Topic Modeling and Fine-Grained Sentiment Analysis of Movie Reviews

Matthew Lucht
University of Rochester
Rochester, New York, USA, 14627
Email: mlucht@rochester.edu

***Abstract*—This paper presents a detailed pipeline for analyzing movie reviews by extracting thematic elements and classifying categorized sentiments with machine learning models. Initially, textual reviews are processed using TF-IDF vectorization and enriched by integrating semantic embeddings via GloVe. A binary sentiment classifier using support vector machines (SVM) distinguishes positive from negative sentiment. Subsequent gradient-boosted trees further refine sentiment intensity into 4 total categories. Nonnegative matrix factorization (NMF) discovers key themes within each sentiment class through text analysis, which are then fed into a GPT-4 prompt which creates concise, human-readable summaries. The number of reviews are also fed into a simple formula to calculate a predicted overall point score for the movie. The pipeline was trained on IMDB and Rotten Tomatoes review datasets with over 100,000 reviews and tested on two individual movie review datasets containing approximately 1000 reviews each. The results of this test demonstrate the effectiveness and effectiveness of this methodology in creating informative summaries of the average sentiment and calculating an accurate average score compared to the ground truth.**

I. Introduction

Machine learning-powered sentiment analysis has emerged as a powerful tool to quickly gauge public opinion without needing to manually sift through numerous lengthy reviews. For consumers, sentiment analysis offers immediate access to an aggregated overview of public sentiment, clearly highlighting general attitudes toward media such as films or television series. It simplifies the identification of important and frequent praises and criticisms, enabling users to quickly determine the collective sentiment toward a movie.

Sentiment analysis is also helpful to producers and marketers, providing valuable insight into audience and market feedback by streamlining the process of identifying precisely what viewers appreciate or dislike about their productions. This rapid feedback mechanism can help decision makers address critical viewer concerns, refine marketing strategies, and help producers create a better overall product.

Despite its utility, automated sentiment analysis often neglects nuanced opinions, traditionally favoring simplistic binary distinctions. Also, conventional topic modeling techniques frequently overlook sentiment-specific contextualization, losing critical subtleties related to the actual reasons behind the sentiments. This paper aims to address these limitations by combining classical NLP techniques with modern embedding approaches and providing more granular sentiment labels.

Moreover, the general structure and methodology of this pipeline can easily be adapted and expanded beyond movie reviews to analyze customer sentiment in various domains, such as Amazon product reviews or Yelp store reviews. This broad applicability enables businesses and service providers to directly understand customer preferences, pinpoint areas that need improvement and enhance overall customer satisfaction [**?**].

II. Method

The pipeline is made up of several interconnected stages:

A. Feature Extraction

Raw review text from datasets undergo initial preprocessing by stripping HTML tags and converting to lowercase. Text data is vectorized using TF-IDF (Term Frequency-Inverse Document Frequency), which quantifies the significance of terms based on their frequency across the corpus. Simultaneously, 100-dimensional GloVe (Global Vectors for Word Representation) embeddings are integrated. These embeddings capture semantic relationships among words by leveraging co-occurrence statistics from the corpus. An Embedding Vectorizer averages these embeddings, and a unified feature space is formed by combining TF-IDF and embedding vectors using a Feature Union.

B. Binary Sentiment Classification

A linear-kernel Support Vector Machine (SVM) trained on the IMDB dataset separates the data into positive and negative reviews. The SVM constructs an optimal hyperplane that maximizes the margin between sentiment classes within the high-dimensional feature space, leveraging its robustness and generalizability, particularly effective in text classification scenarios.

C. Strong vs. Weak Sentiment Subclassification

To distinguish nuanced sentiment intensity within positive and negative subsets, Gradient Boosting classifiers are trained separately for each subset. This is trained on the Rotten Tomatoes dataset, which contains a 1-10 score for each review instead of simply positive or negative overall sentiment, allowing more nuance to the sentiment. Gradient Boosting incrementally builds an ensemble of decision trees by sequentially minimizing residual errors from previous iterations. This yields four more precise sentiment categories: strong negative, weak negative, weak positive, and strong positive.

D. Topic Modeling

Within each sentiment subclass, Non-negative Matrix Factorization (NMF) is applied to reveal latent thematic structures. NMF factorizes the TF-IDF representation into low-dimensional topic components by constraining factors to non-negative values, ensuring interpretability. Extracted topics are characterized by their most significant terms, providing clear thematic distinctions between sentiment classes.

E. Score Calculation

To quantify overall sentiment, a weighted scoring system assigns values to each sentiment category: strong negative (0.0), weak negative (0.33), weak positive (0.66), and strong positive (1.0). The average review score is calculated as the weighted mean of these scores, mapped onto a 0–10 scale. The ground truth average score is also calculated by simply averaging the score column of the review dataset, and the predicted vs. ground truth score is printed.

F. Summary Generation

Finally, the resultant topic terms for each subcategory of sentiment are compiled and provided to GPT-4, a large language model (LLM), tasked with a prompt that asks it to create a coherent summary. GPT-4 generates a concise, informative description that highlights reviewers' core positive and negative sentiments, adding quantitative analysis on top of the score's qualitative insight.

III. Experiments

A. Datasets

The pipeline was trained using the IMDB dataset (approximately 50,000 reviews) and the Rotten Tomatoes dataset (approximately 50,000 reviews). Two individual movie datasets, each containing approximately 1000 reviews (text and scores), were used for testing of the pipeline.

B. Classification Results

The binary SVM achieves an accuracy of approximately 90% on the IMDB dataset. Gradient Boosting classifiers effectively differentiate strong and weak sentiments with accuracy of 73% to 64% on negative and positive reviews respectively. Confusion matrices confirm robust performance, despite slightly imbalanced sentiment sub-category distributions.

C. Topic Quality

NMF-generated topics exhibit clear thematic coherence. Strong positive reviews prominently feature aspects such as "visuals," "acting," and "epic storytelling," whereas weak negative reviews commonly highlight issues with "pacing," "plot holes," and "dialogue." Qualitative assessments of several passes over the test datasets validate the thematic alignment with sentiment intensity.

D. Score accuracy

Potentially the most impressive part of the results of testing on these datasets was the accuracy of the predicted score compared to the ground truth score. Example dataset 1 had a predicted score of 8.47 vs a ground truth score of 8.56, within 9%. Example dataset 2 had a predicted score of 7.61 vs a ground truth score of 7.63, within 3%.

E. Summary Examples

An example GPT-4 summary output from the pipeline can be shown to succinctly summarize prevalent sentiments: "Denis Villeneuve's adaptation of "Dune" is a stunning visual and cinematic masterpiece that brings Frank Herbert's intricate sci-fi world to life with remarkable depth and authenticity. The performances by Timothée Chalamet, Zendaya, and Rebecca Ferguson are compelling, and the film's sound design and Hans Zimmer's score enhance the epic narrative, making it a must-see on the big screen. However, some may find the pacing slow and the adaptation of the complex source material a bit challenging to follow, with certain plot elements feeling rushed or underexplained, potentially leaving those unfamiliar with the book lost."

IV. Additional Testing

An additional test of the model was run against direct input of the data into ChatGPT 4o and 4.5, asking it to perform the same task that the pipeline is designed to. Both ChatGPT models were off by over 1 full score point when tested on both testing dataset, significantly less accurate than the pipeline. This gives credit to the pipeline's ability to perform efficiently at the specific task it was designed for over a very large general model.

V. Conclusion

This comprehensive pipeline integrates classical NLP (TF-IDF), semantic embeddings (GloVe), robust classification methods (SVM, Gradient Boosting), interpretable topic modeling (NMF), and advanced natural language generation (GPT-4). The resulting framework provides nuanced, sentiment-aware insights into movie reviews, demonstrating significant potential for broader applicability and extension into dynamic or multi-aspect sentiment domains.

References

[1] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *ECML*, 1998.

[2] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, 1972.

[3] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation," *EMNLP*, 2014.

[4] D. Lee, H. Seung, "Algorithms for Non-negative Matrix Factorization," *NIPS*, 2001.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation," *JMLR*, 2003.

[6] A. Maas et al., "Learning Word Vectors for Sentiment Analysis," *ACL*, 2011.