

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 the natural language processing community, there has been much work done on sentiment analysis [8, 11]. This type of analysis is widely applied to different platforms e.g., Twitter [6], Yelp [5], Amazon [2], and IMDB movie reviews [15], which our project focuses on. In the very beginning, researchers mainly explored sentiment analysis only on document-level; in other words, all content of a particular document. Since document-level label might not be true for each sentence in the document, sentence-level analysis began to gain more attention in the community [10, 13]. With sentence-level annotations, it became possible to investigate the entirety of a document. Tanev et al. proposed a method to monitor the emotion trajectory of Ted talks, which sparked our interest to investigate on 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.

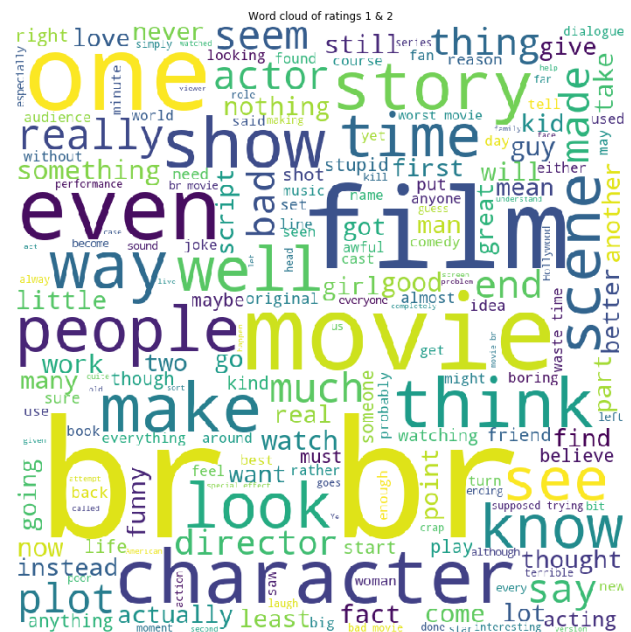
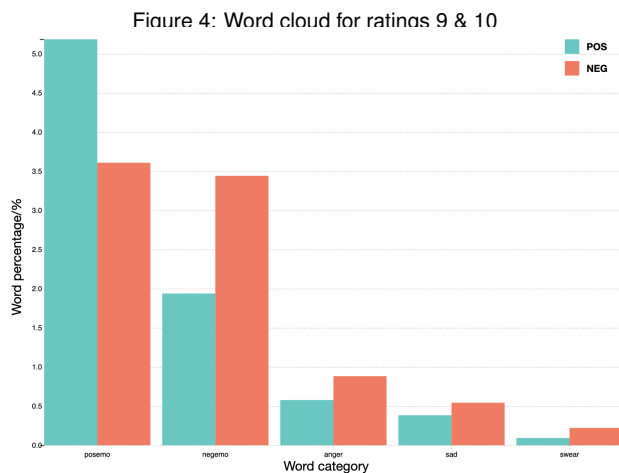


Figure 1: Word cloud for ratings 1 & 2

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.



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.

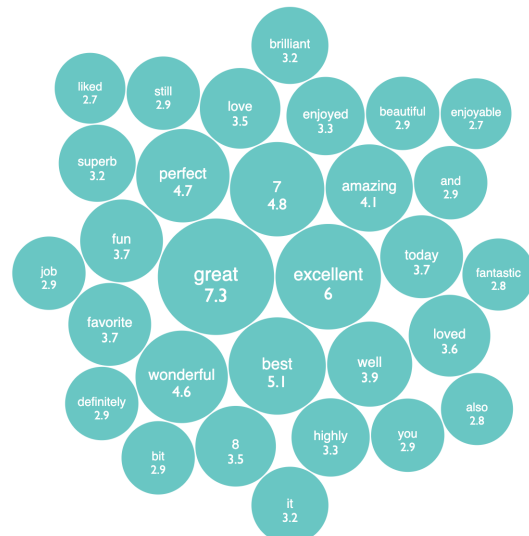
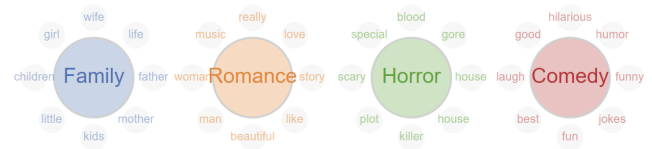


Figure 7: Top 30 words with highest weights

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.

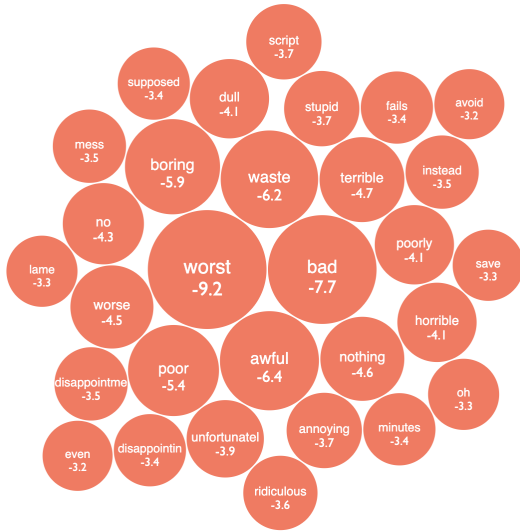


Figure 8: Top 30 words with lowest weights

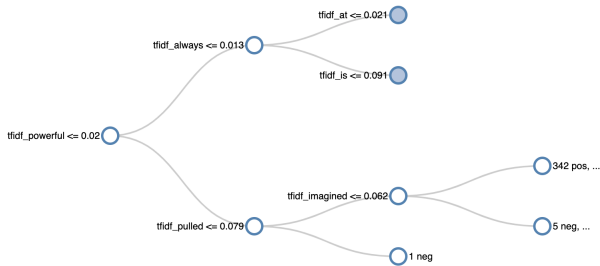


Figure 9: Decision tree from random forest (the top 4 levels)

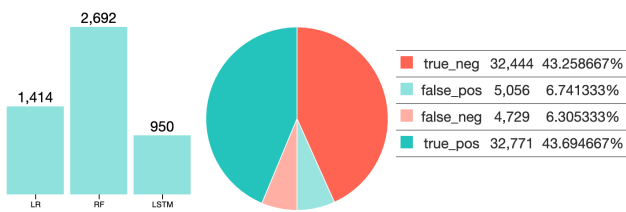


Figure 10: Grouped by confusion matrix values

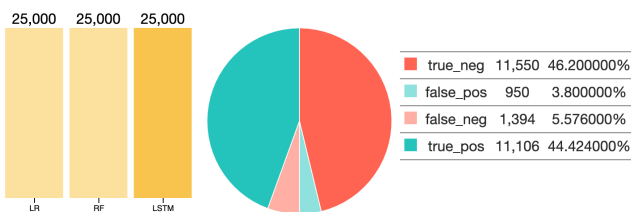


Figure 11: Grouped by model

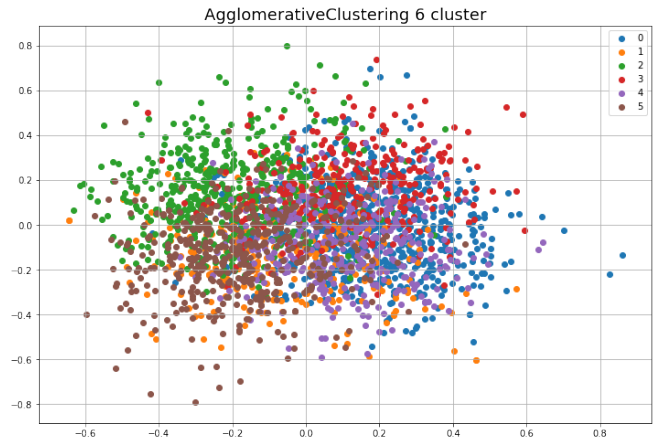


Figure 12: Agglomerative Clustering of sentiment trajectories.

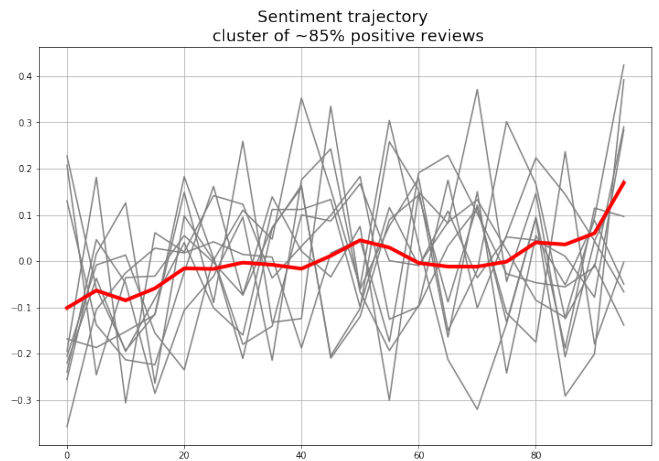


Figure 13: Line chart of the cluster that has ~85% sentiment trajectories of positive reviews. Red line means the mean of all sentiment trajectories in this cluster.

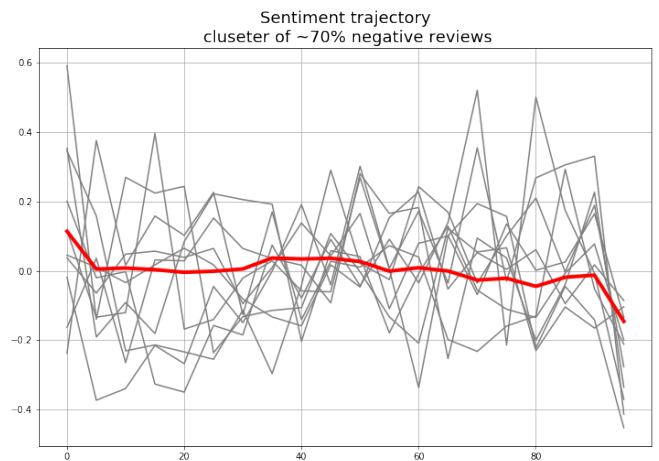


Figure 14: Line chart of the cluster that has ~70% sentiment trajectories of negative reviews. Red line means the mean of all sentiment trajectories in this cluster.

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.