Sentiment Analysis and Topic Modeling on Amazon Movie Reviews

CSE-572 Final Project Portfolio Report

Rakshith Chandrashekar
Arizona State University
Tempe, AZ, USA
rchand44@asu.edu

1 Project Goals

Movie reviews, scattered across platforms in informal texts, lack coherence for traditional analysis. Sophisticated methods are essential for uncovering latent patterns and sentiments. Sentiment analysis[4], pivotal for movie reviews, extracts and categorizes sentiments—positive, negative, or neutral—offering insights into audience reactions. Its applications extend to social media, news, and customer feedback, spanning various industries. Topic modeling[15] complements sentiment analysis, revealing recurring motifs in reviews and enhancing audience understanding. Automating topic extraction streamlines data organization and search. Challenges persist with unstructured texts[13], where traditional methods falter due to polysemy, context, and noise. To tackle these challenges, we propose a comprehensive approach utilizing machine learning models, word embeddings, deep learning methods, innovative sampling, and advanced evaluation metrics.

Our methodology endeavors to enhance the precision and efficiency of sentiment analysis and topic modeling through a comprehensive comparison of cutting-edge machine learning algorithms and word embeddings against traditional approaches. Our primary focus centers on the relatively unexplored Amazon Movie Reviews dataset, boasting an extensive collection of over 8 million reviews. Our approach entails training algorithms on meticulously defined features and leveraging word embeddings to capture the intricate nuances of language and context embedded within movie critiques. To extract features, we delve into the realms of TFIDF vectorizer[10], Count vectorizer[2], and Glove word embeddings[12]. In our exploration, we scrutinize various algorithms such as Multinomial Naive Bayes[14], Logistic Regression[16], Support Vector Machine[9], and Multilayer Perceptron Network[1] for sentiment analysis. Similarly, for the task of topic modeling, we investigate methodologies including Latent Dirichlet Allocation[5], Latent Semantic Analysis[13], and Non-negative Matrix Factorization[2]. By amalgamating time-tested practices with contemporary innovations, our goal is to navigate the complexities inherent in deciphering unstructured and succinct textual data.

2 Individual Contributions

The project epitomized collaborative teamwork, where my contributions encompassed active engagement in group collaborations with fellow members and the completion of individual tasks seamlessly integrated into the overall project. My individual contributions involve the following:

- **Data Preparation** : The dataset available was in the form of a text file where each record is organized in row form with multiple lines of key-value pairs per record. I performed a line-by-line processing to extract and aggregate data. This transformed the dataset into a structured CSV file, facilitating analysis.
- **Data Preprocessing**: Preprocessing data, for our task was a critical step to ensure that the models receive clean and standardised input. This involved punctuations and stopwords removal, tokenization and lemmatization.
- **Feature Engineering**: I converted text data to numerical features using TF-IDF, Count vectorization, and Glove embeddings. Standard Scaler was applied for normalization. To optimize efficiency, I reduced the feature set from 357K to 2.5K by selecting the most common words across documents.
- **Sentiment Analysis**: To classify review texts into positive and negative classes, I implemented the following models for each vectorization method: TF-IDF (with n_features of 500, 1500, and 2500), Count vectorizer (with max_features of 500, 1500, and 2500), and Glove embeddings (ranging from 50D to 300D).
  1. **Naive Bayes (NB)** is a probabilistic learning method commonly used in Natural Language Processing. It calculates the probability, using Bayes' theorem, of a document belonging to a class based on the presence of terms, assuming independence between features. Two variants of this model (Multinomial Naive Bayes and Gaussian Naive Bayes) were utilized depending on the vectorization technique. The model achieved its highest accuracy of **0.883** with TFIDF vectorizer having n_features = 2500.
  2. **Logistic Regression (LR)** is a statistical method for analyzing datasets with one or more independent variables determining an outcome, measured with a dichotomous variable with only two possible outcomes. It attained its highest accuracy of **0.894** with TFIDF vectorizer having n_features = 2500 and using 'l2' penalty.
  3. **Support Vector Machine (SVM)** is a supervised learning model used for classification and regression. It aims to find a hyperplane in an N-dimensional space (where N is the number of features) that distinctly classifies data points. The model achieved its highest accuracy of **0.885** with the Count vectorizer having max_features = 2500 and hinge loss function.

(4) **Multilayer Perceptron (MLP)** networks are feedforward artificial neural networks consisting of multiple layers, including input, hidden, and output layers. Capable of learning complex patterns through supervised learning, my model was a Sequential neural network with two hidden layers (each with 128 units) and ReLU activation, followed by an output layer with sigmoid activation. Utilizing the Adam optimizer with binary cross-entropy loss function, the model achieved its highest accuracy of **0.943** with the Count vectorizer having max_features = 2500 over 20 epochs.

- **Evaluation Metrics**: Given the highly imbalanced class distribution in our dataset, assessing the model's effectiveness relied not only on the accuracy score but also on computing precision, recall, F1-score[8] and visualization through the ROC-AUC[6] curve.
- **Reporting**: I ensured timely report writing, presented findings effectively, and maintained comprehensive documentation. Through clear communication and meticulous documentation, I facilitated collaboration and transparency among team members, contributing significantly to the achievement of project goals.

3 Project Learnings

Throughout this project, I had the invaluable opportunity to practically apply the knowledge gained from the course, particularly regarding various text vectorization methods such as Count vectorizer, TFIDF vectorizer, and word embeddings like Glove. Additionally, I delved into feature reduction techniques such as PCA (Principal Component Analysis) and Linear Discriminant Analysis, enhancing my ability to preprocess and optimize textual data effectively.

Moreover, I deepened my understanding of classification algorithms including logistic regression, SVM, and Naive Bayes, as well as neural network architectures such as Multilayer perceptrons, Recurrent Neural Networks (RNN)[3], and Long Short-Term Memory Networks (LSTM)[7]. This hands-on experience underscored the importance of hyperparameter tuning to tailor models to specific tasks, optimizing performance and generalization.

Furthermore, I explored various supervised and unsupervised learning algorithms, including Latent Dirichlet Allocation (LDA), LSA (Latent Semantic Analysis), and Non-negative Matrix Factorization, expanding my repertoire of text analysis techniques.

Additionally, I familiarized myself with a range of model evaluation metrics such as accuracy score, precision, recall, F1 score, and ROC-AUC curve, enabling comprehensive assessment of model performance and aiding in informed decision-making throughout the project lifecycle.

4 Challenges

One of the primary technical challenges encountered in the project stemmed from the size of the dataset, comprising 8 million reviews. Handling such a vast amount of data posed difficulties in various aspects, including data preparation tasks like converting the text data from a multi-line row format to CSV, preprocessing, and executing complex models such as MLP, RNN, and LSTM. Due to computational constraints on my local system, performing these tasks locally was impractical. Additionally, the desire to explore BERT[11] embeddings for feature engineering remained unfulfilled due to these limitations.

To address these technical challenges, we utilized the free Tensor Processing Units (TPU) engine provided by Google Colab. However, the availability of this resource was limited and non-guaranteed, which continued to pose difficulties in our workflow.

Moreover, comprehensively understanding the supervised and unsupervised learning algorithms utilized in the project presented a considerable challenge. To overcome this, I delved into various online resources, reviewed suggested reading materials from coursework, and studied relevant research papers.

Non-technical challenges included effectively distributing workload among team members, ensuring timely progress, and maintaining equal workloads. To mitigate these challenges, we established weekly Zoom meetings to discuss progress, address any challenges encountered, and coordinate tasks. This proactive approach fostered open communication, facilitated prompt issue resolution, and ensured alignment among team members regarding project objectives and timelines. Additionally, we utilized project management tools to streamline task allocation and progress tracking, further enhancing team coordination and efficiency.

References

[1] 2013. Deep learning of representations: Looking forward. *CoRR* abs/1305.0445 (2013). http://arxiv.org/abs/1305.0445

[2] Dimitar Angelov. 2020. Top2Vec: Distributed Representations of Topics. *arXiv* arXiv:2008.09470 (2020). https://doi.org/10.48550/arXiv.2008.09470

[3] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya. 2021. ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems* 115 (2021), 279–294. https://doi.org/10.1016/j.future.2020.08.005

[4] N.F. da Silva, E.R. Hruschka, and E.R. Hruschka. 2014. Tweet sentiment analysis with classifier ensembles. *Decis. Support Syst.* 66 (2014), 170–179. https://doi.org/10.1016/j.dss.2014.07.003

[5] L. E. George and L. Birla. 2018. A Study of Topic Modeling Methods. In *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS '18)*. https://doi.org/10.1109/ICCONS.2018.8663152

[6] A. Go, R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. http://www.stanford.edu/alecmgo/papers/TwitterDistantSupervision09.pdf

[7] Z. Jianqiang and G. Xiaolin. 2018. Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access* PP (2018), 1. https://doi.org/10.1109/ACCESS.2017.2776930

[8] V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan. 2012. Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th annual ACM symposium on applied computing*. 459–464.

[9] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 437–442.

[10] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119.

[11] U. Naseem and K. Musial. 2019. DICE: Deep Intelligent Contextual Embedding for Twitter Sentiment Analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. Sydney, NSW, Australia, 953–958. https://doi.org/10.1109/ICDAR.2019.00157

[12] U. Naseem, I. Razzak, K. Musial, and M. Imran. 2020. Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis. *Future Generation Computer Systems* 113 (2020), 58–69. https://doi.org/10.1016/j.future.2020.06.050

[13] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2022. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 34 (2022). https://doi.org/10.1109/TKDE.2020.2992485

[14] J. F. Raisa, M. Ulfat, A. Al Mueed, and S. M. S. Reza. 2021. A Review on Twitter Sentiment Analysis Approaches. In *Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. Dhaka, Bangladesh, 375–379. https://doi.org/10.1109/ICICT4SD50815.2021.9396915

[15] S. Vanaja and M. Belwal. 2018. Aspect-Level Sentiment Analysis on E-Commerce Data. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. Coimbatore, India, 1275–1279. https://doi.org/10.1109/ICIRCA.2018.8597286

[16] Zhigang Yuan, Sixing Wu, Fangzhao Wu, Junxin Liu, and Yongfeng Huang. 2018. Domain attention model for multi-domain sentiment classification. *Knowledge-Based Systems* 155 (2018), 1–10. https://doi.org/10.1016/j.knosys.2018.05.004