DESIGN OF BUSINESS INTELLIGENCE ON GEOSPATIAL DATA USING DEEP LEARNING

Canja, Tricha Maie
Villanueva, Iris

Overview

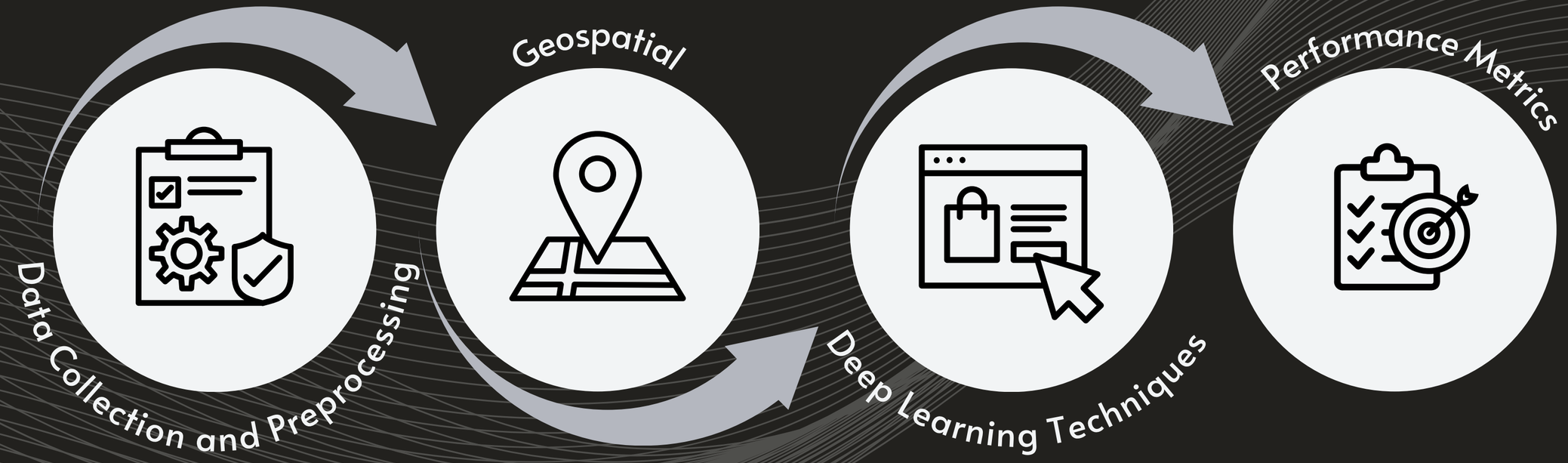
01 INTRODUCTION
02 METHODOLOGY
03 RESULTS AND DISCUSSION
04 CONCLUSION AND FUTURE WORKS

Introduction

One of the main causes of business failure is ignoring customer needs; this kind of mistake is responsible for 14% of small business failures. Understanding and meeting the needs of your target market is essential for growing a successful business since it directly affects consumer engagement and satisfaction. Moreover, poor geographic expansion may be the reason behind 7% of business failures. Before entering new markets, a thorough market analysis is necessary to determine whether the product or service will satisfy local needs and preferences

Methodology

The methodology used in this research was to evaluate the business opportunity in the vicinity through integrating deep learning algorithms with geospatial analysis.



DATA COLLECTION AND PREPROCESSING

Key Variables	Data Types	Example
business_id	string	
name	string	Turning Point of North Wales
latitude	integer	40.210196
longitude	integer	-75.223639
stars	integer	3
date	integer	2018-07-07 22:09:11
text	string	“If you decide to eat here, just be aware it is...”
cleanText	string	“decide eat aware going take hours beginning en... ”
sentiment	string	positive

Table I. Summary of Dataset Characteristics

Yelp Open Dataset

It is a general-purpose dataset used in academic studies to study natural language processing.

The Yelp dataset contains a file named business.json which includes user reviews, star ratings, business details, latitude and longitude values of the area, and user information.

The key focus was on the textual data from customer reviews and in this dataset, it is the ‘cleanText’ feature, which was used to determine the sentiment.

GEOSPATIAL DATA

Figure I. Sample Geospatial Analysis

Figure 1 shows the sample location of a center point where a business might be built or where a company locates and this can be used in collecting the competitors strength and weaknesses using the reviews with a similar type of business for the purpose of analyzing the market and competitors and can be used to assess the nearby competitors.

DEEP LEARNING TECHNIQUES

Sentiment Analysis

- Convolutional Neural Network (CNN)
- Long Short-Term Memory (LSTM)
- Bidirectional Long Short-Term Memory (BiLSTM)

DEEP LEARNING TECHNIQUES

Topic Modeling

- Latent Dirichlet Allocation (LDA)
- Non Negative Matrix Factorization (NMF)
- Latent Semantic Analysis (LSA)

PERFORMANCE METRICS

Sentiment Analysis: Accuracy and Confusion Matrix

Accuracy provides an overall assessment of how well the deep learning models are performing in classifying sentiments. Higher accuracy values signify better performance in correctly predicting sentiment labels for the reviews.

Confusion matrix provides insights into how well the models classify sentiment categories (positive, negative, neutral) based on the input text data. Each row of the matrix represents the actual sentiment labels, while each column represents the predicted sentiment labels.

PERFORMANCE METRICS

Topic Modeling: Coherence

Coherence score is a metric that assesses how human-interpretable the topics are. This is shown as the top n terms that are most likely to be related to that specific topic.

Results and Discussion

Models	Accuracy	Positive Precision	Negative Precision	Neutral Precision
CNN	91.19%	95.52%	**85.86%**	70.19%
LSTM	**92.53%**	**95.88%**	85.31%	**76.27%**
BiLSTM	91.92%	95.38%	79.03%	76.20%

Table II. Sentiment Analysis Results

Results and Discussion

Overall, the Long Short-Term Memory (LSTM) model emerged as the most balanced and accurate, achieving the highest accuracy of 92.53%. It also performed exceptionally well in predicting positive sentiments with a precision of 95.88% and neutral sentiments with a precision of 76.27%.

Results and Discussion

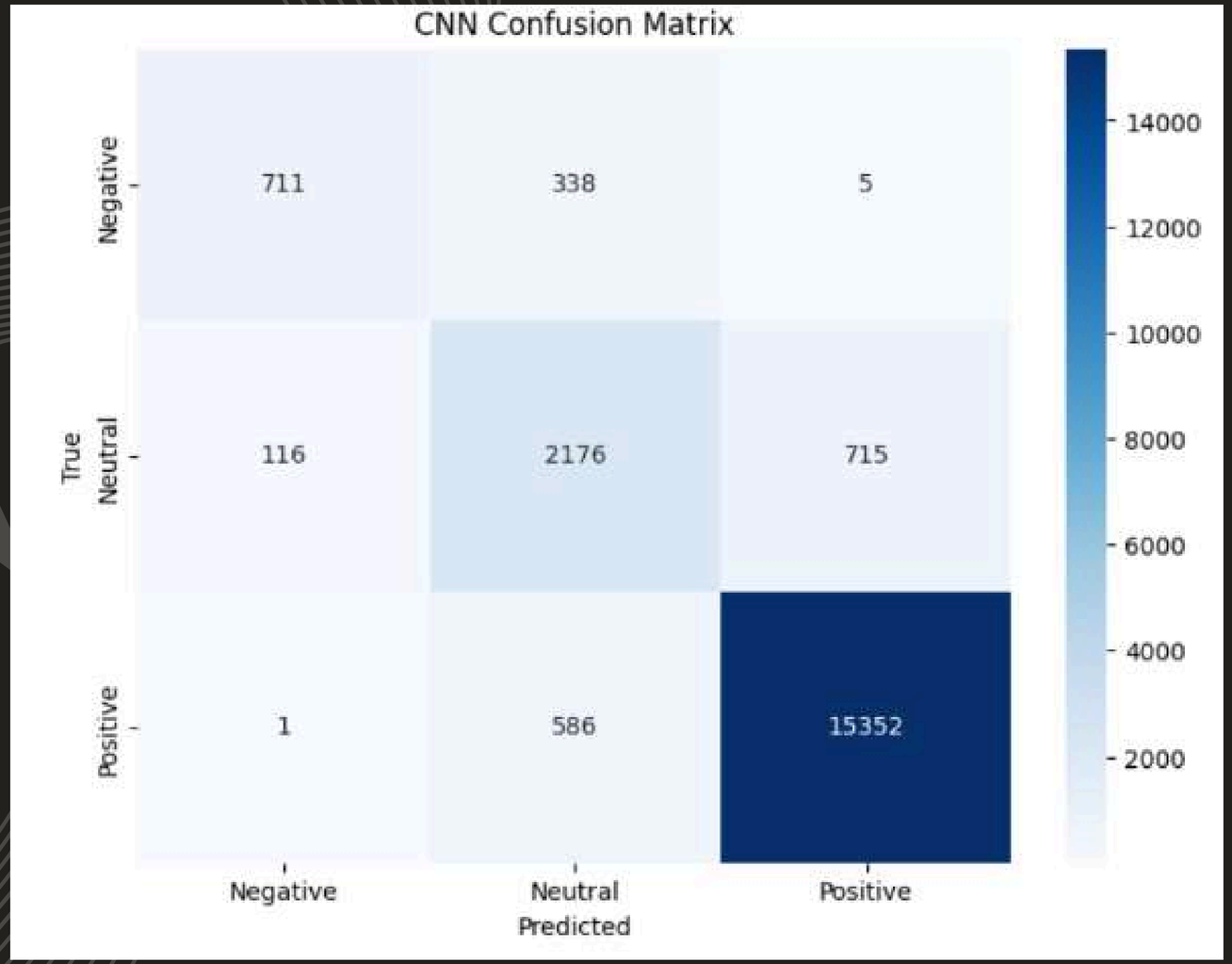


Figure II. CNN Confusion Matrix

It is evident that the CNN model performed well in predicting neutral sentiments, with a majority correctly classified. However, it struggled more with negative and positive sentiments, as indicated by the higher number of misclassifications in those categories.

Results and Discussion

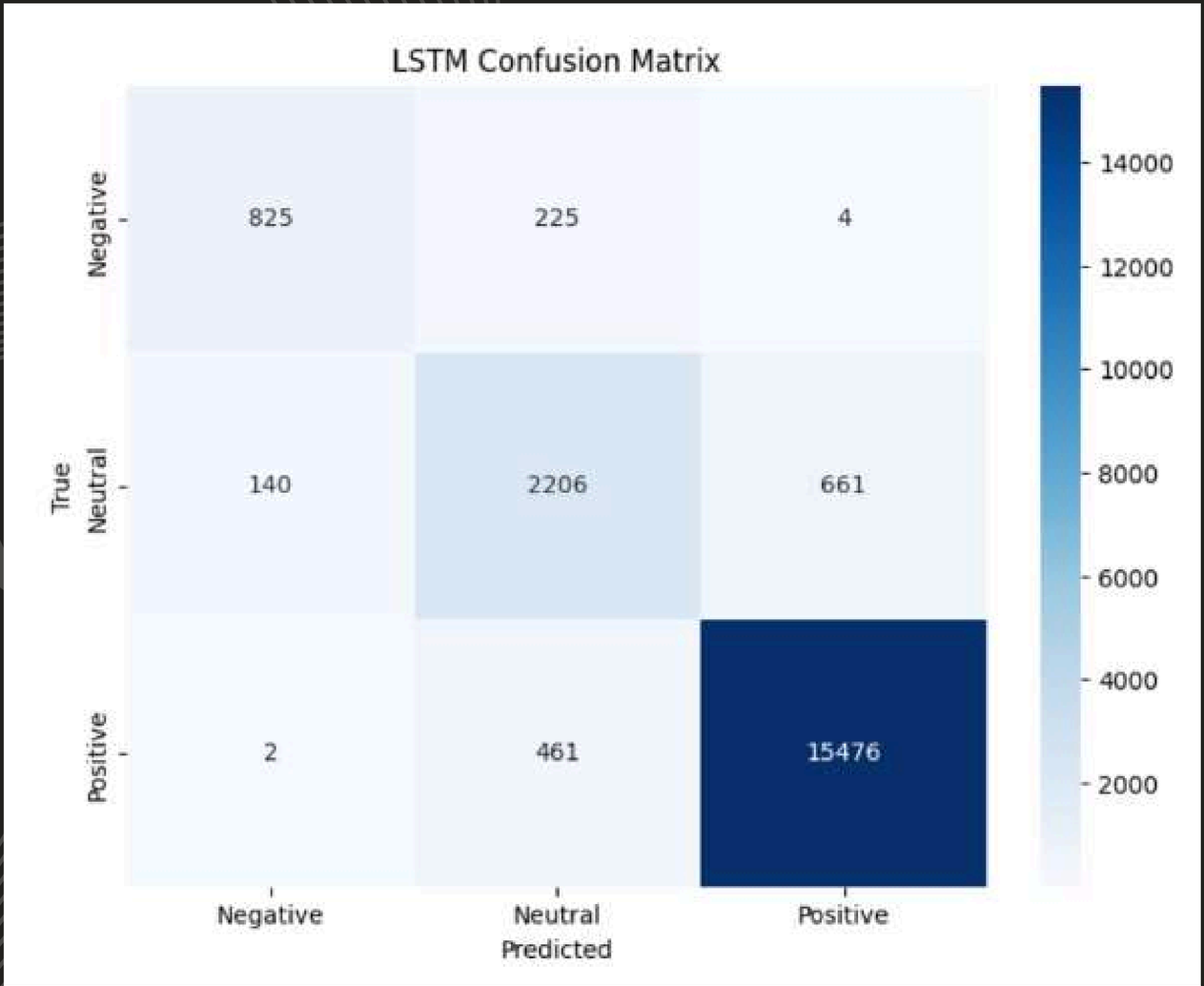


Figure III. LSTM Confusion Matrix

The LSTM model exhibits proficiency in predicting neutral sentiments, with a substantial number of correct classifications. However, it also faces challenges in accurately classifying negative and positive sentiments, particularly the latter, where misclassifications are more prevalent.

Results and Discussion

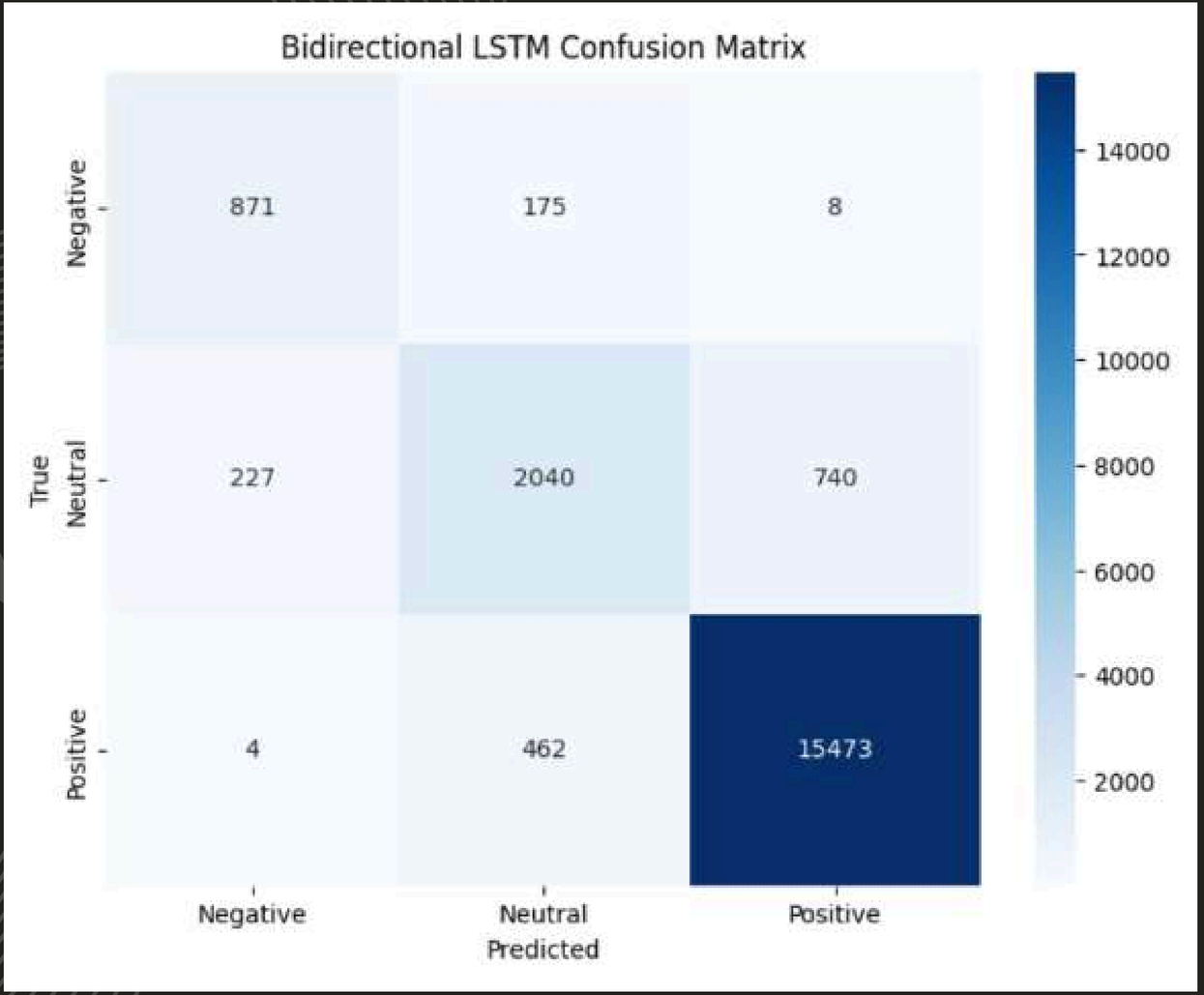


Figure IV. BiLSTM Confusion Matrix

It excels in predicting neutral sentiments, with a significant portion correctly classified. Nonetheless, it encounters difficulties in distinguishing between negative and positive sentiments, as shown by the higher number of misclassifications in those categories.

Results and Discussion

Models	Coherence Score	Number of Topics
Latent Dirichlet Allocation (LDA)	0.60	46
Latent Semantic Analysis (LSA)	0.38	20

Table III. Coherence Score of LDA and LSA

Results and Discussion

Table 3 shows the coherence score of LDA and LSA. Only these two models use the coherence score. The coherence score that achieved the LDA is 0.60, and 46 topics have been identified, while the LSA obtained 20 topics and achieved a coherence score of 0.38.

In terms of these metrics, the higher coherence score indicated that the topics are more interpretable and meaningful. In this case, the LDA model shows a good sign that the model learned more meaningful topics from the data than the LSA.

Results and Discussion

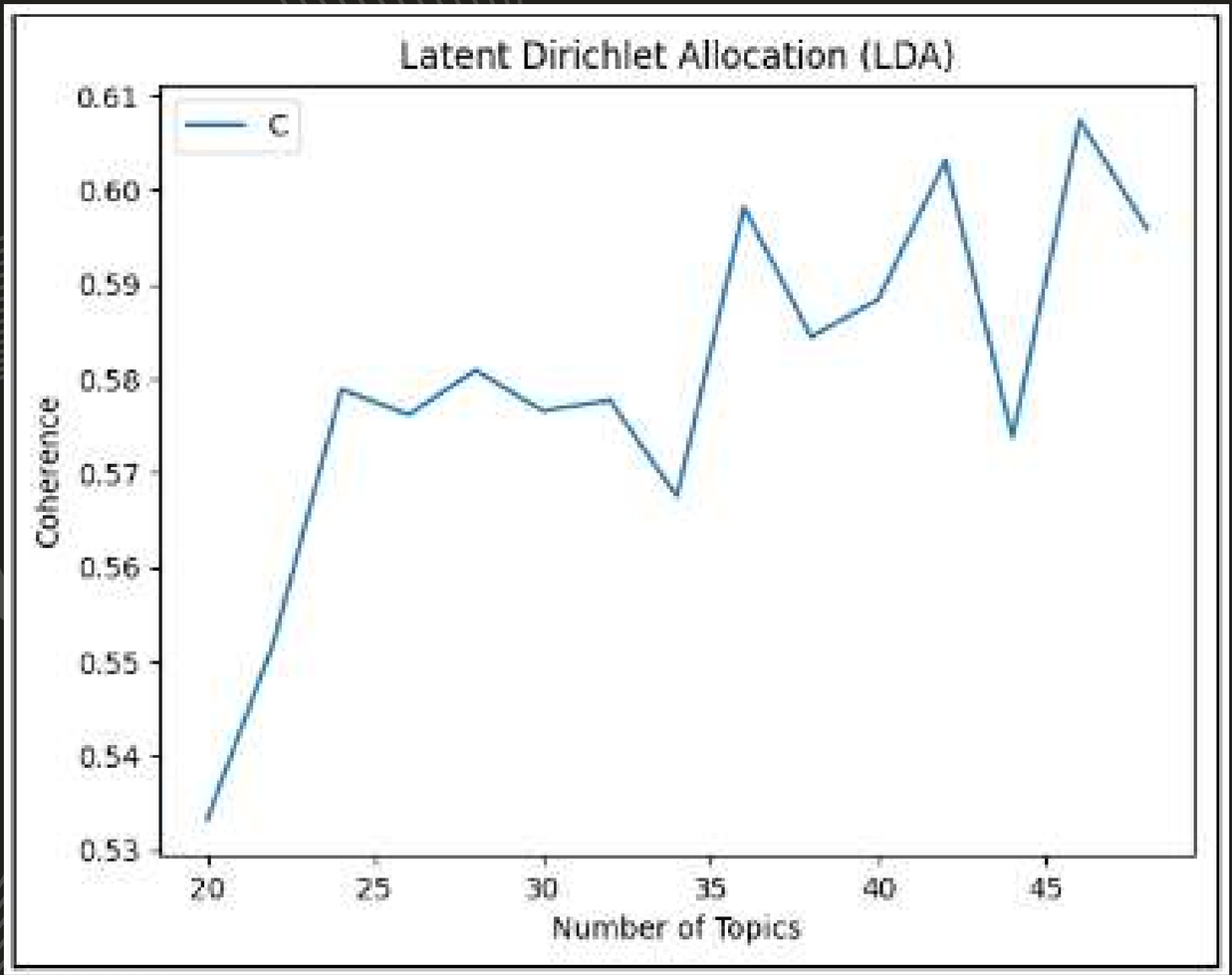


Figure V. LDA Coherence over Number of Topics

Figure 5 shows the coherence score of latent Dirichlet allocation for the different number of topics. In the graph, the coherence score starts at a low and then increases as the number of topics increases. This indicates that the more topics, the more interpretable and meaningful they are.

Results and Discussion

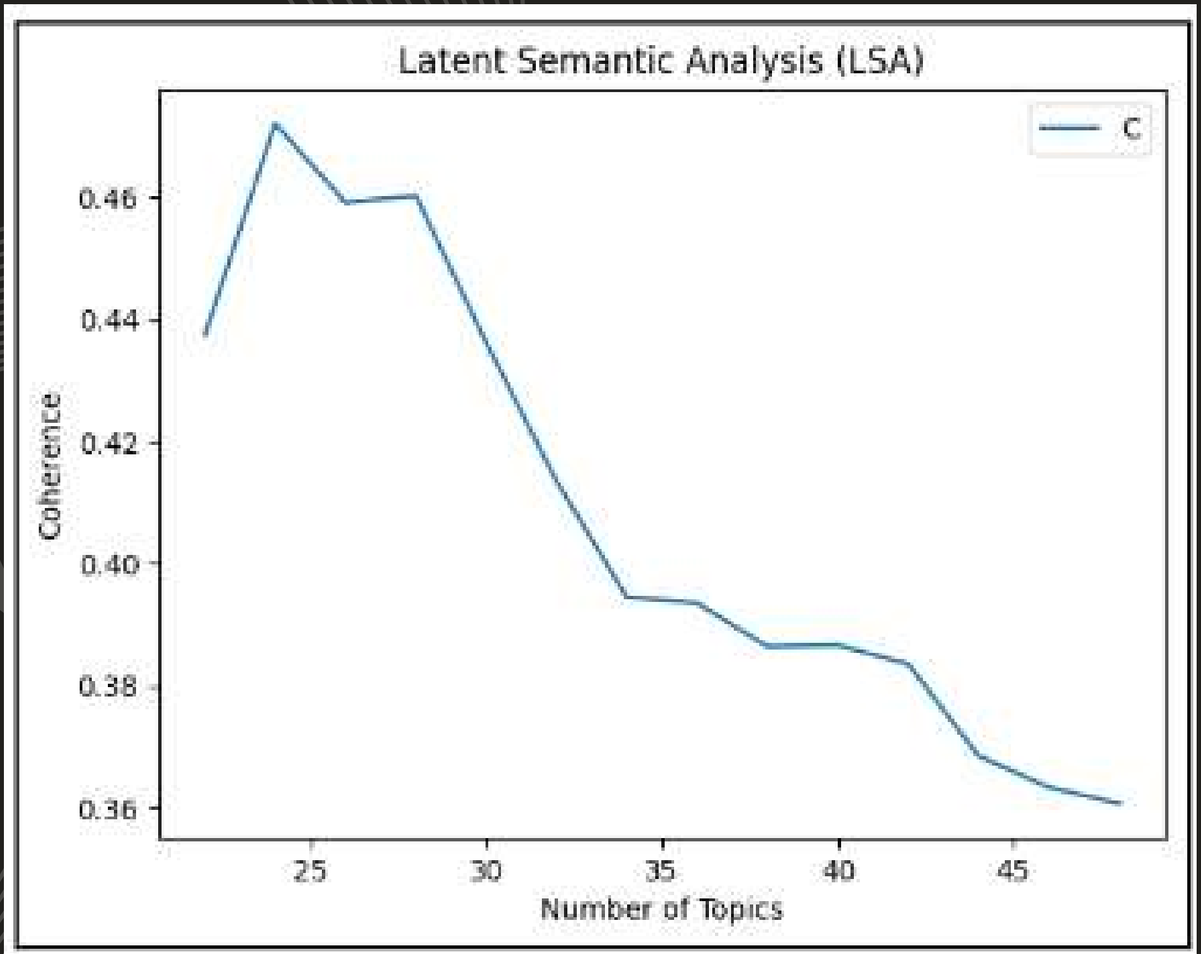


Figure VI. LSA Coherence over Number of Topics

Figure 6 shows the coherence score of the LSA model, which starts relatively low and then increases slightly near the 20 topics, and then after that peak, the coherence score starts to decrease over the number of topics.

Conclusion and Future Works

Sentiment Analysis

- Results showed high accuracy across all models, with LSTM slightly outperforming the others in overall sentiment classification.
- Precision scores for each sentiment category highlight the models' effectiveness in distinguishing between positive, negative, and neutral sentiments.

Conclusion and Future Works

Topic Modeling.

- LDA achieved a coherence score of 0.60 and identified 46 topics.
- LSA achieved a coherence score of 0.38 and identified 20 topics.
- The higher coherence score of the LDA model indicates it learned more interpretable and meaningful topics compared to the LSA model.

Conclusion and Future Works

Overall, this research contributes to the field of business opportunity analysis by demonstrating how advanced data science techniques can be employed to analyze complex datasets. The findings underscore the potential of combining geospatial analysis, sentiment analysis, and topic modeling to provide a holistic view of business opportunities, ultimately aiding in the economic growth and development of specific regions. Future work may involve expanding the dataset to include more diverse regions and refining the models to improve their predictive capabilities further.