Lambton College – Mississauga

Unveiling Customer Sentiment:

Analyzing Yelp Reviews with NLP

David Higuera
Cecille Jatulan
Maria Melencio
Michael Montanez
Diana Reyes
Abhikumar Patel

2024W-T3 AML 2304 - Natural Language Processing 01
Bhavik Gandhi
April 11, 2024

Table of Contents

Introduction

Customer satisfaction is pivotal for business success in the service industry, driving efforts to improve service quality and retain loyal customers. This necessitates careful customer feedback analysis, particularly in today's digital era, where platforms like Yelp provide abundant data. Our project focuses on Natural Language Processing (NLP) and sentiment analysis within the Yelp dataset, specifically within the service industry.

Yelp is a valuable source of customer reviews across various businesses, offering insights that can inform strategic decisions and enhance service standards. Through preprocessing, sentiment analysis, and interpretation techniques, we aim to develop a model and user interface capable of categorizing reviews as positive or negative. We strive to provide actionable insights for businesses to improve service quality and customer satisfaction in a competitive market by utilizing advanced analytical methods and user-friendly interfaces.

Business Domain:

The business domain for this project is the service industry, encompassing businesses that offer both tangible and intangible products/services to consumers. This includes a variety of professions, such as auto services, contractors, electricians, HVAC technicians, and plumbers, as well as businesses like restaurants and bars.

Problem and Need:

Service industry businesses rely heavily on customer feedback to maintain and improve their reputation and customer satisfaction. However, analyzing large input volumes from platforms like Yelp poses a significant challenge. Manual review analysis is time-consuming and prone to errors, hindering businesses from effectively extracting actionable insights.

Existing sentiment analysis tools often need more specificity to interpret sentiments accurately within the diverse context of the service industry. The language and terminology used in reviews for different types of businesses vary greatly, making it difficult for generic sentiment analysis models to provide meaningful insights.

Custom/Differentiated Solution:

To address these challenges, a custom sentiment analysis solution explicitly tailored to the nuances of the service industry using the Yelp dataset is needed. This solution would leverage natural language processing techniques to accurately analyze sentiments expressed in customer reviews across various types of service-oriented businesses.

Key features of the differentiated solution include:

By creating a custom sentiment analysis solution using the Yelp dataset tailored to the service industry's unique needs and challenges, businesses can gain valuable insights to enhance customer satisfaction and improve overall service quality, ultimately driving success in a competitive market.

Our solution integrates a Naive Bayes classifier with a deep learning model based on Recurrent Neural Networks (RNNs) using Long Short-Term Memory (LSTM) units. Furthermore, we introduced an ensemble technique called blending to maximize predictive performance.

To create a robust sentiment analysis solution for the service industry leveraging the Yelp dataset, we devised a unique approach that combines the strengths of both traditional machine learning and deep learning models. Our solution integrates a Naive Bayes classifier with a deep learning model based on Recurrent Neural Networks (RNNs) using Long Short-Term Memory (LSTM) units. Furthermore, we introduced an ensemble technique called blending to maximize predictive performance.

Hybrid Model Integration:

We recognized the complementary nature of Naive Bayes and deep learning models in sentiment analysis. While Naive Bayes excels at capturing basic patterns and dependencies in the data, deep learning models like RNN-LSTM are adept at capturing intricate sequential dependencies and nuances in language. By integrating these models, we aimed to leverage their strengths for more accurate predictions.

Weighted Fusion:

Our approach assigned different weights to the predictions generated by the Naive Bayes and RNN-LSTM models based on their performance on validation data. This weighted fusion technique ensured that the final prediction benefited more from the model with superior performance on the given dataset.

By adopting this differentiated solution, we aimed to create a sentiment analysis framework explicitly tailored to the service industry's needs, leveraging the richness of the Yelp dataset. Through the strategic integration of Naive Bayes, RNN-LSTM, and blending techniques, our solution offers businesses in the service industry a powerful tool to extract actionable insights from customer feedback, ultimately driving improvements in service quality and customer satisfaction.

Background and Literature Review

Sentiment analysis, or opinion mining, is a burgeoning field within Natural Language Processing (NLP) that focuses on extracting subjective information from textual data. In recent years, sentiment analysis has garnered significant attention due to its wide-ranging applications in various domains, including customer feedback analysis, social media monitoring, and market research.

In the service industry, where customer satisfaction is not just a goal but a necessity, sentiment analysis plays a pivotal role in understanding customer perceptions and improving service quality. As a prominent platform for consumer reviews, Yelp is a valuable data source for sentiment analysis in the service industry. However, analyzing the vast volume of reviews on Yelp presents challenges, including the need for accurate sentiment classification and interpretation.

Literature Review:

Numerous studies have explored sentiment analysis techniques and methodologies, ranging from traditional machine learning algorithms to advanced deep learning models. Naive Bayes classifiers, a simple yet effective probabilistic model, have been widely used for sentiment analysis tasks due to their simplicity and efficiency, particularly in text classification tasks.

On the other hand, Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units have emerged as powerful tools for sequence modeling and text analysis. They are capable of capturing intricate dependencies and contextual information in sequential data. Applying RNN-LSTM models in sentiment analysis has shown promising results, especially in capturing the nuanced nature of language and sentiment expression.

Ensemble methods, such as blending, have gained popularity for improving prediction accuracy and robustness by combining multiple models' predictions. Blending involves aggregating predictions from diverse models, each trained with different configurations or algorithms, to obtain a consensus prediction that often outperforms individual models.

While previous studies have explored the effectiveness of individual sentiment analysis techniques, limited research has focused on combining Naive Bayes classifiers and RNN-LSTM models for sentiment analysis, particularly in the context of the service industry and Yelp dataset. By integrating these diverse methodologies and leveraging ensemble blending techniques, this study aims to fill this gap and provide a comprehensive sentiment analysis solution that holds immense potential to enhance the understanding of customer sentiment in the service industry.

Overall, the literature underscores the importance of sentiment analysis in understanding customer sentiment and improving service quality while highlighting the potential of hybrid models and ensemble techniques to enhance prediction accuracy and robustness in sentiment analysis tasks. This study uses the Yelp dataset to build upon existing research and contribute to advancing sentiment analysis methodologies in the service industry.

Methodology

Dataset Collection and Preprocessing:

The Yelp reviews dataset is collected and preprocessed to ensure its suitability for sentiment analysis. The preprocessing steps include removing punctuation and stop words and performing tokenization, lemmatization, negation handling, contraction handling, and normalization. These steps aim to clean the text data and standardize its format for further analysis.

Model Architecture:

The sentiment analysis model architecture is a unique blend of two distinct approaches: the Naive Bayes classifier and a Long Short-Term Memory (LSTM) neural network. The Naive Bayes classifier leverages TF-IDF (Term Frequency-Inverse Document Frequency) features extracted from the preprocessed text data, while the LSTM network excels in processing the sequential nature of text data, capturing intricate dependencies and contextual information.

Training and Optimization:

The machine learning model, comprising the Naive Bayes classifier and LSTM network, is trained on the preprocessed text data and optimized using appropriate techniques. Hyperparameter tuning, cross-validation, and grid search methods are employed to optimize model performance and prevent overfitting. The training process aims to maximize the model's ability to classify Yelp reviews into positive or negative sentiments accurately.

Evaluation Metrics:

The performance of the trained sentiment analysis model is evaluated using a range of evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify reviews and its effectiveness in capturing sentiment polarity. The evaluation process ensures the robustness and reliability of the sentiment analysis model before deployment.

Model Deployment and Integration:

Once trained and evaluated, the sentiment analysis model is deployed for practical use. The deployment process involves integrating the model into a user-friendly interface using Streamlit. Additionally, the model is deployed on Render.com to enable its integration into a website for real-time sentiment analysis of new Yelp reviews. This deployment and integration ensure businesses can easily access and utilize the sentiment analysis model to enhance customer satisfaction and service quality.

MECE Table

Category	Details
Dataset Collection	Cleaning and preprocessing the Yelp reviews dataset, including punctuation removal and tokenization. Leverages techniques such as lemmatization, negation handling, contraction handling, and normalization.
Feature Extraction	Extracting features from the preprocessed text data using TF-IDF (Term Frequency-Inverse Document Frequency).
Model Architecture	Combining two approaches: 1. Naive Bayes classifier utilizing TF-IDF features. 2. LSTM neural network for sequential text analysis.
Evaluation Metrics	Evaluating model performance using accuracy, precision, recall, and F1-score to assess classification effectiveness and robustness.
Model Deployment and Integration	The trained model is deployed for practical use via a user-friendly Streamlit interface. Integrating it into a website for real-time analysis via Render.com.
Ensemble Model	Utilizing a weighted ensemble model that combines Naive Bayes and LSTM, with Naive Bayes assigned a lower weight (30%) compared to LSTM.
Limitations and Future Considerations	Blending ensemble technique performs poorly on predicting class 0 sentiment. Overfitting of class 1 in the ensemble model despite high training accuracy. Lack of interpretability in blending technique. Future optimization of blending coefficients and model selection to address limitations.
Data Limitations	Class imbalance, especially in predicting class 0 sentiment, affects model performance. TF-IDF implementation preferred over Word2Vec due to potentially lower computational complexity.
Future Directions	Further exploration of advanced LSTM architectures (e.g., bi-directional, attention-based) for improved sentiment analysis accuracy. Continuous refinement of text preprocessing techniques and incorporation of advanced NLP methods (e.g., dependency parsing, named entity recognition). Evaluation of alternative vectorization techniques and consideration of ensemble learning methods for leveraging diverse model strengths.

Project Work Table

Category	Naive Bayes	LSTM	Ensemble (Blending)
Interpretation	Naive Bayes model achieves an accuracy of 88.27% on the test dataset. It demonstrates good precision, recall, and F1-score for both classes. However, it struggles in correctly identifying instances of class 0 (negative sentiment), resulting in lower precision and recall compared to class 1.	LSTM model outperforms Naive Bayes with a test accuracy of 95.22%. It exhibits high precision, recall, and F1-score for both classes, particularly excelling in identifying instances of class 0.	The blending ensemble technique achieves a test accuracy of 74.41%. However, it performs poorly in predicting instances of class 0, indicating an issue with the ensemble's ability to capture negative sentiment effectively.
Model Tuning	Hyperparameter tuning was performed to optimize Naive Bayes' smoothing parameter and prior probabilities. Grid search and cross-validation were used to find the optimal values.	LSTM architecture was tuned by experimenting with various network architectures, including different numbers of layers and units, activation functions, and dropout rates. Hyperparameter tuning was conducted using grid search and validation data.	Blending coefficients were adjusted to find the optimal combination that maximizes the ensemble's performance. Grid search or random search techniques were utilized to search for the best blending weights.
Issues Encountered	Naive Bayes struggles in correctly identifying instances of class 0 (negative sentiment), resulting in lower precision and recall for this class. Class imbalance may exacerbate this issue.	Overfitting observed in the LSTM model, as indicated by the significant drop in performance between training and test datasets. LSTM may overfit to instances of class 1, leading to suboptimal performance on unseen data.	The ensemble model performs poorly in predicting instances of class 0, indicating a limitation in capturing negative sentiment effectively. Overfitting of class 1 observed despite high training accuracy, suggesting the need for regularization techniques.

Solutions	Address class imbalance by applying techniques such as oversampling, undersampling, or using class weights to give more weight to the minority class during training. Further refine text preprocessing techniques to improve the model's ability to capture negative sentiment.	Implement regularization techniques such as dropout or L2 regularization to mitigate overfitting in the LSTM model. Use techniques like early stopping to prevent overfitting during training.	Investigate and address the root cause of poor performance in predicting class 0 sentiment in the ensemble model. Implement regularization techniques or adjust the blending coefficients to prevent overfitting of class 1.

Dataset and Data Collection

Data Source: https://www.yelp.com/dataset/download

Definition of Columns

business.json:
business_id: Business ID
name: business's name
address: the full address of the business
city: the city
state: state code, if applicable
postal_code: the postal code
latitude: latitude
longitude: longitude
stars: star rating, rounded to half-stars
review_count: number of reviews
is_open: 0 or 1 for closed or open, respectively
attributes: business attributes to values. note: some attribute values might be objects
categories: an array of strings of business categories
hours: an object of key day to value hours, hours are using a 24hr clock

review.json:
character: unique review id
review_id: character unique review id
user_id: character unique user id, maps to the user in user.json
business_id: character business id, maps to business in business.json
stars: star rating
date: date formatted YYYY-MM-DD
text: the review itself
useful: number of useful votes received
funny: number of funny votes received
cool: number of cool votes received

Project Board: https://aml-2304-final-project.atlassian.net/jira/software/projects/KAN/boards/1

Github Repository: https://github.com/dars180602/Sentiment-Analysis-Interface

Discussion of Models and Results

Naive Bayes:
Test Accuracy: 88.27%
Training Accuracy: 83.26%
Precision:
Class 0: 77% (support: 321,703)
Class 1: 92% (support: 935,548)
Recall:
Class 0: 78%
Class 1: 92%
F1-score:
Class 0: 77%
Class 1: 92%

LSTM:
Test Accuracy: 95.22%
Training Accuracy: 93.79%
Precision:
Class 0: 88% (support: 321,703)
Class 1: 98% (support: 935,548)
Recall:
Class 0: 94%
Class 1: 96%
F1-score:
Class 0: 91%
Class 1: 97%

Observations:

The LSTM model outperforms the Naive Bayes model regarding accuracy, precision, recall, and F1-score on both test and training datasets.

The Naive Bayes model demonstrates good performance, but it's notably surpassed by the LSTM model, particularly in precision and recall for class 0, which corresponds to negative sentiment.

The LSTM model is better at correctly identifying instances of class 0 (higher precision and recall), reflected in its higher F1-score for class 0 compared to Naive Bayes. So, the LSTM model demonstrates that it can identify negative sentiment better than the naive Bayes model.

Based on these results, the LSTM model is the better choice for this classification task.

Ensemble Technique

The ensemble technique, blending, was employed to enhance the predictive performance of a sentiment analysis model. Blending combines predictions from two distinct models, Naive Bayes and Long Short-Term Memory (LSTM), using predetermined blending coefficients. The ensemble aims to leverage the diverse strengths of each model to improve overall accuracy and robustness.

The blended model performs poorly on the test dataset, particularly in predicting class 0, as evidenced by the extremely low precision, recall, and F1 score.

The model overfits class 1, as indicated by the high accuracy of the training dataset but the poor performance of the test dataset.

Optimization Strategies:

Hyperparameter Tuning: The blending coefficients (weights) could be optimized using grid or random search techniques to find the combination that maximizes performance metrics.

Model Selection: Instead of blending Naive Bayes and LSTM, experimenting with different models or using more advanced ensemble techniques like stacking or boosting could improve performance.

Feature Engineering: Considering additional features or enhancing existing ones might improve the models' ability to capture patterns in the data.

Cross-Validation: Evaluating the model using cross-validation can provide a more reliable estimate of its performance and help detect overfitting.

Model Interpretation:

Blending makes it challenging to interpret the contributions of individual models to the final predictions. Assessing the ensemble's interpretability and implications for the specific application is crucial.

Data Imbalance:

The class imbalance (especially class 0) might influence the model's performance. Techniques such as oversampling, undersampling, or using class weights could address this issue.

While blending can be a powerful technique, it's essential to carefully optimize the blending process and consider other strategies to improve model performance and generalization.

Challenges and Future Directions

Further exploration of advanced LSTM architectures, such as bi-directional or attention-based models, could yield even more significant improvements in sentiment analysis accuracy by capturing richer contextual information.

Our journey towards enhancing model performance is a continuous one. By refining and experimenting with text preprocessing techniques and incorporating advanced NLP methods like dependency parsing and named entity recognition, we can stay at the forefront of sentiment analysis.

While TF-IDF proved effective, periodic evaluation of alternative vectorization techniques such as Word2Vec and contextual embeddings (e.g., BERT) could provide insights into potential performance gains.

Consider increasing the complexity of the RNN architecture by incorporating more units and layers alongside techniques like dropout regularization to strike a balance between computational efficiency and model sophistication.

Explore ensemble learning techniques, such as combining multiple RNN architectures or incorporating other machine learning algorithms, to leverage diverse model strengths and improve overall sentiment analysis accuracy.

Conclusion

The LSTM model outperformed the Naive Bayes model, indicating its ability to capture deeper relationships among words and thereby enhance sentiment analysis accuracy.

Neither the Naive Bayes model nor the RNN exhibited overfitting, as evidenced by similar performance on training and test datasets, suggesting robustness in generalization.

Text preprocessing, particularly handling negations and contractions, significantly boosted the Naive Bayes model's accuracy by at least 10%, underscoring the critical role of meticulous preprocessing in enhancing model performance.

Although TF-IDF and Word2Vec were explored, ultimately, TF-IDF implementation was preferred due to comparable results and potentially lower computational complexity.

The implemented RNN architecture, with its basic structure and low-dimensional token mapping, was deliberately designed to maintain low computational demands. However, employing a more complex RNN architecture could further enhance model performance.

References

Pang, B., & Lee, L. (2008). "Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

Liu, B. (2012). "Sentiment Analysis and Opinion Mining." Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

Zhang, B., & Wallace, B. C. (2015). "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification."

Dos Santos, C. N., & Gatti, M. (2014). "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." Proceedings of COLING 2014, 69–78.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data." SIGKDD Explorations Newsletter, 6(1), 20–29.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 16, 321–357.

Dietterich, T. G. (2000). "Ensemble Methods in Machine Learning." In Multiple Classifier Systems (pp. 1–15). Springer.