Yelp Dataset Sentiment Analysis

Literature Review Survey

Cole Hanrahan and Julia Sober
COSC 5540 — Text Mining & Analysis
Fall 2025

GEORGETOWN UNIVERSITY

Background and Motivation

- Online platforms like Yelp contain millions of reviews, but the link between text and star ratings is not straightforward.
- Automating star prediction helps users and businesses understand sentiment quickly, and researchers analyze which aspects most affect ratings.
- Literature shows high accuracy for binary sentiment, but more nuanced star prediction is difficult due to imbalanced data, reviewer subjectivity, and noisy language.

> *Restaurant: XYZ, Kitchener, N2G4Z6, Canada*
> *Rating:* ● ● ● ○ ○
> *I've been to XYZ a bunch of times. It's a decent place. Nice food, lots of variety! The place is really small though, so you almost never find a spot to sit and eat. The service is also slow at times.*

Source: Yelp Dataset Challenge: Review Rating Prediction, Nabiha Asghar

Research Directions in Literature

1. **Binary Sentiment Classification** – positive vs. negative reviews.
2. **Star Rating Prediction** – predict 1–5 stars from text.
3. **Bias Analysis** – explore inconsistencies between text and rating.
4. **Aspect-based sentiment analysis (ABSA)** — how does the reviewer feel about certain “aspects” of the product/restaurant (e.g. good food, bad service). Can we use these extracted opinions in our model?

Objectives

- Accurately predict a restaurant's star rating (1–5) from **review text alone**.
- **Compare multiple feature extraction methods** (n-grams, TF-IDF, sentiment lexicons, embeddings).
- Explore **supplemental features** (review length, punctuation usage) and **address class imbalance**
- Benchmark baseline ML models against deep neural networks and ensemble approaches.

Dataset

- Yelp Dataset Challenge (~7–8 million reviews).
- Each review includes text and star rating (+ many more features); known imbalance with most ratings ≥4 stars.
- Similar datasets in research: Amazon, Twitter, IMDB for method comparison.
  - Some papers used these

Preprocessing Methods

- Lowercase conversion, punctuation/special character removal, stop word elimination.
- Extract features (length, punctuation) before cleaning.
- Negation handling (“not good” → “not_good”), optional stemming.
- Part of Speech tagging

Feature Engineering Approaches

- **Bag of Words (BoW):**
  - Baseline—simple but effective
- **TF-IDF** and **n-grams** (unigrams, bigrams, trigrams):
  - Define rare/meaningful words/phrases
- **Sentiment Lexicons:**
  - Compare text with predefined word lists (e.g., Bing Liu)
  - Struggle with slang/typos
- **Part-of-Speech Tagging:**
  - e.g. use only adjectives
  - improves interpretability and model performance
- **Contextual Embeddings:**
  - e.g. BERT
- Try adding extracted features like review length, positivity/negativity scores, user votes if feasible

Modeling Approaches

- **Baseline classifiers:** Logistic Regression, Naive Bayes, SVM (linear/non-linear), Passive-Aggressive (similar to SVM).
- **Tree-based ensembles:** Random Forest, AdaBoost.
- Shallow/deep **Multi-Layer Perceptrons** with contextual embeddings.
- **Ensemble methods:** weighted voting, stacking.
- **Ordinal models** (ordered logistic regression) to honor rating order (i.e. 1 is closer to 2 than 4).
- **Regressors** (RandomForest, SVM, etc.), layered strategies (classify sentiment then regress rating).

Best Results

Model	Dataset	Classification Type	Preprocessing	Feature Extraction	Results	Notes/ Suggestions
Naïve Bayes	Amazon	**binary**	tokenization, stop word / punctuation removal, stemming	**BoW → opinion lexicon →** sentiment score	**98%**	ABSA (e.g. camera's quality, megapixel, etc.)
Logistic Regression	Yelp	**1-5** star rating	capitalizations / stop word / punctuation removal	top 10,000 **unigrams + bigrams**	64%	Ordinal linear regression, PoS tagging
Ensemble: LR + NB + SVM + MLP	Yelp	**1-5** star rating	unspecified	TF-IDF for baseline, BERT for MLP	58%	**Worst accuracy with 2 & 3 star ratings**
Weighted: Bi-LSTM + CNN	Yelp	**binary**	punctuation / non-alphabetic /special character removal, NLTK word_tokenize()	BERT	0.90 **(F1-score)**	Shallow models perform better
Linear Regression	Yelp	**1-5** star rating	unspecified	BoW (top k), PoS (adjective) extraction	0.64 **(RMSE)**	Treated as a **regression** problem

Review of Findings

- Binary sentiment models (Naive Bayes, Random Forest) good at **positive/ negative**
- Multi-class and regression models (Logistic/SVM, Linear Regression) weaker for all 5 star ratings (**accuracy ≤ 65%**).
- **TF-IDF and n-gram** features are standard; context-sensitive embeddings (BERT) have emerged for fine-grained results.
- Lexicons help but struggle with **slang, typos, noisy data; sarcasm** and nuanced language largely unsolved.
- **Ensemble methods** sometimes improve results; deep learning gaining ground, especially for multi-aspect analysis and context.
- **Trends observed:**
  - More features ≠ better performance (can cause overfitting).
  - Simple > Complex Models

Interesting Findings

- Standard preprocessing and Bag of Words are still strong baselines.
- Too many features can hurt linear regression results; keep it simple when possible.
- Context-aware models (e.g., BERT) for sarcasm, multi-aspect sentiment are promising.
- Aspect-Based Sentiment Analysis worth trying.

Challenges and Gaps

- **Context & Sarcasm:** “Great job burning my pizza!” — an unsolved problem
- **Lexicon mismatch:** limited by real-world data noise, slang, and sarcasm
- *****Imbalanced data***:** most reviews are 4 or 5 stars
- **Explaining middle ratings:** very difficult to discern between 2 & 3 star ratings

Our Expected Outcomes

- Benchmark accuracy (60–65%) for unigrams/bigrams plus logistic regression or SVM, matching published results.
- Potential performance gain from contextual embeddings and ensemble techniques.
- Deeper insights: which features most help, and which star ratings are toughest to classify.

Papers Read

Sentiment Analysis on Product Reviews Using Machine Learning Techniques; Rajkumar S. Jagdale, Vishal S. Shirsat and Sachin N. Deshmuk

Yelp Dataset Challenge: Review Rating Prediction; Nabiha Asghar

Sentiment Analysis: A Systematic Case Study with Yelp Scores; Wenping Wang Et al.

Ensemble Sentiment Analysis Using Bi-LSTM and CNN; Puneet Singh Lamba Et al.

Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM; Naimul Hossain Et al.

Sentiment Analysis of Yelp Reviews by Machine Learning; Hemalatha S, Ramathmika

Sentiment Analysis on Food Review using Machine Learning Approach; Nourin Islam, Ms. Nasrin Akter, Abdus Sattar

Sentiment Analysis using Machine Learning Techniques on Python; Ratheee Et al.

Sentiment Analysis of Yelp's Ratings Based on Text Reviews; Xu Et al.

Predicting a Business' Star in Yelp from Its Reviews' Text Alone; Fan, Khademi