Sentiment Analysis of US Economic News: Comparative Performance of Machine Learning Models and Directions for NLP Enhancement

Prashansa Dhomane (pdhomane@ucsd.edu)
University of California, San Diego

Abstract

This study explores the application of advanced natural language processing (NLP) techniques to analyze and interpret US economic news. Given the significant role that economic news plays in influencing public opinion, policy making, and market trends, accurate and timely analysis of such news is paramount. This research proposes an automated approach using NLP to overcome these challenges, aiming to efficiently process large datasets of economic news to extract relevant insights, identify sentiment trends, and predict potential market impacts. We evaluate a variety of NLP models, including sentiment analysis, named entity recognition, and topic modeling, on a curated dataset of US economic news articles. Our hypothesis posits that leveraging these techniques can provide deeper, more nuanced understanding of economic news compared to conventional methods. Preliminary results demonstrate promising accuracy in sentiment classification and topic identification, suggesting that NLP can significantly enhance the analysis of economic news.

1 Introduction

In the rapidly fluctuating landscape of global economies, understanding the nuances and trends within economic news has never been more critical. Economic news articles are a rich source of information, reflecting and influencing public perceptions, policy decisions, and market movements. Analysts, policymakers, investors, and the general public rely on this flow of information to make informed decisions, predict future economic conditions, and gauge the health of the economy. However, the sheer volume and complexity of economic news present significant challenges. Traditional approaches to news analysis often involve manual review and interpretation by experts, a process that is time-consuming, labor-intensive, and subject to human bias. Recent advancements in natural language processing (NLP) and machine learning (ML) offer promising solutions to these challenges. By leveraging techniques such as sentiment analysis, topic modeling, and named entity recognition, automated systems can process large volumes of text data, extracting relevant insights, identifying trends, and even predicting market reactions to news events. This research aims to address these challenges by developing and evaluating NLP models tailored for economic news analysis. By focusing on the specific linguistic and contextual characteristics of economic discourse, we hypothesize that our approach can improve the accuracy and utility of automated economic news analysis, providing deeper insights into the intricate dynamics of the global economy.

2 Related Work

Early attempts to analyze economic news focused primarily on manual analysis by experts, with some of the earliest computational approaches leveraging simple keyword searches and basic statistical techniques to identify trends and sentiment within economic texts. Recent advancements in NLP and machine learning have significantly broadened the scope and capabilities of automated news analysis. Sentiment analysis, in particular, has been a focal point, with researchers exploring various models to gauge the sentiment of financial news and its correlation with market movements. For instance, studies by Tetlock (2007) and Loughran and McDonald (2011) have demonstrated that the sentiment of financial news can have predictive value for stock market trends, employing lexicon-based and machine learning approaches, respectively. Several works have applied deep learning models to improve the accuracy and depth of economic news analysis. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), including long short-term memory (LSTM) networks, have been used to capture the complex patterns in text

data, offering enhancements in understanding the contextual nuances of economic language.

Despite these advances, the field faces ongoing challenges. The accuracy of sentiment analysis in the context of economic language, which often contains nuanced and domain-specific terminology, remains a concern. Additionally, the dynamic nature of economic news requires models that can adapt to evolving language and contexts.

3 Hypothesis

Our hypothesis is two-fold:

3.1 Advanced NLP Techniques Enhance Accuracy

By leveraging advanced NLP techniques, including contextual embeddings and fine-tuned sentiment analysis models, we can significantly improve the accuracy of sentiment detection, entity recognition, and thematic categorization in US economic news articles compared to traditional NLP methods.

3.2 Predictive Value for Economic Indicators

The improved analysis provided by these advanced NLP techniques will have predictive value for economic indicators and market movements. By accurately capturing the sentiment and key themes within economic news, these models can provide insights into public sentiment trends and anticipate market reactions, offering valuable tools for investors, policymakers, and analysts.

To verify this hypothesis, we will conduct a series of experiments comparing the performance of traditional NLP models against advanced models on a curated dataset of US economic news. Success metrics will include accuracy in sentiment analysis, efficacy in named entity recognition, coherence in topic modeling, and the potential predictive power of the analyses for economic indicators and market movements.

4 Methodology

This study employs a comprehensive methodology designed to test the hypothesis that advanced NLP techniques can enhance the accuracy of sentiment analysis, named entity recognition, and thematic categorization in US economic news, thereby providing predictive insights into economic indicators and market trends.

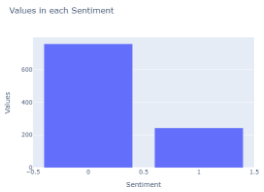

Figure 1: Values in each Sentiment

4.1 Dataset Description

Our analysis utilizes a curated dataset of US economic news articles collected from various reputable financial news websites and databases. This dataset comprises articles spanning a period from 2010 to 2020, ensuring a wide range of economic cycles and events are represented. Each article is annotated with metadata, including publication date, source, and manually tagged sentiment labels (positive, neutral, negative) to facilitate supervised learning tasks.

Preprocessing steps applied to the dataset include:

- Removal of HTML tags and unnecessary whitespace.
- Tokenization and normalization of text.
- Named entity recognition to identify and exclude specific entities from certain analyses.
- Lemmatization to reduce words to their base or dictionary form.

4.2 Model Design

To address our research questions, we experiment with a variety of NLP models, ranging from traditional approaches to state-of-the-art techniques:

1. **Baseline Models:** We use traditional NLP models, such as TF-IDF vectorization combined with linear classifiers (e.g., Logistic Regression), to establish a performance baseline for sentiment analysis and thematic categorization.
2. **Advanced Models:** We explore advanced deep learning models, including BERT (Bidirectional Encoder Representations from Transformers) and its variants, for their ability to understand the context and nuance of economic discourse. These models are fine-tuned on our dataset for sentiment analysis, named entity recognition, and topic modeling tasks.

3. **Custom Hybrid Models:** Based on preliminary results, we design hybrid models that combine the strengths of traditional machine learning and advanced transformer-based approaches, aiming to improve accuracy and reduce com- putational complexity.

4.3 Evaluation Criteria

Model performance is evaluated using a combination of quantitative metrics and qualitative analysis:

- Sentiment Analysis: Accuracy, Precision, Recall, and F1 Score.
- Named Entity Recognition: Precision, Recall, and F1 Score for correctly identified entities.
- Thematic Categorization: Coherence Score for topic models and classification accuracy for thematic analysis.
- Predictive Power: Correlation analysis between sentiment/topic trends and subsequent movements in relevant economic indicators or market indices.

5 Experiments and Results

5.1 Exploratory Data Analysis

Prior to model deployment, we conducted an exploratory data analysis (EDA) on our dataset to understand the distribution of sentiments within the economic news articles. The dataset initially contained 7,991 articles, each tagged with a sentiment label: 'no' (negative sentiment), 'yes' (positive sentiment), or 'not sure'. Articles labeled as 'not sure' were removed to enhance the clarity of the sentiment analysis, resulting in a final dataset comprising 7,991 articles.

The sentiment distribution within the dataset is heavily skewed, with negative sentiments ('no') accounting for approximately 82.23% of the articles and positive sentiments ('yes') representing about 17.77%. This imbalance is indicative of the predominant narrative tone in the economic news articles and presents a potential challenge for machine learning models, which typically perform better with balanced datasets.

5.2 Text Preprocessing

Sentiment Mapping: We converted the 'relevance' field, which contains sentiment labels, into a binary format suitable for classification tasks. Positive sentiment labels ('yes') were mapped to 1, and negative sentiment labels ('no') were mapped to 0, simplifying the output variable for our predictive models.

Feature Selection: We narrowed down our features to two essential columns: 'text', which contains the news article, and 'relevance', our target variable representing the sentiment.

5.3 Text Cleaning

The integrity of our dataset heavily relies on the quality of the text data. To this end, we undertook a meticulous text cleaning process using spaCy, nltk, and the stop-words library to prepare our dataset for analysis. Each step in our cleaning process was designed to reduce noise and normalize the text for processing:

1. **Named Entity Removal:** We used spaCy to identify and remove named entities from the text. This step ensures that our analysis focuses on the linguistic structure and sentiment of the text, rather than specific entities which may skew the results.
2. **Case Normalization:** The text was converted to lowercase to maintain consistency and avoid duplication of tokens differing only in case.
3. **Tag Removal:** HTML break tags (</br>) were replaced with spaces to clean the text of extraneous HTML content.
4. **Hyphen Handling:** Hyphens were replaced with spaces to ensure compound words are treated as separate tokens.
5. **Punctuation and Digit Removal:** All punctuation and digits were removed to focus on textual data.
6. **Stopword Removal:** We compiled a comprehensive list of stopwords using both spaCy and the stop-words library. Tokens identified as stopwords were excluded from the text to remove commonly used words that offer minimal informative value.
7. **Lemmatization:** The nltk library's WordNetLemmatizer was employed to reduce words to their base or dictionary form, ensuring that different inflections of a word are analyzed as a single item.

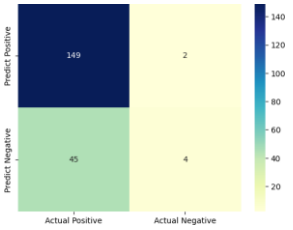

Figure 2: Confusion Matrix for Naive Bayes Classifier

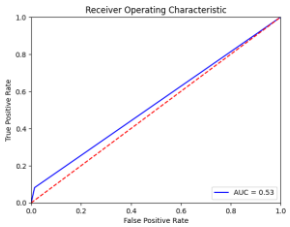

Figure 3: The ROC curve with an AUC of 0.53 indicates a performance close to the random chance for the Gaussian Naive Bayes classifier.

6 Model Training

6.1 Data Partitioning

The dataset was partitioned into training and test sets to both train our models and evaluate their performance on unseen data. We used an 80-20 split, allocating 80% of the data for training and the remaining 20 % for testing. This split was stratified based on the sentiment labels to maintain the distribution of sentiments across both sets, which is crucial given the imbalanced nature of our dataset. A fixed random seed was used to ensure the reproducibility of our results

6.2 Feature Extraction

We employed Tfidf Vectorizer from the sklearn feature extraction text module to convert the text documents into a matrix of TF-IDF features. The TF-IDF model was restricted to the top 20,000 features, selected based on term frequency across the corpus, to reduce dimensionality and improve computational efficiency.

7 Results and Conclusion

7.1 Naive Bayes Classifier

: Following feature extraction using TF-IDF, we proceeded to train a Gaussian Naive Bayes classifier — a probabilistic model well-suited for high-dimensional data. Given the dichotomy in our target variable, Gaussian Naive Bayes serves as an appropriate choice, assuming that the features follow a normal distribution. The performance of the Gaussian Naive Bayes classifier was evaluated on both the training and test datasets. It achieved a high accuracy score of 99.75% on the training data, while the accuracy on the test data was 76.5%.

- Training Accuracy : 99.75%
- Testing Accuracy : 76.59%
- Precision (Not Relevant) :0.77
- Recall (Not Relevant) :0.99
- F1-Score (Not Relevant): 0.86
- Precision (Relevant): 0.67
- Recall (Relevant): 0.08
- F1-Score (Relevant): 0.15
- AUC:0.53

To assess the model's discriminative ability, we plotted the Receiver Operating Characteristic (ROC) curve and computed the Area Under the Curve (AUC):

The AUC of 0.53 suggests that the classifier does not discriminate between the 'relevant' and 'not relevant' classes better than random chance.

The matrix shows a high number of true positives but also reveals that the classifier is biased towards predicting the majority class, as indicated by the low number of true negatives.

7.2 Multinomial Naive Bayes Classifier

: With its rapid training time, the Multinomial Naive Bayes classifier facilitates fast iterations and evaluations. The Multinomial Naive Bayes classifier exhibited a slight decrease in training accuracy compared to the Gaussian model but maintained similar test accuracy. Notably, the model achieved a test accuracy of 75.5%, demonstrating consistency with the Gaussian Naive Bayes model. However, the classification report indicated that while the model predicted the 'not relevant' class with high accuracy, it failed to identify any 'relevant' articles, evidenced by a recall and precision of 0.00 for the 'relevant' class.

- Training Accuracy : 75.75%
- Testing Accuracy : 75.5%
- Precision (Not Relevant) :0.76

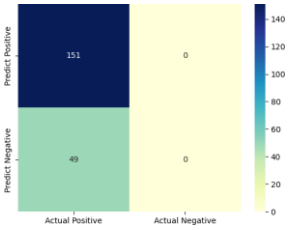

Figure 4: Confusion Matrix for Multinomial Naive Bayes Classifier

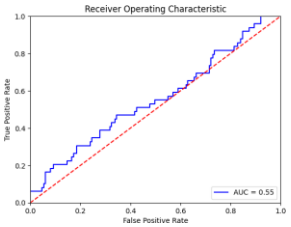

Figure 5: The ROC curve for the Multinomial Naive Bayes classifier.

- Recall (Not Relevant) :1.00
- F1-Score (Not Relevant): 0.86
- Precision (Relevant): 0.00
- Recall (Relevant): 0.00
- F1-Score (Relevant): 0.00

The confusion matrix further visualized the model's performance, confirming its bias towards the majority class and its failure to correctly classify any of the minority class. Confusion matrix for the Multinomial Naive Bayes classifier showing a high true positive rate for the majority class but no correct predictions for the minority class.

The ROC curve for the Multinomial Naive Bayes classifier, while slightly better than the Gaussian model, still indicated room for improvement with an AUC of 0.55.

7.3 Logistic Regression Classifier

: The Logistic Regression classifier achieved a training accuracy of 81.13% and a testing accuracy of 76.5%. While the training accuracy is high, the drop in testing accuracy indicates that the model may be overfitting to the training data but still maintains a relatively high level of generalization when applied to the test data. The model demonstrates high precision (0.77) and recall (0.99) for the majority 'not relevant' class. This suggests that the

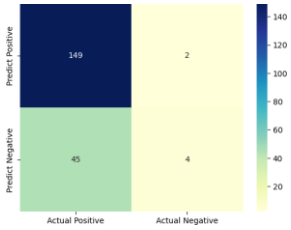

Figure 6: Confusion Matrix for Logistic Regression

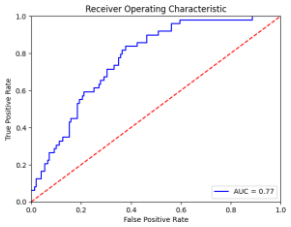

Figure 7: The ROC curve for the Logistic Regression classifier

model is quite effective at identifying true negatives (articles without relevant sentiment), with a very low rate of false negatives.

- Training Accuracy : 81.13%
- Testing Accuracy : 76.5%
- Precision (Not Relevant) :0.77
- Recall (Not Relevant) :0.99
- F1-Score (Not Relevant): 0.86
- Precision (Relevant): 0.67
- Recall (Relevant): 0.08
- F1-Score (Relevant): 0.15
- AUC:0.77

The precision for the 'relevant' class is moderate at 0.67, indicating that when the model predicts an article to be relevant, it is correct about two-thirds of the time. However, the recall for the 'relevant' class is very low at 0.08, indicating that the model only identifies 8percent of the actual relevant articles. This is a significant issue as it means that the model misses 92percent of the relevant articles, which could be critical in a real-world application. The F1-score for the 'not relevant' class is high at 0.86, indicating a good balance between precision and recall.

The Area Under the Curve (AUC) for the ROC is 0.77, suggesting that the Logistic Regression classifier has a fair discriminative ability between the

Figure 8: Confusion Matrix for Support Vector Machines

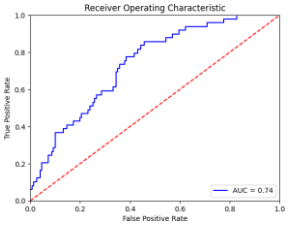

Figure 9: The ROC curve for SVM

positive and negative classes. An AUC score closer to 1.0 would be ideal, but 0.77 indicates a reasonable performance, especially when compared to a random classifier, which would have an AUC of 0.50.

7.4 Support Vector Machines

: The SVM is known for its effectiveness in high-dimensional spaces and its versatility in handling both linear and non-linear data. In our study, the SVM classifier was applied to the sentiment classification task within the economic news dataset. The SVM model demonstrated perfect accuracy on the training data; however, this did not translate to the test data, where the accuracy dropped to 74percent. On the test set, the SVM's precision, recall, and F1 scores for the 'not relevant' class were strong, indicating good model performance for the majority class.However, for the 'relevant' class, the scores were significantly lower, reflecting the model's struggle with the imbalanced dataset. The recall score of 0.41 for the 'relevant' class implies that the SVM was only able to correctly identify 41percent of the actual positive cases.

- Training Accuracy : 100%
- Testing Accuracy : 74%
- Precision (Not Relevant) :0.82
- Recall (Not Relevant) :0.85
- F1-Score (Not Relevant): 0.83
- Precision (Relevant): 0.47
- Recall (Relevant): 0.41
- F1-Score (Relevant): 0.43
- AUC:0.74

The confusion matrix for the SVM classifier showed a substantial number of true positives and true negatives, yet there were considerable false negatives, highlighting the model's challenge with the minority class.

The ROC curve's AUC of 0.74 is reasonably good, signifying that the SVM classifier has a fair discriminative ability to distinguish between the positive and negative classes.

7.5 Decision Tree Classifier

: The Decision Tree Classifier achieved perfect training accuracy, indicating the model was able to fully capture the training data's patterns. However, similar to previous models, a notable drop in accuracy to 71percent on the test data suggests overfitting. On the test set, the Decision Tree Classifier displayed balanced precision and recall for both the 'not relevant' and 'relevant' classes, with a score of approximately 0.41 for the latter. This balance indicates a more even-handed approach compared to the Naive Bayes and Logistic Regression models, which were skewed towards the majority class.

- Training Accuracy : 100%
- Testing Accuracy : 71%
- Precision (Not Relevant) :0.81
- Recall (Not Relevant) :0.81
- F1-Score (Not Relevant): 0.81
- Precision (Relevant): 0.41
- Recall (Relevant): 0.41
- F1-Score (Relevant): 0.41

The confusion matrix revealed a more equitable distribution of predictions across the four quadrants for the Decision Tree model, although false negatives and false positives remain a concern, reflecting challenges in correctly classifying both positive and negative sentiments.

Figure 10: Enter Caption

7.6 Ensemble

: Our ensemble method mirrored the high training accuracy seen in other models, achieving a perfect score. On the testing set, the ensemble achieved an accuracy of 77%, which is on the higher end of the spectrum compared to the individual classifiers previously discussed.

- Training Accuracy : 100%
- Testing Accuracy : 77%

The improved test accuracy suggests that the ensemble method was able to leverage the strengths of individual models while mitigating their weaknesses. This is a key outcome, as ensemble methods are designed to increase predictive performance by integrating diverse perspectives from various models, thereby reducing the risk of overfitting to the idiosyncrasies of the training data.

8 Future Scope

The current study on sentiment analysis of US economic news using NLP techniques opens several avenues for future research.

- **Model Refinement:** Exploring more complex models such as deep learning architectures—including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)—may offer improved accuracy, especially when combined with word embeddings that capture semantic relationships between terms.
- **Ensemble Techniques:** The potential benefits of ensemble methods were touched upon in this study, suggesting that combining the pre- dictions of multiple models can enhance per- formance. Future research should delve into various ensemble techniques, such as stack- ing, boosting, and bagging, to determine their effectiveness in this domain.
- **Transfer Learning:** Leveraging transfer learning, where a model is pre-trained on a large dataset and fine-tuned on a specific task, could also prove beneficial. Models such as BERT and GPT-3, which have been pre-trained on vast corpora, may bring substantial improvements when fine-tuned for economic sentiment analysis.
- **Real-Time Analysis**: Given the fast-paced nature of economic news, developing systems capable of real-time analysis and prediction could provide significant value. Investigating the feasibility and performance of models in a streaming data context would be a valuable contribution.
- **Interdisciplinary Applications:** Applying sentiment analysis tools to interdisciplinary studies, such as the correlation between economic news sentiment and stock market movements, could provide practical insights into the predictive power of news sentiment on economic indicators.