An analysis on IMDB dataset

Chao-chun Hsu Vivian Lai Yichen Wang

Abstract

We perform a set of analyses on the IMDB dataset. We first get an overview of the dataset by performing a couple of analyses. In the second part of our work, we attempt to show directly, and indirectly how coefficients/importance of features may change classifiers' predictions. Finally, we explore sentiment trajectory in reviews.

1 Introduction

In this short paper, we discuss the types of analyses performed on the IMDB dataset [9]. In the first section, we understood the dataset better by performing the following: 1) generate word clouds of different grouped ratings; 2) investigate word percentage by word category; and 3) cluster content reviews by different topics. The second section show the inner workings of two models, namely Logistic Regression, and Random Forest. We then show the differences of actual and predicted values of the models through an interactive chart. Finally, we investigated the sentiment trajectory of the reviews.

2 Related Work

This section will give a brief review of related work for our task. In natural language processing community, there have been a lot of work on sentiment analysis [8, 11]. This technology is widely applied to different platforms e.g., Twitter [6], Yelp [5], Amazon [2], and IMDB movie review [15] which our project focuses on. In the beginning, researchers mainly explore the sentiment analysis on document-level which includes all content of a document. However, document-level label might not be true for every sentence in the article, so sentence-level analysis starts to get more attention in this community [10, 13]. With sentence-level annotations,it becomes possible to investigate the flow in a document. Tanevv et al. proposed a method to monitor the emotion trajectory of Ted talks that stimulate our interest to dig into sentiment trajectory of movie reviews [14].

3 Dataset Overview

We used the IMDB dataset created by Stanford AI lab [9]. Since this is a dataset for binary sentiment classification, ratings 5 and 6 were removed. There are 25 000 movie reviews for training, and 25,000 for testing.

3.1 Word clouds

The subsection describes the rationale behind using word clouds, and how they are generated. We used word clouds to show a bird's eye view of the dataset grouped by ratings. Since there are more than 5 different ratings i.e., 1 - 10, excluding 5 and 6, we decided to group two ratings in one group since content reviews may not differ by too much. In other words, ratings 1 and 2 are grouped together, then ratings 3 and 4, etc.

We used the Python package wordcloud to generate the word clouds. Before generating the word clouds, stop words are first removed from review content. The size of the word in the word cloud corresponds to the frequency of the word; the bigger the word, the higher it appears in the dataset.

The word clouds show that there is not a huge difference between different grouped ratings. However, the word "bad" appeared in 1, and the word "good" appeared in reffig:wordCloud5, suggesting that there is still a small difference between the word clouds. The most commonly used words across the grouped ratings are: movie, character, file, and br br. The last word "br br" is the line break element appearing at the end of each review, which should have probably been removed before generating the word clouds. We hypothesize that sentiment words do not appear as much in the word clouds as the frequency of those words are lower than the more commonly used words.

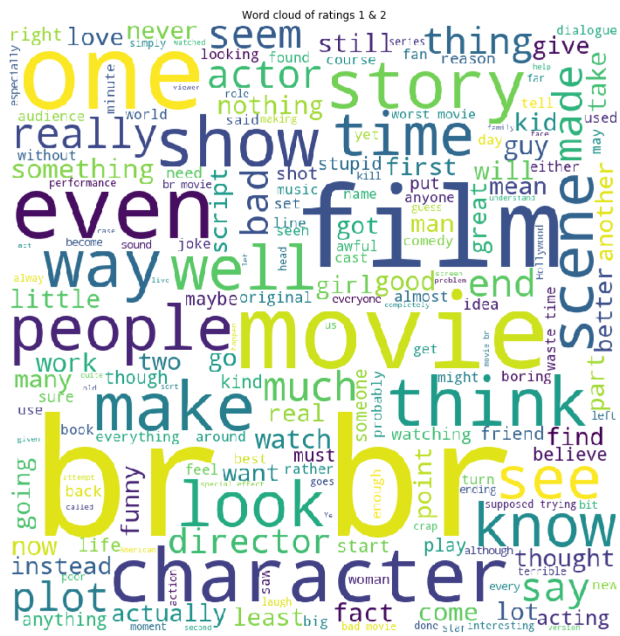

Figure 1: Word cloud for ratings 1 & 2

3.2 Word percentage by word category

To show the detailed word usage in the reviews, we compare and visualize the word usage of 5 different word categories (positive emotion, negative emotion, anger, sad and swear) for positive and negative reviews. The proportion of words in the reviews are analyzed using the Linguistic Inquiry and Word Count software (LIWC [12]), a word frequency-based text analysis tool.
To visualize the proportion of words and the comparison in positive and negative reviews, we use grouped bar chart, as shown in figure 5. Bar chart is a direct way to show the quantitative values using the height of bars. In addition, because the two review types (positive, negative) are grouped and arranged side-by-side, the bar clusters make easy to interpret the differences between them, and even among the 5 categories.

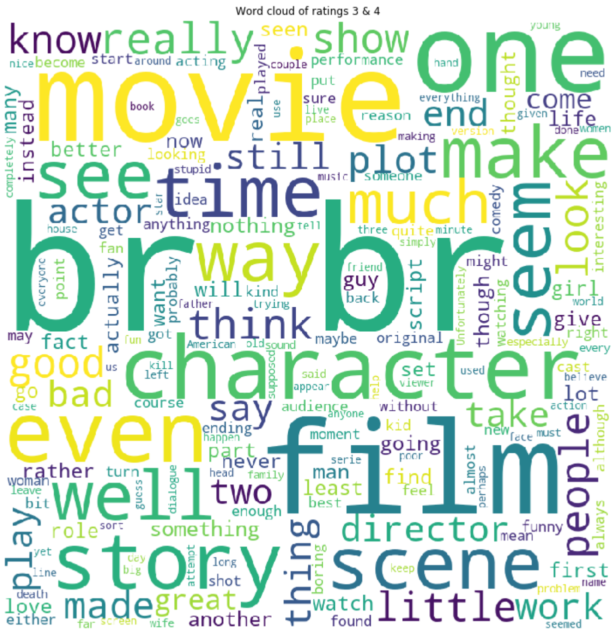

Figure 2: Word cloud for ratings 3 & 4

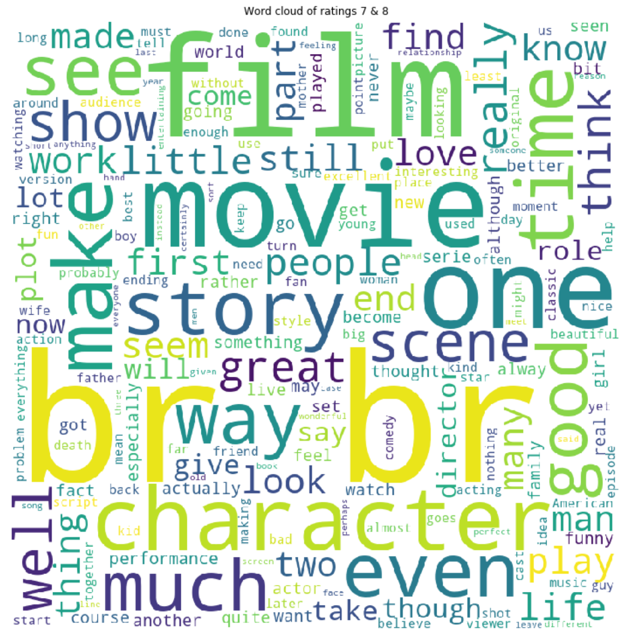

Figure 3: Word cloud for ratings 7 & 8

3.3 Cluster content reviews

Because this IMDB dataset has no category label, we want to explore the topics of reviews with unsuperivsed KMeans Clustering method. The first step is to convert reviews into TF-IDF features and then apply KMeans Clustering to form 20 clusters. We manually check top words of all clusters and finally select four cluster with clear topics to put on the bubble chart. As shown in Figure 6, the topics are family, romance, horror, and comedy. Each selected cluster has eight words that describe the specific topic.

4 Features and Classification Models

In this section, we attempt to show how directly and indirectly how features can affect the model's prediction. We do this by showing the top K feature weights of Logistic Regression, and discussing some important features of the Random Forest classifier. Lastly, we show false positives, true positives, false negatives, and true negatives values of each model through an interactive bar and pie chart.

4.1 Features and Models

Before training and testing the model, we first used the bag-of-words [3] model to extract features from our text data. Words from the dataset are stemmed and tokenized, subsequently the frequency of each token is calculated. For more details on bag-of-words, please read the implementation of Python Scikit-learn TfidfVectorizer.

The three models we used to perform the sentiment classification task is Logistic Regression, Random Forest [7], and LSTM with attention [1, 4]. The rationale behind using these models is the fact that each model represent a different type of classifier. The Logistic Regression represents a linear model, the Random Forest represents a tree model, and the LSTM with attention represents a non-linear model. For each model, we first trained and tuned the model to get the best accuracy using the validation set, the best parameters are used to test on the test set. The accuracy of the classifiers are as follow: 1) Logistic Regression: 88.6%; 2) Random Forest: 81.6%; 3) LSTM with attention: 90.6%. For source code, please check out our Github repository, each model has a Jupyter Notebook named after it.

4.2 Logistic regression with feature weights

We used logistic regression as out linear model to predict the review sentiment. After training, the weights of the logistic regression model can be a good representation of feature importance. To have a better understanding of the role of different features (words) in the prediction, we plotted the top 30 words with highest (and lowest) weights using bubble chart, which are the most important features determining if a review is positive or negative. For bubble chart, the size of bubble is a straightforward representation of the importance. We also used 2 different colors with large contrast for positive and negative words, as shown in figure 7, 8. The words are not beyond our expectation. Words like ”great”, ”perfect” have high weights and words like ”worst”, ”boring” have low weights.

4.3 Random forest with splitting process

We used random forest for our tree-based model, which is another widely used model in different machine learning tasks. To understand the decision process of the model, we selected a single decision tree from the forest and visualized the top 4 levels of it, as shown in figure 9. We used a collapsible tree which user can click the nodes to expand and collapse to visualize the splitting process. The blue nodes can be clicked to expand, and after expansion can be clicked to collapse back. We used this visualization is because it is not only a intuitive representation of a tree (in data structure), but its ”expand and collapse” interaction is a natural way to show the path of a tree.

4.4 Classifier predictions

In this subsection, we show the false positives, true positives, false negatives, and true negatives values of each model. The purpose is to show the difference between actual and predicted values between the models. When hovering over one of the bar charts that represent a particular model, the pie chart on the right shows the confusion matrix values for the particular model. Refer to 11 for more details. On the other hand, when hovering over one of the four slices of the pie chart, the bar chart on the left shows the different number of cases for each of the model. Refer to 10 for more details. For a better user experience on the interactive bar and pie chart, please use our website.

Figure 4: Word cloud for ratings 9 & 10

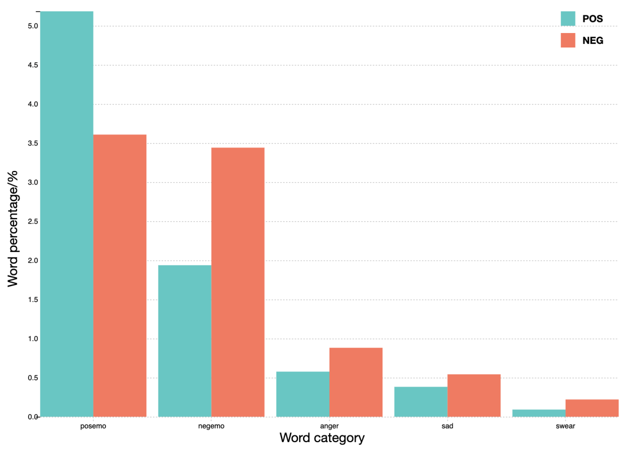

Figure 5: Grouped bar chart for word category comparison

5 Sentiment Trajectory

Inspired by Tanvee et al. [14] which analyzes the emotion trajectory in Ted talks, we try to represent the sentiment trajectory in IMDB movie review to explore the potential patterns. We first train a logistic regression classifier on whole review of training set, and then we test our model on each sentence of a review in testing set to obtain the probability of positive prediction. In this way, we end up with a list of probability for a review as sentiment trajectory. Since sentence numbers may vary across reviews, we first filter out review under twenty sentences and interpolate all sentiment trajectories of reviews to same length for next clustering step. As show in Figure 12, Agglomerative Clustering is introduced to cluster sentiment trajectories. With the clustering result, we find two clusters that represent sentiment trajectories of positive and negative reviews respectively. In Figure 13 that denotes the sentiment trajectories of a cluster having about 85% positive reviews, we can see an increasing trend in the line plot which means a review pattern that keep saying good words to a movie. On the other side, Figure 14 shows an interesting pattern having a peak in the beginning and going down in the end. This could be that when users want to criticize a movie,

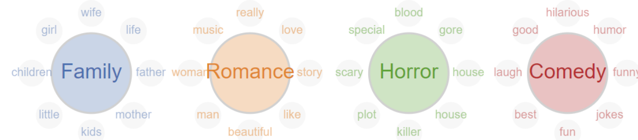

Figure 6: Topic Modeling by KMeans Clustering with TF-IDF feature

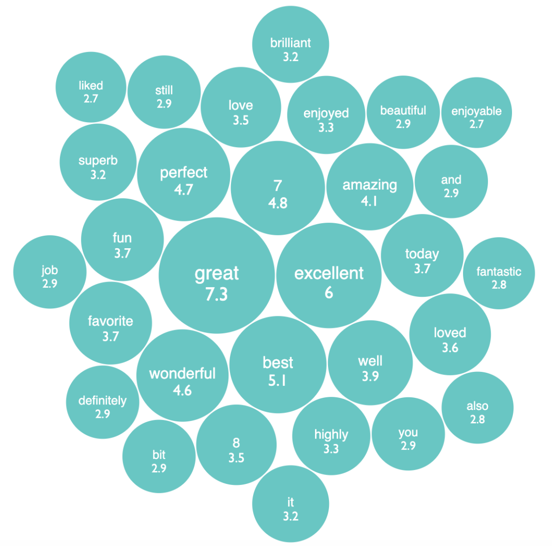

Figure 7: Top 30 words with highest weights

they will start with good words and then come up with a "BUT" in the review.

6 Conclusion

To conclude, we performed a couple of analyses using different methods on the IMDB dataset to understand it better. In first and second sections, we discovered that although we are not able to find the significant different from word clouds of reviews with low rating and high rating, the visualizations of logistic regression and random forest show that models still make decision depending on positive words and negative words. For sentiment trajectory section, we found an interesting pattern of negative reviews that users will start with positive evaluation and then end up with negative criticism.

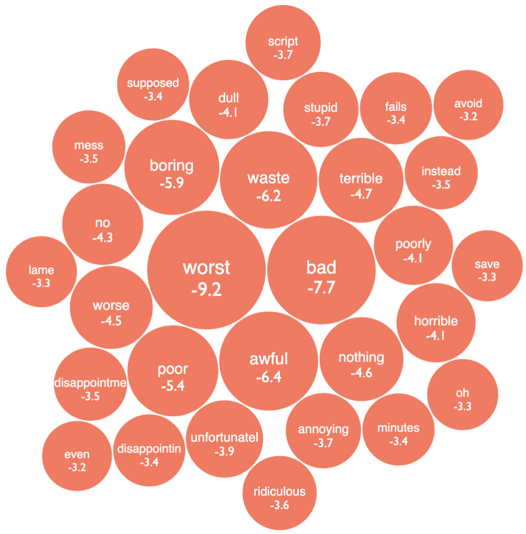

Figure 8: Top 30 words with lowest weights

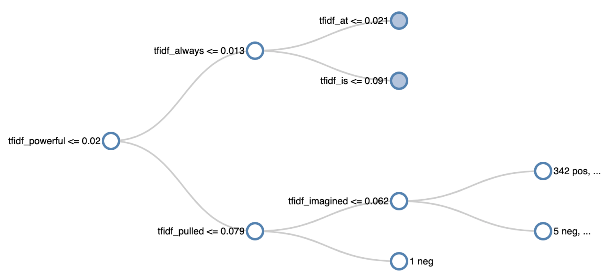

Figure 9: Decision tree from random forest (the top 4 levels)

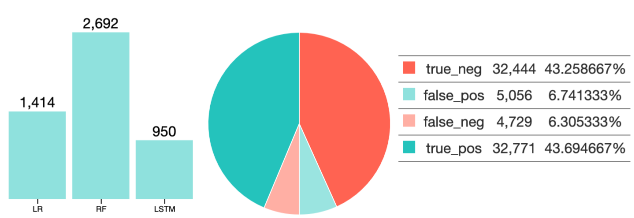

Figure 10: Grouped by confusion matrix values

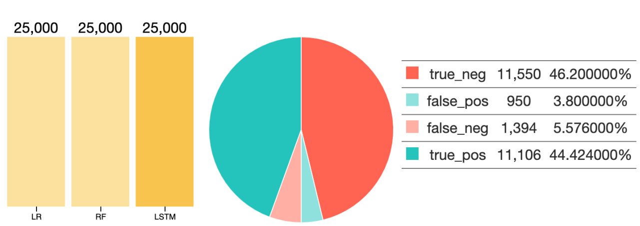

Figure 11: Grouped by model

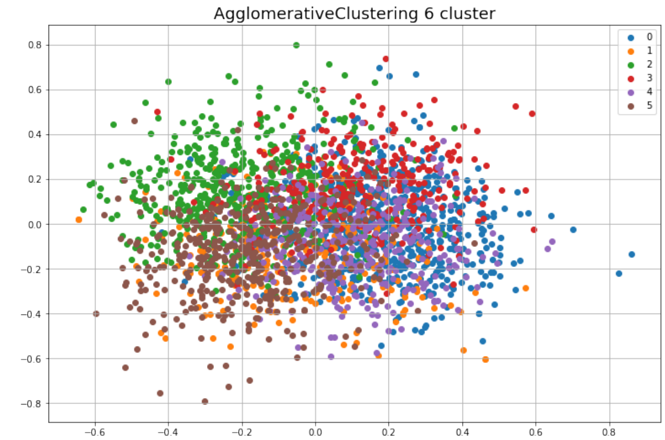

Figure 12: Agglomerative Clustering of sentiment trajectories.

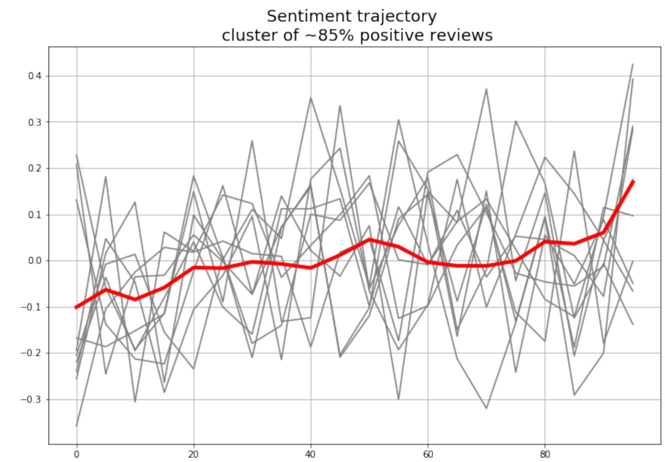

Figure 13: Line char of the cluster that has $\tilde{8}$5% sentiment trajectories of positive reviews. Red line means the mean of all sentiment trajectories in this cluster.

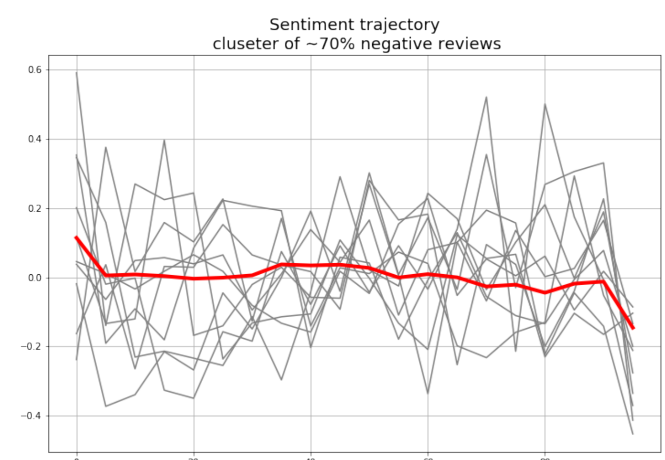

Figure 14: Line char of the cluster that has $\tilde{7}$0% sentiment trajectories of negative reviews. Red line means the mean of all sentiment trajectories in this cluster.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] A. Bhatt, A. Patel, H. Chheda, and K. Gawande. Amazon review classification and sentiment analysis. *International Journal of Computer Science and Information Technologies*, 6(6):5107–5110, 2015.

[3] Z. S. Harris. Structural linguistics. 1963.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442, 2014.

[6] E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*, 2011.

[7] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[8] B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[9] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA, June 2011.

[10] A. Meena and T. Prabhakar. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *European Conference on Information Retrieval*, pp. 573–580. Springer, 2007.

[11] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[12] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[13] A. Shoukry and A. Rafea. Sentence-level arabic sentiment analysis. In *2012 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 546–550. IEEE, 2012.

[14] M. I. Tanveer, S. Samrose, R. A. Baten, and M. E. Hoque. Awe the audience: How the narrative trajectories affect audience perception in public speaking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 24:1–24:12. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173598

[15] T. T. Thet, J.-C. Na, and C. S. Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848, 2010.