Literature Review

Yelp affords researchers a large volume of reviews for free, and several researchers have already put this textual data to good use in the field of natural language processing, including in sentiment analysis.

In two separate analyses focused on Yelp restaurant reviews, researchers first performed topic modeling (also referred to as aspect extraction). This step endeavors to extract detail from nuanced texts. For example, if a reviewer complimented a restaurant's food but complained about the wait time, that should result in two sentiments: a positive sentiment for the food topic and a negative sentiment for the time topic.

Zhang et al. predetermined five categories – price, time, food, service, and location – and used custom-built sentiment dictionaries to identify positive, negative, and negation words (words that change a proceeding negative word to a positive word, or positive to negative). This is helpful in the context of restaurants, as these are five topics relevant to reviewers' experiences and on which restaurateurs can take action to improve their business. However, this lexicon-based method (as opposed to a deep learning method) is static and relies on human input to identify words that much certain sentiments and which topics are important.

Zhang et al. employed Latent Dirichlet allocation (LDA) to model topics, then used the TextBlob Python library for sentiment analysis. LDA is an appropriate choice when working with short documents (Chen et al., 2020). The model tokenizes documents into bags of words (BoW) and assumes that documents with similar topics use similar groups of words, which enables it to identify topics latent in the documents. It outputs matrices detailing the probability distributions over documents and the word probability distributions over topics. LDA uses alpha and beta hyperparameters. The former controls the mixture of topics for a document, and the latter controls the number of words per topic. A high alpha results in a higher mix of topics; a high beta results in more words in topics. These hyperparameters are tuned through experimentation.

In contrast, Alamoudi and Alghamdi used GloVe to isolate four latent topics: food, service, ambience, and price. They considered BOW models, N-grams, Term Frequency-Inverse Document Frequency (TF-IDF), or word embeddings, like word2vec and GloVe (Pennington et al., 2014). They decided on GloVe because other methods ignore "contextual similarities and the semantic similarity between extracted words," whereas in word embedding, "words are represented by vectors of real numbers, and the words of similar meaning have similar representation."

Once the latent topics were identified by GloVe modeling, Alamoudi and Alghamadi performed sentiment analysis using a Bidirectional Encoder Representations from Transformers (BERT) model. This model can be tuned with hyperparameters that control the maximum learning rate and evaluated using accuracy, precision, recall, and F1 score; Alamoudi and Alghamdi also measured LogLoss, though this is less commonly used than the other evaluation metrics.

Furthermore, Alamoudi and Alghamdi utilized the star ratings given by the Yelp reviewers. They made a scatter plot with the Yelp star ratings on the y axis and the average sentiment rating across all topics on the x axis, showing a positive, linear relationship that implied the sentiments extracted by their model matched closely with the sentiment intended by the reviewers.

In a very recent study, Belaroussi et al. investigated the effectiveness of Long Short-Term Memory (LSTM) networks in sentiment analysis of Yelp reviews versus various types of Bidirectional Encoder Representations from Transformers (BERT) models. They found that a domain-specific trained LSTM model outperformed BERT-based models in accuracy, precision, recall, and F1 score. “These findings suggest that well-tuned LSTM architectures offer a lightweight and effective alternative to large pretrained transformer models, particularly for structured and domain-specific datasets like Yelp reviews.” However, they also noted that BERT-based models remain effective, work better with diverse sets of documents, and are more generalizable to less domain-specific contexts.

Belaroussi et al. did not perform any topic modeling, given that their goal was to assess the ability of various models to identify simply positive, neutral, and negative sentiments.

Our use case is to analyze sentiments of Yelp reviews related to touristic sites and identify what cities should better utilize or improve. Given this, topic modeling is a valuable step for us. Performing this step using GloVe to identify latent topics makes more sense than the lexicon-based method Zhang et al. employed since touristic sites are a more diverse group than restaurants, so pre-determining topics and sentiment dictionaries is ill-suited to the task.

BERT-based and LSTM-based models are both well suited to the sentiment analysis task, but LSTM makes the most sense for our use case because of its high performance in accuracy, precision, recall, and F1 score and because generalizability is not needed in this case.

We will measure our model’s performance using accuracy, precision, recall, and F1 score and by plotting our derived sentiments against Yelp reviewers’ ratings.

References

Zhang, Sonya, et al. "Topic Modeling and Sentiment Analysis of Yelp Restaurant Reviews," International Journal of Information Systems in the Service Sector (IJISSS) 14, no.1: 1-16. https://doi.org/10.4018/IJISSS.295872

Ariyasriwatana, W., & Quiroga, L. M. (2016). "A thousand ways to say 'Delicious!' - Categorizing expressions of deliciousness from restaurant reviews on the social network site Yelp," Appetite, 104: 18–32. doi:10.1016/j.appet.2016.01.002 PMID:26774047

Chen, W., Yu, Z., Xian, Y., Wang, Z., & Wen, Y. (2020). Mining Keywords from Short Text Based on LDABased Hierarchical Semantic Graph Model. International Journal of Information Systems in the Service Sector, 12(2), 12. doi:10.4018/IJISSS.2020040106

Hung, Lai Po and Alias Suraya. 2023. "Beyond Sentiment Analysis: A Review of Recent Trends in Text Based Sentiment Analysis and Emotion Detection." Journal of Advanced Computational Intelligence and Intelligent Informatics 27 (1) (01): 84-95. doi:https://doi.org/10.20965/jaciii.2023.p0084. https://link.ezproxy.neu.edu/login?url=https://www.proquest.com/scholarly-journals/beyond-sentiment-analysis-review-recent-trends/docview/2766706720/se-2.

Alamoudi, Eman Saeed, and Norah Saleh Alghamdi. "Sentiment Classification and Aspect-Based Sentiment Analysis on Yelp Reviews Using Deep Learning and Word Embeddings." Journal of Decision Systems 30, no. 2–3 (2021): 259–81. https://doi.org/10.1080/12460125.2020.1864106.

Pennington, J., Socher, R., & Manning, C.D. (2014). GloVe: Global vectors for word representation. EMNLP 2014: 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543. https://doi.org/10.3115/v1/d14-1162

Belaroussi, Rachid, Sié Cyriac Noufe, Francis Dupin, and Pierre Olivier Vandanjon. "Polarity of Yelp Reviews: A BERT–LSTM Comparative Study." Big Data and Cognitive Computing 9, no. 5 (2025). https://doi.org/10.3390/bdcc9050140.