Design of Business Intelligence on Geospatial Data Using Deep Learning

1st Tricha Maie Canja
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City
qtmdacanja@tip.edu.ph

2nd Iris Villanueva
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
qilvillanueva@tip.edu.ph

3rd Roman Richard
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
rrichard.cpe@tip.edu.ph

***Abstract*—This paper presents a comprehensive business opportunity analysis using geospatial data and deep learning algorithms. Geospatial data presents valuable opportunities for identifying and leveraging emerging business opportunities. However, traditional analysis methods often struggle to uncover hidden patterns and insights within large, complex geospatial datasets. This paper proposes a novel approach using deep learning models to predict business success and in the Philippines. The first model employs sentiment analysis on user reviews to categorize feedback into positive, negative, and neutral sentiments using various neural network architectures such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM). This analysis helps in understanding customer perceptions and satisfaction levels across different business categories. The second model involves topic modeling techniques, including Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), to uncover underlying themes within the reviews. This integrated approach enables entrepreneurs and investors to make data-driven decisions, reducing risks associated with poor location choices and mismatched business ventures. By harnessing the power of deep learning, this paper aims to provide a comprehensive framework for business opportunity analysis, contributing to more informed decision-making and economic growth in the Philippines.**

Keywords—geospatial data, deep learning, NLP, sentiment analysis, geographic information systems, economic feasibility, business opportunity analysis

I. Introduction

Starting a business can be quite difficult. It combines calculated decision-making, strategic planning, and taking calculated risks [1]. While many businesses must begin with a business plan which describes how to achieve the goal and contains important details about the business itself, market analysis—which gathers data about the factors and circumstances that might affect your sector—and competitive analysis—which helps entrepreneurs assess the advantages and disadvantages of their competitors—are also important [2].

Entrepreneurs may take into account different factors when establishing certain types of enterprise. The location of the business and the customer reviews are two of these elements that may support the growth and success of the business. These factors may help in decision-making, potential for expansion, product or services improvements and understanding the complexity of local marketplaces [3,4].

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 [5]. Before entering new markets, a thorough market analysis is necessary to determine whether the product or service will satisfy local needs and preferences [6].

Despite the widespread acceptance of the importance of strategic studies and decision-making in identifying and taking advantage of business opportunities. The huge amount of data that combines information on daily and business activities makes it more challenging to carry out effective decision-making [7].

Business intelligence combines data-driven decision-making with an overview of the data about the business. It places a particular focus on managing the huge quantity of data related to how the business is operated. In addition, it contains a competitive and market analysis, which is quite beneficial for business growth and development [8].

Nowadays, the use of geospatial data holds significant potential for the expansion of businesses in identifying and taking advantage of growing market opportunities [9]. Geospatial data is information that defines objects based on how they are located. It usually combines attribute, location, and temporal information to support location-based business strategies [10].

Geospatial analysis is used to give information about the distance between others, which helps in business making decisions. Businesses may gather and evaluate the historical and current business data using geospatial data to determine the strengths, weaknesses, and market drivers. The company may use this information to improve and strengthen its offerings in terms of both products and services [11].

The integration of geospatial and business intelligence can do a lot to help make a critical decision. It gives the company the ability to assess, contrast, and analyze data from competitors. This kind of integration may contribute a lot more to the future of the business.

The aim of this study is to develop a market and competitive analysis for a specific field by leveraging deep learning techniques. By incorporating sentiment analysis and topic modeling from reviews, this study intends to deeply evaluate competitors' strengths and weaknesses. The findings generated will provide valuable insights to support strategic decision-making in business and economic planning.

II. Methodology

The methodology used in this paper was to evaluate the business opportunity in the vicinity through the data gathered and by integrating the deep learning algorithms with geospatial analysis.

A. Data Collection and Preprocessing

The dataset that was used in this paper came from the Yelp Open Dataset [12], a general-purpose dataset used in academic studies to study natural language processing. Table 1 provides an overview of the dataset's attributes, including the primary variables, data types, and example values. The dataset has 9 features and 10,000 records. The dataset was preprocessed to fix missing values and normalize numerical features before analysis. The information is a crucial instrument for competitive and market analysis, as well as for guiding strategic marketing initiatives aimed at establishing a business in a specific location.

TABLE I. Summary of Dataset Characteristics

Key variable	Data Types	Example
business_id	string	7ATYjTIgM3jUlt4UM3IypQ
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	integer	“If you decide to eat here, just be aware it is…”
cleanText	string	“decide eat aware going take hours beginning en… ”
sentiment	string	positive

The Yelp dataset contains a file named business.json and review.json, which houses a plethora of information crucial for business analysis. This 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.

a. Sentiment Labeling

The sentiment of each review was categorized into three classes: positive, negative, and neutral. This was achieved using the TextBlob library, a natural language processing (NLP) tool that provides a simple API for common NLP tasks [13]. TextBlob analyzes the polarity of the text, which is a measure of the positivity or negativity of the text.

i. Reviews with a polarity greater than 0 were labeled as 'positive'.
ii. Reviews with a polarity less than 0 were labeled as 'negative'.
iii. Reviews with a polarity equal to 0 were labeled as 'neutral'.

b. Text Tokenization and Padding

The reviews were tokenized and converted into sequences of integers using the Keras Tokenizer. These sequences were then padded to ensure uniform input length for the deep learning models.

B. Geospatial Data

Geospatial data plays a vital role in enhancing the decision-making process and gaining deeper insights into patterns and trends. Using this kind of data may leverage market opportunities, mitigate risks, and develop targeted strategies. The integration of geospatial and business intelligence as an important part of the data to extract customer insights and drive informed decision-making requires the integration of different techniques.

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.

C. Deep Learning Techniques

Deep learning is a subfield of machine learning and artificial intelligence that finds wide use across multiple fields, including text analytics, image recognition, healthcare, and many more. This study implements a different deep learning approach [14]. Using deep learning models, the performance measurements will be the foundation of the assessment.

a. Sentiment Analysis
   i. Convolutional Neural Network (CNN)
   ii. Long Short-Term Memory (LSTM)
   iii. Bidirectional Long Short-Term Memory (BiLSTM)
b. Topic Modeling
   i. Latent Dirichlet Allocation (LDA)
   ii. Latent Semantic Analysis (LSA)

D. Performance Metrics and Evaluation

To measure the performance of each deep learning algorithm, various metrics serve as a reference to assess how well each deep learning algorithm is performing in the tasks of sentiment analysis and topic modeling.

a. Accuracy

This metric 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.

b. Confusion Matrix

In sentiment analysis, it 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.

c. Coherence Score

In topic modeling, 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.

III. Result and Discussion

The sentiment analysis results from the Yelp dataset were evaluated using three deep learning models: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and

Bidirectional Long Short-Term Memory (BiLSTM). The evaluation metrics used to measure the performance of these models included accuracy and precision for positive, negative, and neutral sentiments. The following table summarizes the results:

TABLE II. SENTIMENT ANALYSIS RESULTS

Model	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%

The sentiment analysis results demonstrate that different models excel in different areas. 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%. The Convolutional Neural Network (CNN) also showed strong performance, particularly in predicting positive sentiments with a precision of 95.52% and negative sentiments with a precision of 85.86%. However, it struggled more with neutral sentiments, achieving a lower precision of 70.19%. The Bidirectional Long Short-Term Memory (BiLSTM) model, while slightly less accurate overall with an accuracy of 91.92%, performed similarly to LSTM in predicting neutral sentiments (76.20%) but lagged in negative sentiment prediction (79.03%). In summary, the LSTM model is the most reliable and balanced for sentiment analysis.

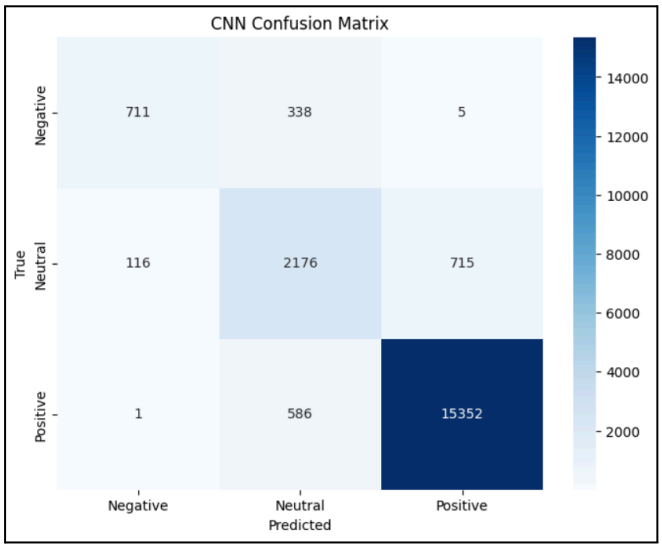

FIGURE II. CNN CONFUSION MATRIX

The confusion matrix for the Convolutional Neural Network (CNN) model reveals its performance across three sentiment classes: negative, neutral, and positive. From the plot, it's 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.

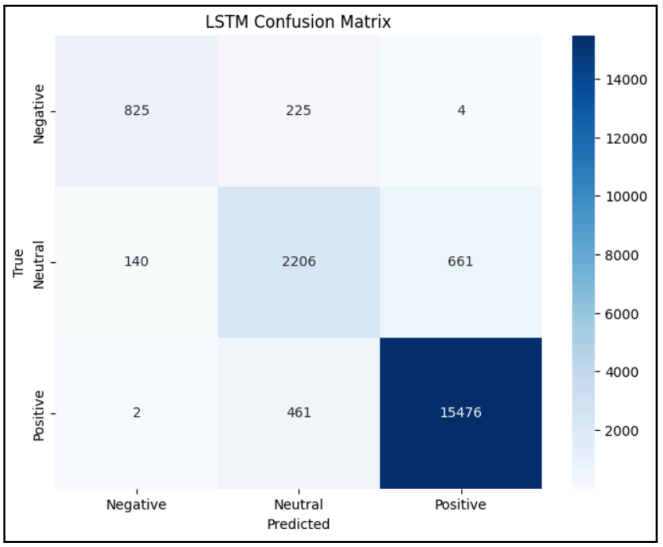

FIGURE III. LSTM CONFUSION MATRIX

Moving on with the LSTM model, it demonstrates a similar trend to the CNN model. 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.

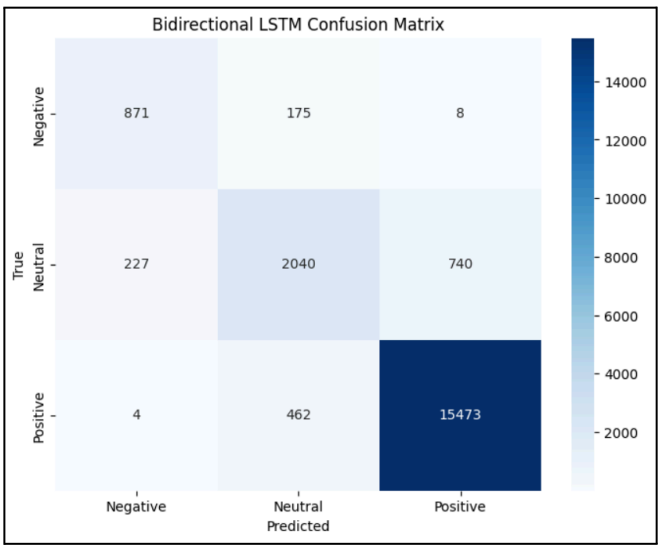

FIGURE IV. BILSTM CONFUSION MATRIX

Finally, the Bidirectional Long Short-Term Memory (BiLSTM) model showcases comparable performance to the CNN and LSTM models. 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.

Topic modeling can be a powerful technique for identifying trends that can be focused on for the business. This type of Natural Language Processing is helpful in the different aspects of the business. Combining the analysis allows the business to understand the specific aspects of positive and negative sentiment and use it as a reference to improve, strategize, and know the market in a specific area.

TABLE III. COHERENCE SCORE OF LDA AND LSI

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

Table 3 shows the coherence score of LDA and LSI. 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 LSI 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.

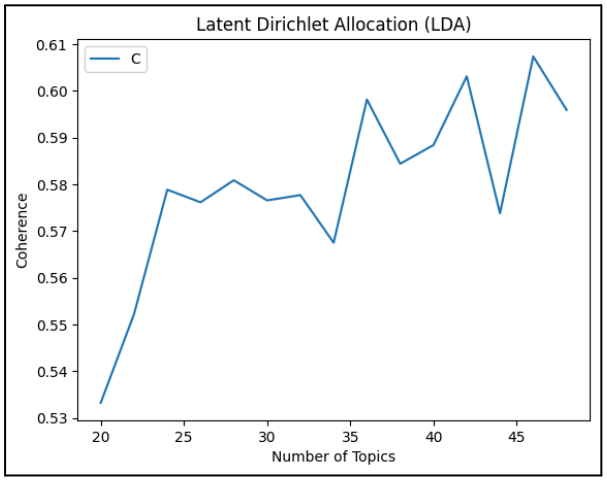

FIGURE V. LDA COHERENCE OVER NUMBER OF TOPICS

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.

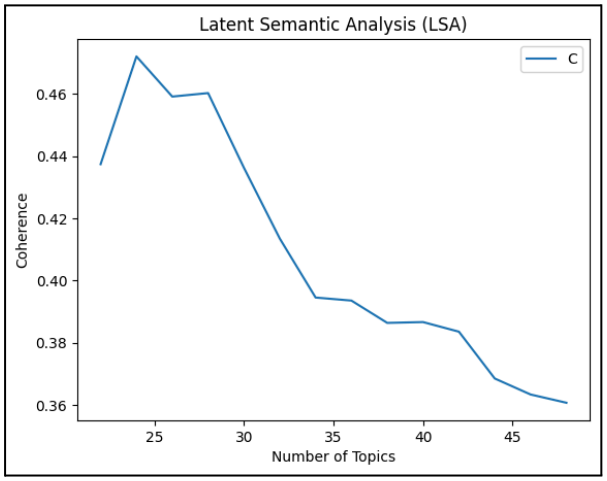

FIGURE VI. LSI COHERENCE OVER NUMBER OF TOPICS

Figure 6 shows the coherence score of the LSI 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.

IV. CONCLUSION AND FUTURE WORKS

This paper shows the effectiveness of using geospatial data, sentiment analysis, and topic modeling to identify and assess potential business opportunities. Creating a comprehensive framework made the study successfully integrate various data sources and analytical techniques to provide valuable insights into market trends, customer sentiments, and optimal business locations.

The sentiment analysis models, including CNN, LSTM, and BiLSTM were implemented to classify customer reviews into positive, negative, and neutral sentiments. The results showed that each model exhibits high accuracy, with LSTM slightly outperforming the others in overall sentiment classification. The precision scores for each sentiment category across different models highlight their ability to distinguish between positive, negative, and neutral sentiments effectively.

In addition to sentiment analysis, the integration of topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) provided deeper insights into the underlying themes within customer reviews. The coherence scores for these models indicate their effectiveness in learning meaningful topics from the data. Specifically, the LDA model achieved a coherence score of 0.60 and identified 46 topics, while the LSA model obtained a coherence score of 0.38 with 20 topics. The higher coherence score of the LDA model suggests that it learned more interpretable and meaningful topics compared to the LSA model.

The inclusion of geospatial data enhances the framework's ability to recommend optimal business locations, contributing to more strategic decision-making for entrepreneurs and investors.

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.

REFERENCES

[1] M. TALMAGE-ROSTRON, "Types & Importance of Risk Taking in Entrepreneurship 2024," www.nexford.edu, Mar. 09, 2024. [Online]. Available: https://www.nexford.edu/insights/risk-taking-in-entrepreneurship

[2] J. Lindzon, "The Importance of a Business Plan: 10 Reasons You Need a Road Map For Your Business," www.waveapps.com, Jan. 02, 2022. [Online]. Available: https://www.waveapps.com/blog/importance-of-a-business-plan

[3] T. Callen, "Gross Domestic Product: An Economy's All," International Monetary Fund, 2022. [Online]. Available: https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/gross-domestic-product-GDP#:~:text=GDP%20is%20important%20because%20it

[4] JLL, "Seven reasons why location is important," www.us.jll.com, 2023. [Online]. Available: https://www.us.jll.com/en/views/seven-reasons-location-important

[5] I. Mitic, "Small Business Failure Statistics to Know in 2020," Fortunly, Jul. 09, 2019. [Online]. Available: https://fortunly.com/statistics/small-business-failure-statistics/

[6] M. Williams, "Market Analysis for Your Business Plan," www.wolterskluwer.com, Aug. 30, 2022. [Online]. Available: https://www.wolterskluwer.com/en/expert-insights/market-analysis-for-your-business-plan

[7] B. Zohuri and M. Moghaddam, “From Business Intelligence to Artificial Intelligence,” Journal of Material Sciences & Manufacturing Research, vol. 1, no. 1, Feb. 2020, doi: https://doi.org/10.32474/MAMS.2020.02.000137.

[8] Handbook on decision support systems 2 : Variations. Berlin: Springer, 2008.

[9] I. Limited, “Geospatial Data Analysis to Aid Business | Infosys BPM,” www.infosysbpm.com. https://www.infosysbpm.com/blogs/geospatial-data-services/significance-of-geospatial-data-sercives.html (accessed May 16, 2024).

[10] IBM, “What Is Geospatial Data,” www.ibm.com, 2020. [Online]. Available: https://www.ibm.com/topics/geospatial-data

[11] matylda, “An introduction to geospatial analysis for business,” Spyrosoft, Apr. 21, 2022. https://spyro-soft.com/blog/geospatial/geospatial-analysis (accessed May 18, 2024).

[12] Yelp, “Yelp Dataset,” Yelp.com, 2019. https://www.yelp.com/dataset

[13] TextBlob, “Making Natural Language Processing easy with TextBlob.” (2021, October 9). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/making-natural-language-processing-easy-with-textblob/

[14] I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” SN Computer Science, vol. 2, no. 6, Aug. 2021, doi: https://doi.org/10.1007/s42979-021-00815-1.

GitHub Repository Link:
CPE 313 Final Project

Model Deployment Presentation:
🗀 Model Deployment