IST 664 Natural Language Processing Final Project Report

Sentiment Classification of Movie Reviews

Mark Stiles, Logan Roach

Introduction

In the information age, the proliferation of online platforms becoming an increasingly large part of an individual’s life has increased allowing them to share opinions and perspectives about various subjects including movies. This decentralization of opinion has generated a diverse array of perspectives on social media, blogs and dedicated movie review websites signaling a large departure from before where movie reviews were strictly in the domain of professional critics. Now, individuals can access large reservoirs of opinions and perspectives allowing for a comprehensive understanding of public sentiment towards a film.

Sentiment analysis, a subfield within Natural Language Processing, allows for a systematic approach to determining sentiment within textual data. In regard to movie reviews, sentiment analysis allows movie reviews to be classified into positive or negative sentiment providing insight into the public audience’s reception to a film. This is important for various stakeholders in the film industry and public audience. Filmmakers can gauge audience reactions while consumers can make more informed decisions about which movies to direct their attentions towards. Sentiment analysis also allows for the understanding of broad social and cultural trends in addition to providing direct information to stakeholders.

In this study, various routes of sentiment analysis were explored to classify sentiment within Stanford’s largest movie dataset of 25,000 reviews. Within the data, half are labeled as positive, and half are labeled as negative. Each review is an open text block about an unnamed movie found within a folder prescribing the positive or negative label. After data preparation and formatting, many experiments were conducted to provide the most accurate way to classify sentiment across the large repository of reviews.

Methodology

All aspects of the study including tokenization, cleaning the data, building the feature sets, and evaluating the experimental models were written to be easily replicated and iterated throughout the experiments or for future experiments. The data was first prepared by taking the text blocks and organizing the labeled text into a data frame for further use downstream in the analysis. Functions were defined to process and clean the data such as tokenizing, removing stop words, and removing unnecessary punctuation. Features were then produced in the notation of the Natural Language Tool Kit such as a “bag-of-words” feature set where the data is represented by counting the frequency of the words in each document, and part of speech tags where words are tagged according to their grammatical categories. Other features generated include a word capitalization count, a review length feature, an average word length feature, bigrams, the number of nots in a review, the number of negations in a review, a VADER feature set using the Vader Sentiment Intensity Analyzer, and a TF-IDF (Term Frequency-Inverse Document Frequency) feature set generated from a TF-IDF tokenizer function. Each of these feature sets was used separately or in conjunction with other sets within the experiments to find the best performing methods within the study. The NLTK Naïve Bayes classifier was used with cross-validation to judge the performance of the experiments by obtaining the precision, recall, F1 score, accuracy and an overall mean accuracy across experimental rounds to get the most comprehensive understanding of the results of each experiment.

Experiments

Experiment 1 was the baseline for the study using the bag-of-words feature set and all tagged parts of speech categories. Experiments 2 and 3 build off Experiment 1 by removing varying amounts of text based on parts of speech tags to identify if there are any relationships between using tagged parts of speech and the performance of the classification. Experiment 2 iterates through each tagged part of speech and removes one tag every iteration to investigate the effect one tagged part of speech has on the performance of the classification. If the performance drops after removing a part of speech, then its assumed that part of speech has a positive impact on classifying the sentiment of the reviews, and conversely, if the performance improves then its assumed that the part of speech has a negative impact on the classification. Experiment 3 uses a broader approach and removes all parts of speech tagged in the bag of words feature set where the only features left were not tagged as a part of speech to see if including tagged parts of speech reduces the model’s performance. Experiment 4 moves away from the bag of words feature set and creates a feature set using the frequency of the tagged parts of speech within the review to classify sentiment investigating if the volume of parts of speech give indication of positive or negative sentiment. Experiment 5 uses a feature set combining a feature set of the number of word capitalizations within a review, the average word length in each review and the average word length of each review for a comprehensive look at how these text statistics effect the performance of the classifier.

Experiment 6 uses bigrams generated from the review corpus as a feature set to see if the inclusion of bigrams results in any movement regarding the classification performance. Experiment 7 is similar to Experiment 8 where the focus is the investigation of negation within the reviews. Experiment 7 uses a feature set counting the number the word “not” appears in each review to see if the focus shares any insight into the classification performance. Experiment 8 uses a feature set of contradicted words to investigate if the negations within a review are a good barometer for classifying the sentiment of the reviews. In this case not all words that were contradicted were included on the most common words by frequency to determine if each review contained a negated or contradicted version of that word. Experiment 9 is similar to Experiment 1-3 where a bag of words feature set is used except with TF-IDF tokenized words. The experiment should be ideal for investigating how tokenization affects the performance of classification. Experiment 10 uses the VADER third-party sentiment library to generate features with a sentiment value to investigate if using features with preassigned sentiment would improve the performance of the classification. Experiment 11 and Experiment 12 combine feature sets from the top performing experiments to see if an improved performance can be seen when classifying the sentiment. Experiment 11 combines the Vader features, TF-IDF features, and the bag of words feature set from Experiment 1 where all words were included. Experiment 12 combines Vader features, TF-IDF features, and the bag of words feature set but with nouns removed from the tagged parts of speech.

Results and Analysis

In Experiment 1, all parts of speech were included in the bag of words feature set of 25,000 reviews. The accuracy ranges from 81.86% to 83.88% across the five rounds. The performance of each round varies slightly across all five rounds where there is a slight dip in the third round but then improves again in the following round of experimentation. The most indicative features used to distinguish sentiment remain consistent across all rounds where words like "pointless," "laughable," and "waste" are strongly associated with negative sentiment. Overall, the mean accuracy across all rounds was 83.14% suggesting the model is reasonably accurate in sentiment classification, but with room for improvement.

In Experiment 2, the model’s performance is evaluated with one part of speech removed from the bag of word features during each iteration. The parts of speech removed for each iteration in order are conjugation, determiner, noun, verb, adjective, adverb, particle, marker, numerical, foreign words, symbols, interjection, to, ex (existential there) and pos (posessive endings).

When removing conjugation parts of speech, accuracy ranges from 82.08% to 84.08% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment "pointless," "laughable," "redeeming," and "worst" are strongly associated with negative sentiment. Overall, the mean accuracy is across all rounds 83.28% suggesting the model is reasonably accurate in sentiment classification when removing conjugation parts of speech but does not show significant signs of improvement.

When removing the determiner parts of speech, accuracy ranges from 81.74% to 83.96% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment "pointless," "laughable," and "worst" are strongly associated with negative sentiment. Overall, the mean accuracy is across all rounds 83.09% suggesting the model is reasonably accurate in sentiment classification when removing determiner parts of speech with comparable results to the model’s results when including all parts of speech.

When removing noun parts of speech, accuracy ranges from 83.7% to 85.6% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment are "3/10," "4/10," and "7/10”, which are assumed to be ratings out of ten with the lower scoring ratings corresponding to negative sentiment and the higher rating corresponding to positive sentiment. Overall, the mean accuracy is 84.61%, which is the highest in this experiment, suggesting when removing noun parts of speech is reasonably accurate in sentiment classification but only slightly improved from using all parts of speech.

When removing verb parts of speech, accuracy ranges from 81.16% to 83.5% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment are "unfunny," "pointless," and "10/10" where the first two are strongly associated with negative sentiment and the last strongly associated with positive sentiment. The mean accuracy is 82.46% following the same trend as before where the model is reasonably good at classifying sentiment but with no improvement when including all parts of speech.

When removing adjective parts of speech, accuracy ranges from 79.9% to 81.58% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment are "10/10," "pointless," and "waste," where the first is associated with positive sentiment and the last two are strongly associated with negative sentiment. Overall, the mean accuracy is 80.86% which is the lowest accuracy in the experiment, still high suggesting a reasonable performance by the model, but slightly lower performance when using all parts of speech.

When removing adverb parts of speech, accuracy ranged from 81.5% to 83.6% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment are "pointless," "laughable," and "redeeming" which are strongly associated with negative sentiment. With an overall mean accuracy of 82.64% suggests the model is reasonably good at classifying sentiment, but still does not perform as well when including all parts of speech.

When removing particle parts of speech, accuracy ranges from 81.86% to 83.88% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment "pointless," “laughable," and "waste" are strongly associated with negative sentiment. The mean accuracy is 83.14% showing the model is reasonable at classifying sentiment, but with no improvement when including all parts of speech.

When removing marker parts of speech, accuracy ranges from 81.86% to 83.88% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment "pointless," "laughable," and "waste" are strongly associated with negative sentiment. The mean accuracy is 83.14% again suggesting the model is reasonable at classifying sentiment, but with no improvement when including all parts of speech.

When removing numerical parts of speech, accuracy ranges from 81.86% to 83.96% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment were again “pointless," "laughable," and "waste" which were strongly associated with negative sentiment. The mean accuracy is 83.12% showing the model is reasonable at classifying sentiment, but with no improvement when including all parts of speech.

When removing foreign parts of speech, accuracy ranged from 81.86% to 83.88% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment were again “pointless," "laughable," and "waste" which were strongly associated with negative sentiment. The mean accuracy is 83.14% showing the model is reasonable at classifying sentiment, but with no change when including all parts of speech.

When removing symbol parts of speech, accuracy ranges from 81.86% to 83.88% across the five rounds. The most indicative features used to distinguish sentiment were again “pointless," "laughable," and "waste" which were strongly associated with negative sentiment. The mean accuracy is 83.14% showing the model is reasonable at classifying sentiment, but with no improvement when including all parts of speech.

When removing interjection parts of speech, accuracy ranges from 81.84% to 84.0% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment were again “pointless," "laughable," and "waste" which were strongly associated with negative sentiment. The mean accuracy is 83.14% showing the model is reasonable at classifying sentiment, but with no change when including all parts of speech.

When removing to parts of speech, accuracy ranges from 81.84% to 84.02% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment were again “pointless," "laughable," and "waste" which were strongly associated with negative sentiment. The mean accuracy is 83.14% showing the model is reasonable at classifying sentiment, but with no change when including all parts of speech.

When removing ex parts of speech, accuracy ranges from 81.82% to 83.86% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy but vary slightly across rounds. The most indicative features used to distinguish sentiment were again “pointless," "laughable," and "waste" which were strongly associated with negative sentiment. The mean accuracy is 83.14% showing the model is reasonable at classifying sentiment, but with no change when including all parts of speech.

When removing pos parts of speech, accuracy ranges from 81.84% to 83.92% across the five rounds. The most indicative features used to distinguish sentiment were again “pointless," "laughable," and "waste" which were strongly associated with negative sentiment. The mean accuracy is 83.14% showing the model is reasonable at classifying sentiment, but with no improvement when including all parts of speech.

In Experiment 3, all tagged parts of speech were removed from the bag of words feature set. The accuracy ranges from 58.86% to 60.26% dropping significantly from the previous experimental results. The precision, recall, and F1 scores are similar with respect to accuracy. The most indicative features used to distinguish sentiment were "winchester," "hanzo," and "zombi" where the first two are associated with positive sentiment and the last associated with negative sentiment. The mean accuracy is 59.53% significantly lower than all parts of speech included in the bag of word features, and lower than removing just one part of speech tag at a time.

In Experiment 4, the features set was composed of counts of tagged parts of speech including conjugations, determiner parts of speech, nouns, verbs, adjectives, adverbs, particles, markers, numerical parts of speech, foreign words, symbol parts of speech, and interjection parts of speech. The accuracy ranges from 52.4 % to 55.48% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy varying slightly across rounds, but the F1 score for classifying negative sentiment was slightly higher than classifying positive sentiment. This suggests the model is more precise and has a greater recall classifying negative sentiment rather than positive sentiment. The most indicative features used to distinguish sentiment were "verb\_count" associated with positive sentiment, "determ\_count" associated with negative sentiment, “noun\_count” associated with positive sentiment. This suggests the model finds reviews with high number of verbs and noun parts of speech to be associated with positive sentiment and higher determiner parts of speech to be associated with negative sentiment. The mean accuracy was accuracy across all rounds is approximately 54.74%, showing that removing noun parts of speech on the bag of words feature set still has a better performance when classifying sentiment.

In Experiment 5, the feature statistics were used to generate the feature set. These statistics are composed of capitalized word count, review length, and an average word length. The accuracy ranges from 51.38% to 52.76% across the five rounds. Again, this range drops from the initial set of experiments but the precision, recall, and F1 scores are similar with respect to accuracy. Again, the F1 score for classifying negative sentiment was slightly higher than classifying positive sentiment suggesting the model is more precise and has a greater recall classifying negative sentiment rather than positive sentiment. The most indicative features used to distinguish sentiment were “review\_length” followed by “cap\_count” where each is not specifically associated with positive or negative sentiment, but the model identified them as determinative when classifying the reviews. The overall accuracy of all rounds is 51.87% decreasing from the previous experiment, and only slightly better than a coin flip.

In Experiment 6, a bigram finder is used to find the top bigrams iterated across all reviews. The bigrams are then used as features to classify within the model. The accuracy ranges from 48.88% to 49.98% across the five rounds. In an unusual case, one class of sentiment had perfect precision and the other had a precision of 0.0 where the class depended on the round of the experiment. Where the model precision was perfect, the recall hovered around 59% and the F1 score was higher than the accuracy score found usually around 66%. When the precision of one of the sentiment classes was 0.0 the recall and the F1 score were also 0.0. This could be because the bigrams were comprised of only negated bi-gram features or positive bigram features. The most indicative features were found to be these negated bigram features where the model determined them to have a strong association with negative sentiment implying the model with this feature set is biased towards predicting instances as negative. It’s important to remember, the mean accuracy is 49.33%, indicating that the model's performance is not much better than random chance.

In Experiment 7, the number instances of the word “not” were counted in each review across both sets of reviews to generate a feature set used to classify sentiment. The accuracy ranges from 57.74% to 59.12% across the five rounds. The precision, recall, and F1 scores are similar with respect to accuracy varying slightly across rounds, but the F1 score for classifying negative sentiment was slightly higher than classifying positive sentiment. This suggests the model is more precise and has a greater recall classifying negative sentiment rather than positive sentiment. This logically follows when creating a list of features comprised of the number of times “not”, a negation word, is used across reviews. The most indicative features were the number of “not\_counts” found within the reviews where they were all associated with negative sentiment. The overall mean accuracy 58.33% was higher than the previous experiment, but still not as high as the initial experiments removing noun parts of speech and the bag of words feature set.

In Experiment 8, the feature set used to classify sentiment is composed of top words preceded by the word “not” across the global word distribution set. The accuracy ranges from 48.88% to 49.98% across all five rounds. Again, the behavior of the precision, recall and F1 score results are similar to the bi-gram features where only one class of sentiment had perfect precision and the other had a precision of 0.0. Where the model precision was perfect, the recall hovered around 49%, and the F1 score was about 66%. When the precision of one of the sentiment classes was 0.0 the recall and the F1 score were also 0.0. The same phenomenon that occurred when using the bigram feature set suggesting sets could be composed of wholly negated pairs of top word or perhaps double negated words when the positive class precision is 100% accurate. There are no indicative features that include double negation, but they do include features like “not\_&” and “not\_the” which are associated with negative sentiment. These results suggest the model is better at classifying negative sentiment. The mean accuracy was found to be 49.33%, the same when using the bigram feature set and not improving from previous experiments.

In Experiment 9, the TF\_IDF vectorizer is used to compute the top terms while considering their frequency across the entire corpora. This feature set is then used to classify sentiment where the accuracy ranges from 80.86% to 82.68%. The precision, recall, and F1 scores are similar with respect to accuracy with the F1 scores for positive sentiment classification being slightly higher than the negative sentiment classification. This suggests the slightest favorability when classifying positive sentiment using the model when using this feature set. The indicative features used to distinguish sentiment were "waste," "pointless," and "worst" which are strongly associated with negative sentiment. The mean accuracy across all rounds was 82.14% which is much better than the previous experiments but still slightly worse than the first experiment using all parts of speech and the bag of words feature set.

In Experiment 10, the feature set used to classify across was generated using Vader Sentiment Intensity Analyzer. Everything else being equal, the accuracy ranges from 69.08% to 70.54%. The precision, recall, and F1 scores are similar with respect to accuracy but there were slight differences between classes. For example, for the negative class, precision ranges from approximately 0.544 to 0.554, recall ranges from approximately 0.778 to 0.793, and F1-score ranges from approximately 0.640 to 0.649. For the positive class, precision ranges were higher from approximately 0.840 to 0.862, the recall scores were lower from approximately 0.643 to 0.663, and the F1 scores were thus higher from approximately 0.728 to 0.748. This could indicate a slight favoritism towards classifying positive sentiment when using the model with this feature set. The most informative features where sentiment analyzed features were either present or not. Interestingly, the absence of sentiment (S\_vader = False) is more indicative of the negative class, while the presence of sentiment (S\_vader = True) is more indicative of the positive class. The mean accuracy across all rounds was found to be 69.92%, showing a moderate performance but not as good as performance from Experiment 1 or Experiment 2.

In Experiment 11, multiple feature sets were combined including the Vader feature set, the TF\_IDF feature set and the bag of words feature set including all parts of speech. The accuracy ranged from 82.9% to 84.72% across all five rounds. We see similar results to the previous experiment in terms of the precision, recall and F1 scores to be slightly higher in the positive class than the negative class but with less of a difference between the two. The negative class precision ranges from approximately 0.812 to 0.840, recall ranges from approximately 0.839 to 0.866, and F1-score ranges from approximately 0.831 to 0.844. The positive class precision ranges from approximately 0.854 to 0.870, recall ranges from approximately 0.819 to 0.827, and F1-score ranges from approximately 0.836 to 0.850. The most informative features were a combination of features from the merged list of features. For example, TF\_DIF vectorized features like "TF\_waste" and sentiment features like "V\_pointless" were found to be associated with negative sentiment. The mean accuracy across all rounds was found to be 83.92% making this model just slightly outperform the performance seen in Experiment 1.

In Experiment 12, multiple feature sets were combined including the Vader feature set, the TF\_IDF feature set and the bag of words feature set with the nouns removed from the parts of speech included. This was done to build off the promising results of the previous experiment and using results from experiment 2 where nouns were removed from the bag of word features improved the performance. The accuracy ranged from 84.02% to 85.54% across all five rounds. The same pattern emerges in regard to the previous experiment’s precision, recall and F1 scores being slightly higher in the positive class than the negative class. The negative class had precision ranges from approximately 0.811 to 0.851, recall ranges from approximately 0.852 to 0.877, and F1-score ranges from approximately 0.838 to 0.858, and the positive class had precision ranges from approximately 0.866 to 0.881, recall ranges from approximately 0.827 to 0.858, and F1-score ranges from approximately 0.852 to 0.865. This indicates a slight favoritism towards the positive class sentiment in terms of classifying the reviews, but again very slight. The most informative features appear to be numerical ratings like in the experiment where nouns were removed previously. The same features appear "3/10," "4/10," and "7/10”, with the lower scoring ratings corresponding to negative sentiment and the higher rating corresponding to positive sentiment. The mean accuracy across all rounds is found to be 85.07%, the most promising results from the model across all experiments and iterations.

Conclusion

Through all experiments a prediction accuracy of 85% was the highest achieved. It was through Experiment 12 which combined the features sets with the overall highest accuracy individually. These features were Vader, TF-IDF and Bag-of-Words with nouns removed. The improvement was modest, 1-2 points from using the feature sets as standalone feature sets. Though multiple experiments showed suboptimal results, they provided many different insights into what kinds of features would suit the classification.

With more time other avenues could've been attempted to mix feature logic or spend more time within the review text to identify other aspects of the corpus that indicated movie sentiment with better accuracy. Overall, the results were respectable, but we wished we could've designed features that reached at least 90% accuracy.