# Sentiment Analysis with Naive Bayes Classifier

*Grant Croft*
*ISyE 6420*
*(Dated: November 30, 2023)*

**Abstract**

*This project focuses on sentiment analysis using a Naive Bayes classifier applied to the IMDB Dataset of 50K Movie Reviews. The goal is to predict the sentiment (positive or negative) based on the text content of the reviews. The Naive Bayes classifier is implemented using natural language processing techniques, including text preprocessing and Laplace smoothing.*

## Introduction:

Sentiment analysis is a crucial aspect of natural language processing, with applications ranging from product reviews to social media comments. The IMDB Dataset provides an extensive collection of movie reviews labelled with sentiments, making it an ideal dataset for training a sentiment analysis model.

## Methodology

The project employs a Naive Bayes classifier, a probabilistic machine learning model known for its simplicity and effectiveness in text classification tasks. The classifier is trained on a preprocessed subset of the IMDB Dataset, with a focus on feature extraction using Count Vectorization and Laplace smoothing to handle unseen words.

## Data

The IMDB Dataset of 50K Movie Reviews is utilized for this project. The dataset contains reviews labeled as positive or negative sentiments.

## Analysis/Implementation

1) Text Preprocessing:
- HTML tags, URLs, and non-alphanumeric characters are removed from the reviews.
- Text is converted to lowercase.
- Lemmatization is applied to find the root form of words.
- Stop words are removed (common words that don't add to the meaning i.e: for, an, nor, but, or, yet, so…)
2) Naive Bayes Classifier:
- The foundation of the Naive Bayes classifier lies in Bayes' Theorem, a fundamental concept in probability theory. Bayes' Theorem is expressed as follows:

$$P(A|B)=P(B|A)\cdot P(A) / P(B)$$

- Let A be the sentiment label (positive or negative) and B be the words present in a movie review. The goal is to calculate the probability of a particular sentiment given the observed words in a review; $P(A|B)$.
- The "Naive" assumption the Naive Bayes algorithm makes is that the features (words in this case) are conditionally independent given the class label. This allows for a more tractable calculation, greatly simplifying the process.
- The classifier computes the log likelihood of each label given the words in a review and combines them with the log prior probabilities.

- Laplace smoothing is applied to handle words not present in the training set, preventing zero probabilities.
- The fit method calculates the log prior probability for each label based on the training data.
- The fit_word_counts method extracts word counts for each label using Count Vectorization.

3) Prediction:
- The predict method utilizes the trained classifier to predict the sentiment of input texts.
- Predictions are made based on the label with the highest log probability.

4) Evaluation:
- The model is evaluated on a test set, and accuracy is calculated using scikit-learn's accuracy_score.

## Discussion

The Naive Bayes classifier is grounded in Bayes' Theorem, leveraging conditional probabilities to make predictions about sentiment based on observed words in a movie review. The model's assumptions, particularly the independence assumption, simplify the computation and allow for efficient training and prediction.

In the context of this project, Bayes' Theorem is employed to calculate the probability of a particular sentiment given the words present in a review. The log likelihoods of each sentiment class are combined with the log prior probabilities to make a final prediction. Laplace smoothing is applied to address the issue of zero probabilities for words not present in the training set.

While Naive Bayes classifiers are known for their simplicity, they often perform surprisingly well in text classification tasks. The approach adopted in this project showcases the application of Bayes' Theorem in a practical and effective manner, providing a solid foundation for sentiment analysis on textual data.

## Results

- Training Set Accuracy:

The Naive Bayes classifier achieved an accuracy of 85.24% on the split IMDB data. This indicates that the model performs well on the seen data, effectively capturing the sentiment patterns present in the IMDB Movie Reviews dataset.

- Robustness of the Model:

The model exhibits robust performance, particularly in cases where the language used is straightforward and unambiguous. It performs best when the review is conveyed in a straightforward manner. However, the model might encounter challenges when faced with nuanced language or instances where the sentiment is expressed in a counterintuitive manner, such as the phrase "the movie is so bad it is good." In such cases, where the sentiment is inherently contradictory, the model may struggle to accurately predict the sentiment.

- User Input Section:

The user input section provides an interactive way to test the model in real-time. The output demonstrates the model's capability to predict sentiments for various input movie reviews.
Positive Predictions:

Example: "Because the film is about Kishi Bashi and his inspiration for musical composition, the first hurdle an audience must clear is totally subjective: Do you like Kishi Bashi's music? I did, and this song film was thoughtful and inspiring."

Negative Predictions:

Example: "the movie is great if you want to be bored out of your mind and wishing you were elsewhere."

Challenges:
The model struggles with reviews containing contradictory sentiments or complex language.

Example: "terribly fantastic movie" may confuse the model due to the conflicting terms.

Overall Observations:
The model tends to perform well on reviews with clear sentiments, making it suitable for applications where straightforward opinions are prevalent.

## Conclusion

The model works quite well and achieves a high accuracy on the training set. However, caution is advised when dealing with reviews containing contradictory language or nuanced expressions. This could likely be improved by providing more training examples of similar language. The interactive user input section provides insights into the model's real-world application, showcasing its strengths and potential limitations.

Further exploration and refinement could enhance the model's ability to handle diverse linguistic nuances and improve overall accuracy.

## References

IMDB Dataset of 50K Movie Reviews.
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Scikit-learn Documentation.
https://scikit-learn.org/stable/index.html

## Appendices

#### Code Output:

Accuracy of prediction on the test set: 0.8524

Enter a movie review (type 'exit' to stop): Because the film is about Kishi Bashi and his inspiration for musical composition, the first hurdle an audience must clear is totally subjective: Do you like Kishi Bashi's music? I did and this song film was thoughtful and inspiring
Predicted sentiment: positive

Enter a movie review (type 'exit' to stop): Captivating, heartfelt and illuminating. A powerful protest against hate.
Predicted sentiment: positive

Enter a movie review (type 'exit' to stop): the movie is great if you want to be bored out of your mind and wishing you were elsewhere
Predicted sentiment: negative

Enter a movie review (type 'exit' to stop): terriblly great movie
Predicted sentiment: positive

Enter a movie review (type 'exit' to stop): terrible movie
Predicted sentiment: negative

Enter a movie review (type 'exit' to stop): Essentially plotless action film has two good guys (Fong and Roundtree) pitted against two bad guys (Mitchell and Pierce). Fong is perhaps the most uncharismatic action lead of the 80s, Roundtree's small part is a far cry from his 'Shaft' days, and Cameron Mitchell adds another shameful role to his career, one to sit right next to his laughable turn in 'The Toolbox Murders' (this man was a respected actor once, now he has come down to wearing flowers in his hair and complaining about people bleeding on his carpet). Only Stack Pierce acts with some dignity. As for the violence, don't worry: most of it is too badly done to offend anyone
Predicted sentiment: negative

Enter a movie review (type 'exit' to stop): I grew up on this movie and I can remember when my brother and I used to play in the backyard and pretend we were in Care-a-lot. Now, after so many years have passed, I get to watch the movie with my daughter and watch her enjoy it. If you are parent and you have not watched this movie with your children, then you should, just so you hold them in your arms and watch them get thrilled over the care bears and care-a-lot! The songs, especially 'Forever Young' are very sweet and memorable. Parents, I highly recommend this movie for all kids so they can learn how enjoyable caring for others

can be! When it comes down to all the trash that is on TV, you can raise your children to have the right frame of mind about life with movies like these.
Predicted sentiment: positive

Enter a movie review (type 'exit' to stop): I try to catch this film each time it's shown on tv, which happily is quite often. But I keep forgetting to video it. As it is, I practically know the script by heart, but that doesn't stop me having a good cry, in fact it probably adds to it as I cry knowing what's coming next. It's such a lovely film - well made, well cast, good photography. I love it. One of my top ten films.
Predicted sentiment: positive

Enter a movie review (type 'exit' to stop): Missed it at the cinema, but was always slightly compelled. Found it in the throw-out bin at my local video shop for a measly two bucks! Will I now give it away to anyone who wants it? Probably! No purposeful plot, one dimensional characters, plastic world ripped off from many far better films, no decent dialogue to speak of. You know that empty feeling when you come down off ecstasy? Its that feeling right here. Sad thing is, the Australia I know is heading in this direction, minus the melodrama and simple answers. Interesting only to see the older Aussie actors (who had to ACT back in their day to get by) vs the newer Aussie actors (who have to LOOK GOOD to get by). Like some horribly garish narrative introduction to a film clip that never actually starts... Poor Kylie, started her career as an actress as well...
Predicted sentiment: negative

Enter a movie review (type 'exit' to stop): terribly good movie
Predicted sentiment: negative

Enter a movie review (type 'exit' to stop): terribly great movie
Predicted sentiment: negative

Enter a movie review (type 'exit' to stop): terribly fantastic movie
Predicted sentiment: positive

Enter a movie review (type 'exit' to stop): this moive is so bad it's good
Predicted sentiment: negative