in a sentence is counted and their frequencies is used to denote the sentence in vector form.

The rationale behind each preprocessing step was to clean and standardize the text data to make it suitable for machine learning models. Tokenization and lowercasing ensure uniformity, removal of stop words reduces noise, and stemming ensures that different forms of a word are treated as a single feature. TF-IDF and Bag of Words vectorization transforms text data into numerical form that the models can process.

Note: Lemmatization was not applied in this analysis. Lemmatization, which converts words to their base forms based on context, can be an alternative or complementary approach to stemming for further improving text normalization.

Data Description

The dataset used in this analysis consists of news headlines related to various companies along with the corresponding stock price movements. Each record includes:

- **Headline**: The text of the news article headline.
- **Stock Movement**: The binary indicator of stock price movement (e.g., increase or decrease).
- **Date**: The date is given which has been used to divide the dataset for testing and training.

Model Selection

Chosen Models

1. **Naive Bayes Classifier**
    - **Advantages**:
        - Simplicity: Easy to implement and interpret.
        - Efficiency: Computationally efficient and works well with large datasets.
        - Effectiveness: Performs well with text data and is often used for document classification tasks.
    - **Limitations**:
        - Assumption of Independence: Assumes that features are independent, which is not always true in real-world data.
        - Handling of Rare Words: May not perform well with rare words or phrases.

2. **Random Forest Classifier**
    - **Advantages**:
        - Robustness: Handles a large number of features well and is less likely to overfit compared to single decision trees.
        - Versatility: Can capture complex interactions between features.
    - **Limitations**:
        - Computationally Intensive: Requires more computational resources compared to simpler models.
        - Interpretability: More challenging to interpret compared to simpler models like Naive Bayes.

NOTE: Both bag of words and TF-IDF has been used in the given models as a vectorizer.

Model Evaluation

Performance Metrics

1. **Accuracy**: The proportion of correctly classified instances out of the total instances. It provides a general sense of how well the model is performing.
2. **Precision**: The ratio of true positive predictions to the total predicted positives. It indicates the accuracy of the positive predictions.
3. **Recall**: The ratio of true positive predictions to the total actual positives. It measures the model's ability to identify all positive instances.
4. **Confusion Matrix**: The confusion matrix is also drawn so as to tell the number of correct and wrong predictions of each class.

These metrics are relevant because they provide a comprehensive view of the model's performance. Accuracy gives an overall performance measure, precision shows how many of the positive predictions were actually correct, and recall indicates how well the model identifies positive instances.

Analysis and Insights

Model Predictions and Insights

- **Correlation**: There is a noticeable correlation between the sentiment of news headlines and stock price movements. Positive news often correlates with stock price increases, while negative news correlates with stock price decreases.
- **Sentiment Trends**: The model's predictions can help identify sentiment trends in news headlines. For instance, a series of positive headlines might indicate a bullish trend in the stock market.
- **Stock Market Prediction**: By analyzing the sentiment of news headlines, investors can gain insights into potential stock price movements and make informed decisions.

Conclusion

In conclusion, the Naive Bayes Classifier, Random Forest Classifier with TF-IDF Vectorizer and Bag of Words vectorizer were used for sentiment analysis of news headlines and the best accuracy was with **Naive Bayes and TF-IDF and Random Forest with Bag of words** both of which gave **84.92 % accuracy**. Various preprocessing techniques were applied to clean and standardize the text data. The models' performance was evaluated using accuracy, precision, and recall, which provided a comprehensive view of their effectiveness. The analysis of model predictions revealed a correlation between news headlines and stock price movements, offering valuable insights for stock market prediction.