Machine Learning & Predictive Analytics
ADSP 31009

Natural Language Processing on McDonald's Yelp Reviews

Vincent Feng
May 2024

GitHub Link:
https://github.com/vfeng6704/Yelp-NLP--Machine-Learning

Executive Summary

I. Business Problem

- Customer insights are key for building a competitive advantage[1] in business
- Understanding customer sentiment and their causes at scale is difficult with traditional methods

II. Scope of Work

- (1) Build a deep learning model to predict McDonald's Yelp reviews
- (2) Use explainable-AI techniques (e.g. SHAP) to identify key words that influence review ratings

III. Model Results

- Test Metrics: Accuracy: 0.93, F1 Score: 0.93, Recall: 0.93, Precision: 0.93
- Train Metrics: Accuracy: 0.98, F1 Score: 0.98, Recall: 0.98, Precision: 0.98

IV. Future Work

- Improve overfitting and interpretability
- Aspect Based Sentiment Analysis (ASBA)
- Sentiment trend analysis
- Expand dataset to include reviews from X, Google Reviews, Reddit, etc.

Notes: [1]Kahn, I., Leinwand, P., & Mani, M. M. (2022, October 20). How to Gain a Competitive Advantage on Customer Insights. Harvard Business Review

Overview: Analytical Methodology

1 Create Dataset

- Used The Yelp Open Dataset, filtering for McDonalds
- There are ~17K reviews spanning from 2007 to 2022

2 Preprocessing

- Data cleaning included stripping white spaces and removing special characters and punctuations
- Implemented NLTK Stopwords, Tokenize, and Lemmatizer for preprocessing

3 EDA

- Leveraged BERTopic package to extract topics from reviews
- Ran Flair package to understand sentiment on topics

4 Proposed Approaches

- Built Logistic Regression, Naïve Bayes, and deep learning models to predict yelp ratings based on review text
- Experimented with explainable AI packages to understand key words that influenced review ratings

5 Model Evaluation

- Measured precision, recall, F1-score, and accuracy scores for review ratings across test and training sets

Assumptions/Hypotheses about data and model

Data Assumptions

- There is strong correlation between the content of reviews and their star ratings
- While Yelp reviews may not fully represent the entire customer base due to their tendency to reflect extreme opinions (e.g., dissatisfied customers), this analysis remains valuable for our business case
- By simplifying the problem from multi-classification to binary classification (only 1 and 5 stars), I will remove noise and improve model performance

Model Hypotheses

- By capturing nuances in reviews that simpler models miss, deep learning models will achieve superior performance
- Logistic regression and Naïve Bayes algorithms will be used to create a baseline for comparing the deep learning model's performance

EDA: Negative Word Sentiment

Word Cloud on Positive Sentiment,
McDonald’s Reviews on Yelp (2007-2022)

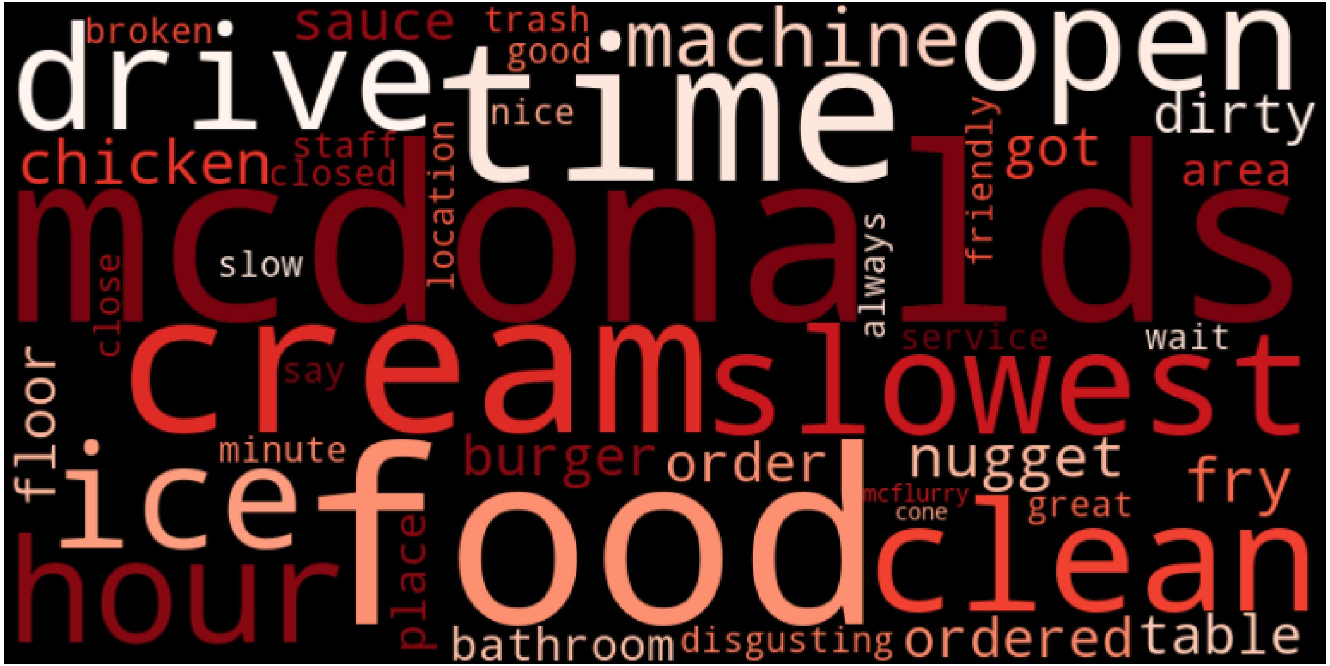
Commentary

- Leveraged BERTopic modeling and Flair package to understand words associated with **positive** sentiment

- Positive words include topics around customer service, cleanliness of the restaurant, breakfast items (e.g. McMuffin, egg, hash, bacon, sausage) and whether this specific McDonalds location was better than others

EDA: Negative Word Sentiment

Word Cloud on Negative Sentiment,
McDonald's Reviews on Yelp (2007-2022)



Commentary

- Leveraged BERTopic modeling and Flair to understand words associated with **negative** sentiment

- Negative sentiment include ice cream (i.e. McFlurry), how dirty the restaurant was, how slow the restaurant was compared to other locations, drive time, and if the restaurant was closed

Data Cleaning and Feature Engineering

1

Text Preprocessing

Data Cleaning: removed special characters, punctuations, and white space

Tokenization*:* split the text into individual tokens

Stop Words*:* removed common words that may not carry significant meaning

Lemmatization: reduced words to their base or root form

2

Feature Engineering and Transformation

TF-IDF: weighed words by their frequency in a document relative to their frequency in the entire corpus

Target Variable Filtering: focused on binary classification for 1 and 5 star reviews

Standard Scaler: scaled the values of the TF-IDF sparse matrix for modeling

Proposed Approaches and Solution

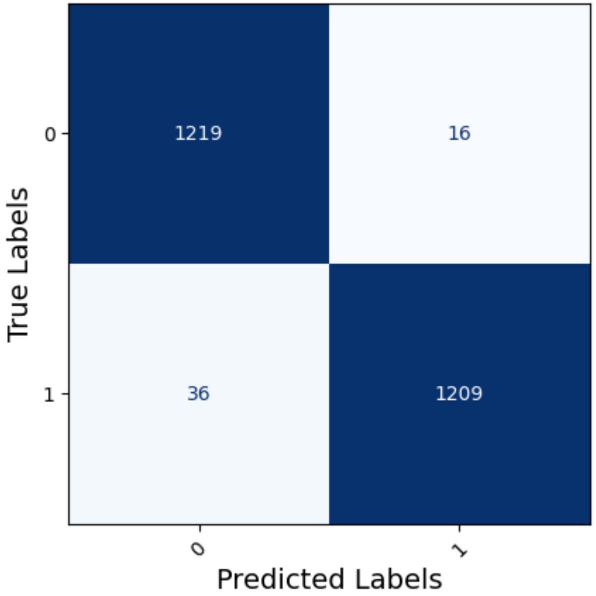
	Logistic Regression	vs	Naïve Bayes	vs	*Proposed Model* LSTM
Description	Linear model where a linear decision boundary is fit		Probabilistic classifier based on Bayes' theorem		Deep learning model
Metrics	**Test** • Accuracy: 1.0 • F1 Score: 1.0 • Recall: 1.0 • Precision: 1.0		**Test** • Accuracy: 0.89 • F1 Score: 0.89 • Recall: 0.89 • Precision: 0.89		**Test** • Accuracy: 0.90 • F1 Score: 0.90 • Recall: 0.90 • Precision: 0.90
	Train • Accuracy: 0.88 • F1 Score: 0.88 • Recall: 0.88 • Precision: 0.88		**Train** • Accuracy: 0.92 • F1 Score: 0.92 • Recall: 0.92 • Precision: 0.92		**Train** • Accuracy: 0.98 • F1 Score: 0.98 • Recall: 0.98 • Precision: 0.98

Notes: F1 Score, Recall, and Precision show the weighted avg results

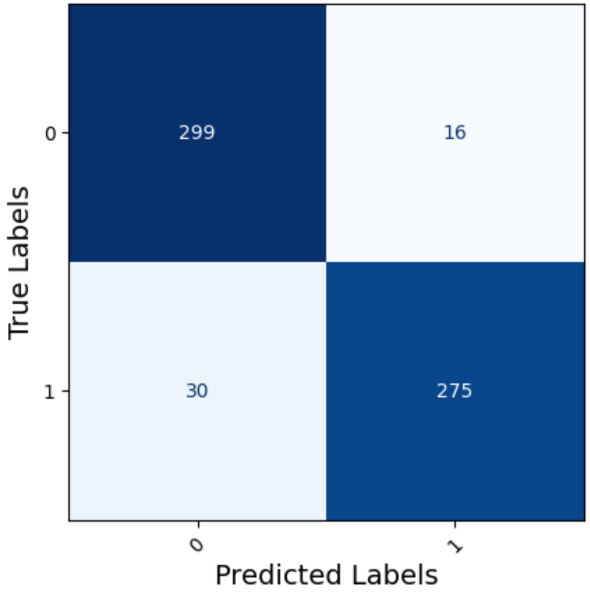
Final Model Results

Confusion Matrix,
Final Neural Network

Test
Accuracy: 0.93, F1 Score: 0.93, Recall: 0.93, Precision: 0.93



Train
Accuracy: 0.98, F1 Score: 0.98, Recall: 0.98, Precision: 0.98



Commentary

- Deep learning looked the most promising based on results, but there is still clear signs of overfitting
- To improve performance, I ran regularization and Bayesian tuning
- Final Model: Neural Network with one embedding layer, global average pooling layer, dropout layer, dense layer, and output layer

Notes: F1 Score, Recall, and Precision show the weighted avg results

Lessons from the Methodology

I. NLP

- This was my first time working with NLP, allowing me to explore many topics
- Text data requires preprocessing and feature extraction (e.g. word embeddings) before modeling
- I attempted to use BERT model for contextual embeddings, but was unsuccessful due to computation constraints (e.g. ran for 10+ hours)

II. Deep Learning Techniques

- Explainable AI (XAI) techniques enables interpretability for AI models
- BERT has deep contextual understanding, allowing it to accurately capture the sentiment expressed in a review and have better prediction power

Future Work

I. Improve overfitting and interpretability

- Despite regularization and reducing model complexity, my model still overfit
- Results from Lime and SHAP were not as insightful as I hoped for

II. Aspect-Based Sentiment Analysis (ABSA)

- Given the criteria of this project, I decided to prioritize the analysis performed because it had clear evaluation metrics
- ASBA enables sentiment of a text with respect to a specific aspect, including things like food quality service, etc.

III. Sentiment Trend analysis

- Analyze sentiment trends over time to detect patterns of shifts in customer satisfaction
- Use time series analysis to correlate these trends with external factors (e.g., new menu items, promotional campaigns)

IV. Dataset

- I would like to create a more robust dataset to properly account for the voice of the customer across all major review platforms, including X, Reddit, and Google Reviews