Classifying Movie Reviews into the Six Ekman Emotions

Victor Verma
Boston University
vpverm@bu.edu

Abstract

1 Introduction

I plan to determine the most prevalent emotions expressed in user movie reviews for 65 of the top 250 highest rated movies on Letterboxd. Specifically, I will consider Ekman's six basic emotions: happiness, sadness, anger, fear, disgust, and surprise **[CITATION NEEDED HERE]**. To collect the source data, I will write a python script that scrapes the reviews (several thousand for each movie) from Letterboxd. In total, I expect to aggregate about 78,000 total reviews for analysis, and I can extend to a larger set of movies if more reviews are needed.

Although excellent text-to-emotion models already exist, I will make my own from scratch. This will be created by training a model using supervised learning on a large text dataset labeled with the six basic emotions. I aim to experiment with logistic regression, support vector machines, random forests, and neural networks, and use the model that performs the best. After identifying the most effective text-to-emotion classification model, I intend to apply it to the unlabeled dataset of movie reviews. As someone who loves movies, I am curious to learn if there is a common set of emotions that audiences express after watching highly rated movies.

There will be two parts to my final report. The first section will focus on evaluating the effectiveness of the emotion classification models that I will have created based on the standard metrics of accuracy, precision, recall, and F1 score. Here I will also try to account for any significant differences in model performance and any potential limitations. The second section will interpret the results of the emotional classification of the Letterboxd movie reviews. This will include summarizing simple statistics, such as the frequencies of each emotion, as well as identifying less obvious patterns that might emerge when grouping emotions together or stratifying by movie metadata (e.g. release year). For example, I am interested to see if there are traditionally unrelated genres that generate similar emotional reactions.

Although the text-to-emotion classification methods are not novel, their applications to Letterboxd movie reviews will draw significant interest. I actively maintain a movie recommendation website that generates recommendations using machine learning based on a user's Letterboxd profile, and has accumulated nearly 1,800 users to date. I intend to present my findings on the website since there is so much topical overlap, and the results may enable users to better understand their own movie-watching identity.

2 Data

2.1 Letterboxd Movie Reviews

The unlabeled movie review data was sourced from Letterboxd's Official Top 250 Narrative Feature Films. Although the list contains 250 movies, data was only used from the 65 that I had previously seen. The official list is updated weekly, but the collected data is from February 24th, 2024.

For each movie, the first 1,200 reviews sorted in order of decreasing activity were scraped. Each individual review was organized as a JSON object containing the author's rating of the movie, the date of the review, the number of comments left on the review, the review text itself, and the number of likes the review received. These were aggregated in a CSV with the JSON fields as columns.

Several steps were taken to preprocess the

movie review data. First, all of the None and NaN reviews were dropped. Next, the emojis were converted into their text equivalent, enclosed by a colon on each side. Non-English language reviews were then removed for convenience, as well of those solely consisting of punctuation. Finally, unnecessary whitespaces and newline characters were deleted to minimize the length of each review.

The processed dataset contained 65,803 movie reviews, with each receiving 153 likes and 3 comments on average. The distribution of ratings given by the user to the movie being reviewed, measured on scale of 0.5 - 5.0, is summarized in Table 2.1.1. From the counts, it can be seen that most of the ratings associated with reviews were high, which was expected given that the source movie data is a list of highly rated movies. This led to the baseline expectation that positive emotions would likely be expressed most often in the dataset.

Rating	Number of Reviews	Percentage
None	1,858	2.82%
0.5	400	0.61%
1.0	237	0.36%
1.5	132	0.20%
2.0	366	0.56%
2.5	471	0.72%
3.0	1,209	1.83%
3.5	2,329	3.54%
4.0	8,586	13.05%
4.5	13,589	20.65%
5.0	36,321	55.20%

Table 2.1.1: Distribution of Movie Ratings

2.2 Twitter Emotion Corpus

The labeled emotion data was sourced from the Twitter Emotion Corpus (TEC). The raw .txt file was parsed into a CSV with the text and emotion label as columns, and preprocessed in a similar manner as the movie review data. First, the emojis were converted into their text equivalent, enclosed by a colon on each side. Next, non-English language reviews were removed for convenience, as well of those solely consisting of punctuation. Finally, unnecessary whitespaces and newline characters were deleted to minimize the length of each review.

The processed dataset consisted of 19,333 text-label pairs, and the distribution of labels is shown in Table 2.2.1. It would have been ideal to have an approximately equal share of samples for each emotion label, but no such dataset was found.

Emotion	Count	Percentage
Happiness	7,892	40.82%
Sadness	3,613	18.69%
Anger	1,489	7.70%
Fear	2,563	13.28%
Disgust	733	3.79%
Surprise	3,043	15.74%

Table 2.2.1: Distribution of Emotion Labels

3 Baseline

As a starting point, an Ekman Classifier (EC) model from HuggingFace was used for emotion inference. It was built using the BertForMultiLabelClassification model architecture and finetuned using Google's GoEmotions dataset. The model supported labeling the standard Ekman emotions, plus "neutral".

To get an initial baseline, the EC model was evaluated upon the processed TEC dataset. Labels predicted as "neutral" were discarded because the label is not present in the TEC dataset. Surprisingly, the model only had an accuracy rate of 43.94%, much lower than expected from a pre-trained model.

The precision, recall, and F1-score are summarized in Table 3.1. Only two out of six emotions, fear and sadness, had a precision score greater than 0.5. This suggested that the baseline model produced a large number of false positives. Only one emotion, happiness, had a recall score greater 0.5, suggesting that the baseline model produced a large number of false negatives too. Consequently, only happiness had an F1-score greater than 0.5, indicating that the baseline model was objectively not very good at classifying the Ekman emotions.

The failures of the EC model were further investigated using the confusion matrix displayed in Figure 3.2, which allowed for insight into how emotions were being mislabeled. For example, it

Emotion	Precision	Recall	F1-Score
Happiness	0.47	0.74	0.58
Sadness	0.56	0.31	0.40
Anger	0.28	0.42	0.34
Fear	0.76	0.21	0.33
Disgust	0.26	0.09	0.14
Surprise	0.26	0.22	0.24

Table 3.1: EC Model Summary Statistics

was seen that happiness was commonly mistook for surprise, which potentially makes sense because there can sometimes be overlap between the two emotions. On the other hand, happiness and sadness were sometimes confused for each other, which was not expected since they are semantically opposite. Interestingly, the model had a hard time identifying anger and fear, with somewhat uniform mislabel distributions, which was surprising because one might consider these emotions to read as relatively distinct in text. The EC model served as a good starting point, but it did not work as well as expected.

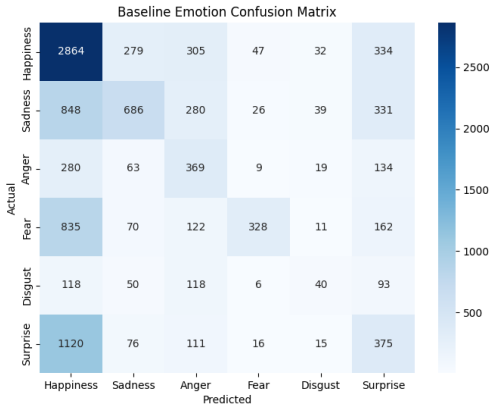

Figure 3.2: EC Model Confusion Matrix

4 Models

5 Application

6 Limitations

7 Acknowledgements

8 References