HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Movie Review Sentiment Analysis: Supervised Learning Versus Large Language Model

Natalia Kochut

Rye Country Day School, USA

ABSTRACT

Sentiment analysis is frequently used to derive insights from natural language. Examples include analysis of textual data to measure brand perception, social media trends, or customer opinion about products. This paper evaluates the performance of three supervised machine learning methods and compares them with the next-generation large language model (LLM), which recently gained popularity with the release of OpenAI ChatGPT. Specifically, we explore the application of Decision Tree, Random Forest, and Support Vector Machine classifiers to a representative sample of 100K movie reviews collected by a well-known website, IMDb.com. Reviews are tagged with numeric ratings, allowing the formulation of a supervised learning problem and exploring the ability to differentiate sentiment between strongly opinionated positive and negative reviews and also, a more challenging problem of differentiating between weakly opinionated positive and negative reviews. Models are tuned to optimize recall and precision in this application, achieving an accuracy score of 0.89 for strong reviews and 0.63 for weak reviews. We then compare the results with ChatGPT, without specialized training, which reaches a perfect accuracy score of 1.00 for strongly opinionated reviews and 0.75 for weakly opinionated reviews, concluding that it outperforms supervised learning approaches but is also imperfect in distinguishing more subtle sentiment in weakly opinionated reviews.

Introduction

Movie reviews are commonly used to decide what to watch. Reviews usually consist of textual descriptions the user provides and a numerical value representing the reviewer's overall opinion. This paper focuses on evaluating the performance of machine learning algorithms to identify whether a given movie review is positive or negative. This is a very relevant area since it is an example of sentiment analysis, which focuses on identifying opinions based on natural language. Sentiment analysis applies to various fields, such as consumer opinion and market analysis, evaluating the effect of marketing campaigns, or checking opinions about political topics. It can be applied to such data as tweets, reviews, and other online media content.

This study focuses on a representative sample of movie reviews collected by the well-known IMDb.com website (IMDb.com, 2023). Each review contains natural language descriptions (short and longer forms) and a numerical value representing the reviewer's sentiment (IMDb score). Because of this, it is an optimal data set to analyze using supervised machine learning methods. The data is first preprocessed to make it ready for machine learning, and then we explore the application of Decision Trees, Random Forests, and Support Vector Machines to determine sentiment based on the text of the review. Each of the models has its parameters optimized for this case to maximize the accuracy of the classification.

In recent years, a completely new approach, called Large Language Models (Introduction to Large Language Models, 2023), was developed to improve natural language processing. The most well-known example is OpenAI ChatGPT (OpenAI, 2023), a conversational system able to interpret natural language and answer complex questions on wide ranging topics. The system does not require training and can be directly used to determine sentiment. We use a sample of the reviews and ask ChatGPT to assess whether they are

positive or negative, and then compare the results with the accuracy of the traditional supervised machine learning models.

All the data analysis is done in Python in Jupiter Notebook (Jupyter, 2023). The numerical implementation of the supervised machine learning algorithms used in this paper is done using Scikit-learn Machine Learning toolkit (Scikit-learn, 2023). It also provides an extensive library of the natural language processing tools that we use to prepare the raw movie review data set for supervised machine learning. Finally, the evaluation of ChatGPT is done using OpenAI API and Python library.

Data Preparation

The data set used in this study is a sample from a body of 5.5M movie reviews (IMDb Review Dataset, 2023) obtained from (Kaggle, 2023), a data science platform. These are public user reviews collected by a well-known movie review website Internet Movie Database (IMDb, 2023). Each review contains the review date, the user identifier of the reviewer, the title of the movie, a numeric rating between 0 and 10, a short review summary, and a longer detailed review. We focus on a sample of 100K reviews covering 37,846 movies submitted to the IMDb.com website by 65,236 reviewers between April 2014 and September 2020. The average number of reviews per movie is 2.64. The most popular movie, in terms of the number of reviews, is the romance "Dil Bechara", with 748 reviews. The number of words in reviews for the detailed part varies between 1 and 1926, with a mean of 145 words. For the short review summary, the maximum is 35, with a mean of 5. Clearly detailed reviews should carry more information, and it is one of the aspects we evaluate.

Table 1. Summary statistics of the sample movie reviews data set. Most commented movies have more than 500 reviews each. The number of words in each review varies between 1 and almost 2000 for the detailed part of the review, with a mean of 145, and between 1 and 35 in review summary, with a mean of 5 words.

MOVIE TITLE	NUMBER OF REVIEWS
Dil Bechara (2020)	748
Wonder Woman 1984 (2020)	650
小丑 (2019)	635
Laxmii (2020)	485
STAR WARS：天行者的崛起 (2019)	483

(a)

STATISTIC	NR_WORDS_SUMMARY	NR_WORDS_DETAIL
count	100,000	100,000
mean	5	145
std	4	167
min	1	1
25%	3	36
50%	4	94
75%	7	183
max	35	1,916

(b)

In order to prepare the data for analysis, we preprocessed each review, both short summary and longer review, by:

1. Removing extra white spaces and converting all letters to lowercase.
2. Tokenizing the reviews by splitting them into lists of words.
3. Removing stopwords using Natural Language Processing Toolkit (NLTK) (nltk.org, 2023). Stopwords include words such as "just," "further," "then," "this," "only", etc. Such words do not carry useful information for the analysis.

4. Lemmatizing the words using WordNetLemmatizer class from NLTK. Lemmatization reduces words to their root form. For example, "running" is changed to "run". This enables the system to group together multiple versions of a word with the same meaning into a common form that can be recognized by machine learning systems.

Table 2. Example review before and after data preprocessing. Stopwords and white spaces are removed, words are lemmatized.

Original review	Preprocessed review
I lost interest when they killed off Antonio. He was the Ying to the Yan of the show. Why couldn't they let him go away somewhere. The show overall is a good one but....	lost interest killed antonio ying yan show couldnt let go away somewhere show overall good one

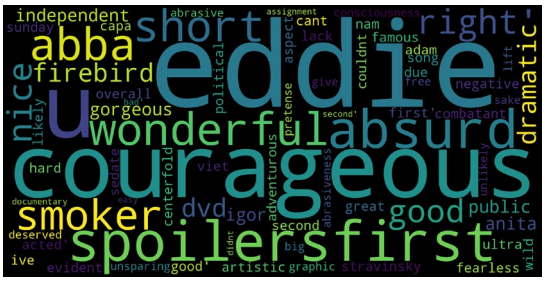
We explore how the opinion strength in the review affects the effectiveness of the machine learning approach for sentiment analysis. We expect that reviews expressing strongly opinionated negative and positive sentiment are easier to distinguish than those with weaker strength of opinion. To explore this question, we have divided the data into four groups, based on strength of sentiment as indicated by the IMDb.com score, a numerical value assigned by the reviewer. Table 3 summarizes how the data set is divided, including the relative fraction of reviews in each of the four categories. Positive and strong positive reviews dominate, with 70% of total reviews falling in these categories.

Table 3. Categorization of movie reviews into strongly and weakly opinionated. The data set is biased towards positive reviews for strong and weak opinionated reviews.

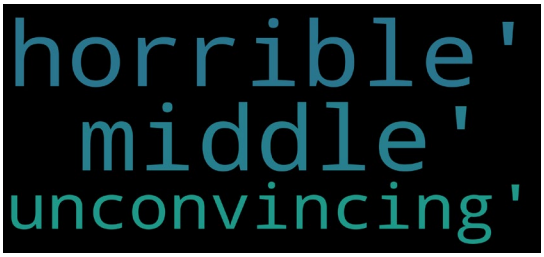
Strength of opinion	Category	IMDb.com score	Fraction of reviews	Number of reviews
Strong opinion	STRONG_POSITIVE	score ≥ 7	58%	57,813
	STRONG_NEGATIVE	score ≤ 3	20%	19,562
Weak opinion	POSITIVE	5 < score < 7	12%	12,334
	NEGATIVE	3 < score ≤ 5	10%	10,291

It is common to associate sentiment in review with adjectives used by the reviewer. It is interesting to see the most frequent adjectives in both positive and negative reviews. To do so, we used the NLTK toolkit (NLTK, 2023) that allows tagging of words in the review with part of speech. The results are shown in Figure 1 for both short summaries and long review descriptions. The most common adjectives are in line with expectation.

(a)
strong positive summary reviews



(b)
strong positive detailed reviews



(c)
strong negative summary reviews

(d)
strong negative detailed reviews

Figure 1. Most common adjectives in reviews. Plot (a) shows results for strong positive review summaries and (c) for strong negative summaries. Plot (b) shows results for strong positive detailed reviews and (d) for strong negative detailed reviews.

Given the skewness of the number of reviews towards positive, we balanced the data set by randomly choosing a subset of positive reviews to match the number of negative reviews in both strong and weak categories. That results in a data set consisting of 19,550 strong positive reviews, 19,562 strong negative reviews, 10,292 weak positive reviews, and 10,291 weak negative reviews. Finally, we divide the data set into two parts: 70% of reviews are used for training and the remainder for evaluation.

Application of Supervised Machine Learning

We apply and tune three supervised machine learning algorithms: Decision Trees, Random Forests, and Support Vector Machines. In each case, the data is analyzed separately for strongly opinionated reviews (with STRONG_POSITIVE and STRONG_NEGATIVE rating) and weakly opinionated reviews (with POSITIVE and NEGATIVE rating). This allows us to check how the accuracy is affected by strength of opinion. The classifiers are also applied to long descriptive reviews and short review summaries to evaluate the impact of the detail of text on the prediction accuracy.

The next phase of data analysis is done using the Sklearn machine learning package (Scikit-learn, 2023). The reviewer sentiment values (labeled positive or negative) are converted into binary form using the LabelBinarizer class. Review text itself is analyzed using the bag of words method. The method represents text as a collection of words without considering their relative positions. We use TfidfVectorizer to distill the most relevant features for the body of reviews. Specifically, texts of all reviews are converted into a matrix of counts of words and phrases (up to 2 words in a phrase). Each row represents one movie review, and columns are words and paired word combinations. Each element of the matrix holds the count of a given word (or pair of

words) in a review. To reduce the effect of very common words, the frequencies are scaled using the term frequency method implemented by TfidfTransformer (Scikit-learn, 2023).

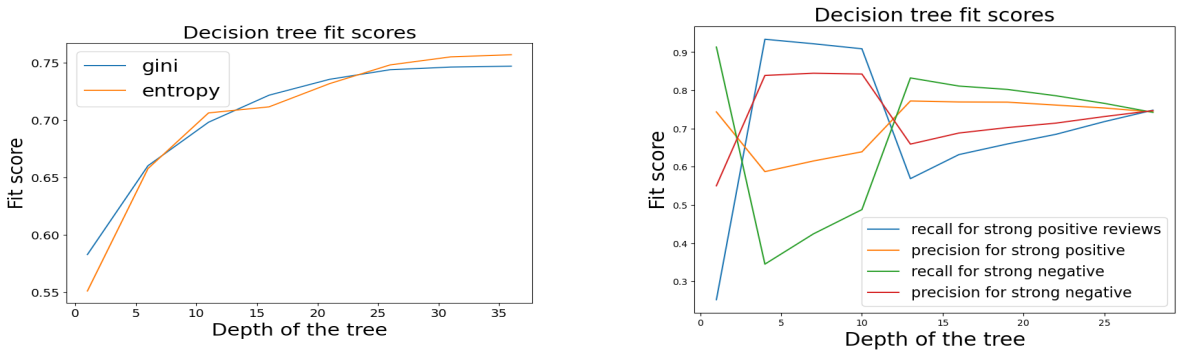
We use standard metrics of precision and recall to assess the quality of predictions made by a classifier. Both metrics range from 0 to 1, with 1 representing a perfect result. Precision captures what proportion of positive identifications was correct. Recall indicates what proportion of actual positives was identified correctly. There is also a composite accuracy score combining both metrics, called the F-1 score, but it is important to maximize both. For example, a classifier that always assigns positive will have perfect recall, but very bad precision. A good description of these metrics can be found in (Classification: Precision and Recall, 2023).

Decision Tree Classifier

Decision Tree classifier depends on a set of decision rules (based on conditions) to classify movie reviews. The rules are derived based on the learning data set. The key decisions while using this classifier are the choice of function to measure the quality of the fit, the maximum depth of the tree, and the minimum number of observations in a leaf of the tree. We set the minimum number of observations per leaf to 20 and then explore the effect of the maximum depth of the tree and type of fit function on the accuracy. Scikit classifier implements "gini" and "entropy" based fit functions, but we have not noticed any difference in the accuracy between these two. However, the accuracy significantly increases with the depth of the tree, up to 35, when it plateaus. It is also worth noting that recall and precision vary significantly as the tree's depth increases. The results are shown in Figure 2.

Figure 3 shows the summary results of using the best parametrization of the Decision Tree (maximum depth of 30) for the movie reviews classification problem. Precision and recall for strong reviews based on long descriptions are in the range of 0.75, with a combined F1-score of 0.75 (Figure 3a). It reduces to 0.62 if only summary reviews are used (Figure 3c). The results are much worse for weakly opinionated reviews, with those based on long reviews achieving a F1-score of 0.57 (Figure 3b) and only 0.51 for short summaries (Figure 3d). It is important to notice that in some cases, the results are very bad. For example, recall for weakly opinionated short reviews is as low as 0.16, making this kind of predictor not useful.

Our results show that single Decision Tree is not a good classifier for movie review sentiment prediction. Another weak element of using this method is the relative lack of stability of both recall and precision as the size of the tree increases (Figure 2b).



(a) (b)

Figure 2. Effect of classifier fit function (gini and entropy) and depth of the Decision Tree on sentiment prediction accuracy (a) and results of fit for the optimal settings of the decision tree (b). While there is almost no difference in accuracy between fit functions, the depth of the tree improves recall and precision up to a depth of 30.

	precision	recall	f1-score	support
STRONG_NEGATIVE	0.75	0.74	0.74	5873
STRONG_POSITIVE	0.74	0.75	0.75	5861
accuracy			0.75	11734
macro avg	0.75	0.75	0.75	11734
weighted avg	0.75	0.75	0.75	11734

(a)
strongly opinionated long reviews

	precision	recall	f1-score	support
NEGATIVE	0.56	0.62	0.59	3026
POSITIVE	0.59	0.53	0.56	3149
accuracy			0.57	6175
macro avg	0.58	0.58	0.57	6175
weighted avg	0.58	0.57	0.57	6175

(b)
weakly opinionated long reviews

	precision	recall	f1-score	support
STRONG_NEGATIVE	0.58	0.85	0.69	5873
STRONG_POSITIVE	0.72	0.39	0.50	5861
accuracy			0.62	11734
macro avg	0.65	0.62	0.60	11734
weighted avg	0.65	0.62	0.60	11734

(c)
strongly opinionated summary reviews

	precision	recall	f1-score	support
NEGATIVE	0.50	0.88	0.64	3026
POSITIVE	0.58	0.16	0.25	3149
accuracy			0.51	6175
macro avg	0.54	0.52	0.44	6175
weighted avg	0.54	0.51	0.44	6175

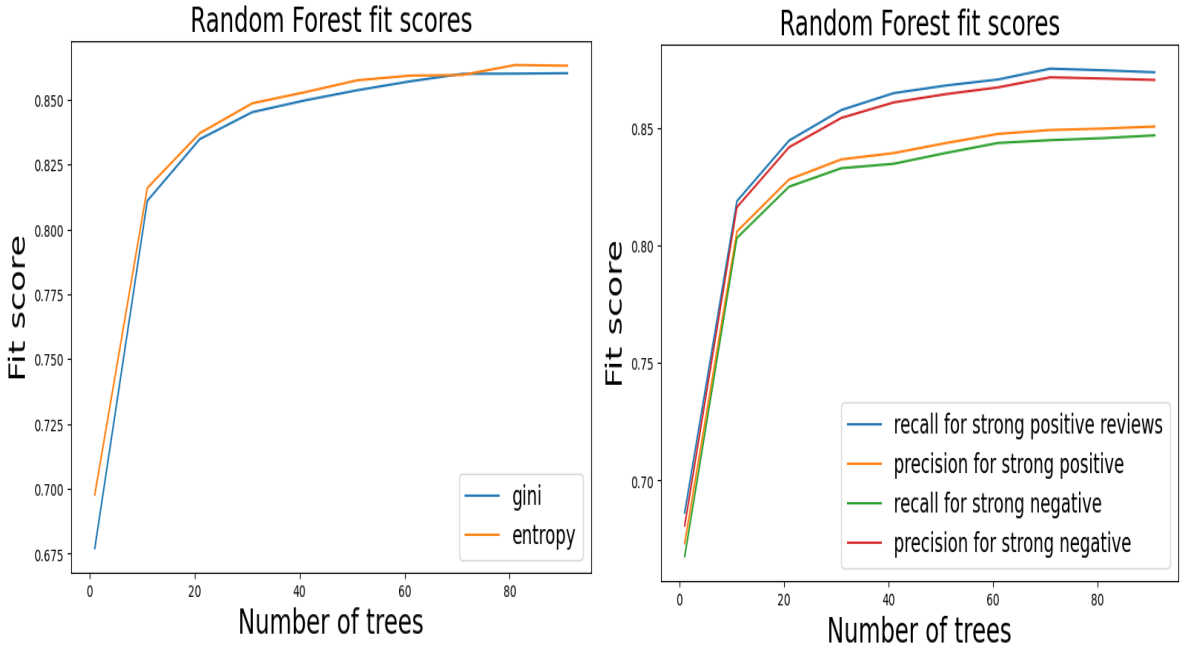
(d)
weakly opinionated summary reviews

Figure 3. Prediction accuracy for Decision Tree classifiers with optimal parameter settings (*gini* fit quality function and maximum tree depth of 30) for strongly and weakly opinionated long and short reviews.

Random Forest Classifier

Random Forest Classifier extends the basic Decision Tree by fitting several different decision tree classifiers on subsets of the dataset. It then uses averages to improve the accuracy of the classifier. A good description of this classifier is in (Random Forest Classifier, 2023). In the case of this classifier, we explore two parameters, the fitting function and the number of trees used. Similarly, as in case of Decision Trees, we did not observe a significant effect of fitting function, results are almost the same for both gini and entropy functions. However, the number of trees in the forest significantly affects accuracy. The accuracy increases to approximately 0.85 as the number of trees approaches 80. We determined this is the optimal choice for the movie review classification problem. The results of this parametrization analysis are presented in Figure 3. It is worth noting that in the case of Random Forest the increase of recall and accuracy with an increase of the size of the model (number of trees) is much steadier than in the case of a single Decision Tree, as shown in Figure 4b versus Figure 2b.

Overall, the results suggest that Random Forest is a much better classifier than Decision tree to use for sentiment analysis.



(a) (b)

Figure 4. Effect of classifier fit function (gini and entropy) and depth of the decision tree on sentiment prediction accuracy (a) and results of fit for the optimal settings of the Random Forest (b). While there is almost no difference in accuracy between fit functions, the number of trees in the forest improves recall and precision up to 80.

	precision	recall	f1-score	support
STRONG_NEGATIVE	0.87	0.86	0.86	5873
STRONG_POSITIVE	0.86	0.87	0.87	5861
accuracy			0.86	11734
macro avg	0.86	0.86	0.86	11734
weighted avg	0.86	0.86	0.86	11734

(a)
strongly opinionated long reviews

	precision	recall	f1-score	support
NEGATIVE	0.61	0.69	0.65	3026
POSITIVE	0.66	0.58	0.62	3149
accuracy			0.63	6175
macro avg	0.64	0.63	0.63	6175
weighted avg	0.64	0.63	0.63	6175

(b)
weakly opinionated long reviews

	precision	recall	f1-score	support
STRONG_NEGATIVE	0.78	0.81	0.79	5873
STRONG_POSITIVE	0.80	0.77	0.78	5861
accuracy			0.79	11734
macro avg	0.79	0.79	0.79	11734
weighted avg	0.79	0.79	0.79	11734

(c)
strongly opinionated summary reviews

	precision	recall	f1-score	support
NEGATIVE	0.55	0.65	0.60	3026
POSITIVE	0.60	0.49	0.54	3149
accuracy			0.57	6175
macro avg	0.57	0.57	0.57	6175
weighted avg	0.57	0.57	0.57	6175

(d)
weakly opinionated summary reviews

Figure 5. Prediction accuracy for Random Forest classifiers with optimal parameter settings (*entropy* fit quality function and maximum number of trees of 80) for strongly and weakly opinionated long and short reviews.

The overall results of applying the best parametrization (80 trees) are presented in Figure 4. Precision and recall for strong reviews based on long descriptions are as high as 0.86 and 0.87, with a combined F1-score of 0.86 (Figure 4a). It reduces to below 0.80 if only summary reviews are used (Figure 4c). The results are much worse for weakly opinionated reviews, with those based on long reviews achieving F1-score of 0.63 (Figure 4b) and only 0.57 for short summaries (Figure 4d). This is in line with our expectations.

Support Vector Machine Classifier

The final supervised learning classifier evaluated in this study is Support Vector Machines. We use the linear version of this classifier (Support Vector Machines, 2023). The key advantage of this classifier is that it is effective in high dimensional spaces, which is the case for sentiment analysis. The reason for this is that we have a large number of potential words in the entire body of all reviews, but only a small number of them are used in any specific review. We initially attempted to use non-linear version of this classifier, but the computation was taking too long. Using LinearSVC (LinearSVC Classifier, 2023) provides much faster results.

We found the accuracy of Support Vector Machines very good, reaching 0.89 F1-score with both recall and precision above 0.88 in the case of strongly opinionated long reviews (Figure 6a). Even while using the short versions of reviews, the results were still above 0.80 for both recall and precision (Figure 6c). As expected, the accuracy decreases while looking at weakly opinionated reviews, with an F1-score of 0.63 for long reviews and 0.58 for short reviews. These are still significantly better than the other two supervised machine learning approaches.

Clearly, the Support Vector Machines is the best traditional classifier for movie review sentiment analysis.

	precision	recall	f1-score	support
STRONG_NEGATIVE	0.90	0.88	0.89	5873
STRONG_POSITIVE	0.88	0.90	0.89	5861
accuracy			0.89	11734
macro avg	0.89	0.89	0.89	11734
weighted avg	0.89	0.89	0.89	11734

(a)
strongly opinionated long reviews

	precision	recall	f1-score	support
NEGATIVE	0.62	0.64	0.63	3026
POSITIVE	0.64	0.63	0.64	3149
accuracy			0.63	6175
macro avg	0.63	0.63	0.63	6175
weighted avg	0.63	0.63	0.63	6175

(b)
weakly opinionated long reviews

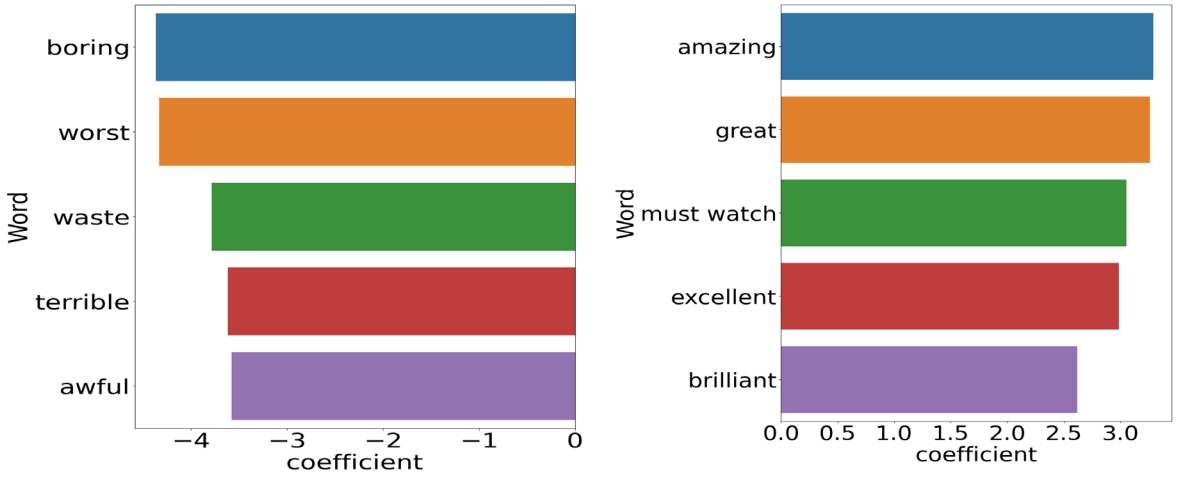
	precision	recall	f1-score	support
STRONG_NEGATIVE	0.81	0.80	0.80	5873
STRONG_POSITIVE	0.80	0.81	0.81	5861
accuracy			0.81	11734
macro avg	0.81	0.81	0.81	11734
weighted avg	0.81	0.81	0.81	11734

(c)
strongly opinionated summary reviews

	precision	recall	f1-score	support
NEGATIVE	0.57	0.56	0.57	3026
POSITIVE	0.58	0.59	0.59	3149
accuracy			0.58	6175
macro avg	0.58	0.58	0.58	6175
weighted avg	0.58	0.58	0.58	6175

(d)
weakly opinionated summary reviews

Figure 6. Prediction accuracy for Linear Support Vector Machines classifier for strongly and weakly opinionated long and short reviews.



(a) words with highest negative coefficients

(b) words with highest positive coefficients

Figure 7. Words with largest and smallest coefficients as identified by the Support Vector Machine classifier. These indicate the most impactful keywords that decide on review classification.

Another interesting aspect of Support Vector Machines is that it is relatively easy to interpret the fit model. In the case of Decision Trees, or Random Forests, the model is relatively complex with many conditional statements and the combining function in Random Forest. However, in Support Vector Machine we can look at the coefficients for each of the features, in our case words. Figure 7 shows these results, presenting the top 5 words with the lowest coefficients that indicate negative sentiment in Figure 7a and the top 5 words with the highest coefficients signifying positive sentiment. The words associated with positive sentiment are "amazing," "great," and "must watch," while those associated with negative sentiment are "boring," "worst," and "waste," well matching our intuition.

Application of OpenAI ChatGPT Large Language Model

We compare the results of traditional supervised machine learning with those of the OpenAI ChatGPT large language model classifier. This newest system, released in late 2022, does not require training and can be applied directly by posing a question. We use ChatGPT Python OpenAI library client (OpenAI Python API, 2023). Specifically, we use the ChatCompletion.create() function and pose the following context question:

- "You will be provided with a movie review, and your task is to classify its sentiment as positive or negative. Answer with one word."

We then send the entire body for movie review and receive classification results. The results are summarized in Figure 6. Our experiment is limited to a sample of 20 requests in each of the strong and weak review groups. ChatGPT had no mistakes for strongly opinionated reviews but misclassified some cases of weakly opinionated reviews. In addition to wrong classifications, we sometimes received two additional types of answers:

- "sorry, but I can't generate a response to that"
- "mixed"

Looking closer at the mistakes for weakly opinionated reviews, these were usually reviews either summarizing the movie, instead of rating it, or having very nuanced phrases.

Classification Report:

	precision	recall	f1-score	support
STRONG_NEGATIVE	1.00	1.00	1.00	13
STRONG_POSITIVE	1.00	1.00	1.00	7
accuracy			1.00	20
macro avg	1.00	1.00	1.00	20
weighted avg	1.00	1.00	1.00	20

(a)
strongly opinionated long reviews

Classification Report:

	precision	recall	f1-score	support
NEGATIVE	0.71	1.00	0.83	12
POSITIVE	1.00	0.38	0.55	8
accuracy			0.75	20
macro avg	0.85	0.69	0.69	20
weighted avg	0.82	0.75	0.71	20

(b)
weakly opinionated long reviews

Figure 8. Prediction accuracy of OpenAI ChatGPT on strongly opinionated reviews (a) and weakly opinionated reviews (b).

The overall results for ChatGPT, presented in Figure 8, show that it achieves perfect accuracy for strongly opinionated reviews, easily beating all traditional supervised learning algorithms. It is, however, a bit surprising that it does not perform so well on more nuanced, less opinionated reviews.

Conclusion

This article explores sentiment analysis for movie reviews using three traditional supervised machine learning classifiers and using large language model implemented by OpenAI ChatGPT. We find that the Decision Tree is not well suited for sentiment analysis, in some cases having very poor results, with recall as low as F1-score of 0.16. Random Forest classifier, that combines larger number of decisions trees, is much better and achieves F1-score 0.86 for long descriptions in strongly opinionated reviews. Support Vector Machines classifier still improves this result to 0.89, with both recall and precision above 0.88. This shows it is the most useful among the three supervised learning methods. Finally, comparison with the large language model confirms its supremacy, especially for strongly opinionated long reviews. In this case, we observed a perfect 1.00 F1-score. However, our results were achieved on a smaller data sample due to limited access to the OpenAI API.

Acknowledgments

I would like to thank my mentor anonymous.

References

Introduction to Large Language Models (2023).
https://developers.google.com/machine-learning/resources/intro-llms

OpenAI (2023).
https://openai.com/
Classification: Precision and Recall (2023).
https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Kaggle Data Science Platform (2023).
https://www.kaggle.com

IMDb Review Dataset – ebD (2023).
https://www.kaggle.com/datasets/ebiswas/imdb-review-dataset/

IMDb.com (2023).
https://www.imdb.com/

NLTK Natural Language Processing Tookit (2023).
https://www.nltk.org/

TfidfTransformer (2023).
https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

Scikit-learn Machine Learning in Python (2023).
https://scikit-learn.org/

Random Forest Classifier (2023). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Suport Vector Machines (2023).
https://scikit-learn.org/stable/modules/svm.html

LinearSVC Classifier (2023).
https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC

OpenAI Python API (2023).
https://github.com/openai/openai-python

Jupyter (2023).
https://jupyter.org/