TEXTUAL SENTIMENT ANALYSIS AND STEGANOGRAPHY DETECTION IN IMDB REVIEWS

PRESENTED BY:

- **Sashidhar Chary Viswanathula**

INTRODUCTION

Our project focuses on the Sentiment Analysis and Steganography Detection in IMDB Reviews :

1. Natural Language Processing (NLP)

Definition: NLP is a field at the intersection of computer science, artificial intelligence, and linguistics. It focuses on enabling computers to understand, interpret, and respond to human language in a valuable way.

Applications:

- Text classification
- Language Translation
- Chatbots
- Voice assistants

2. Sentiment Analysis

Definition: Sentiment Analysis is a technique used in NLP to identify and categorize opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral.

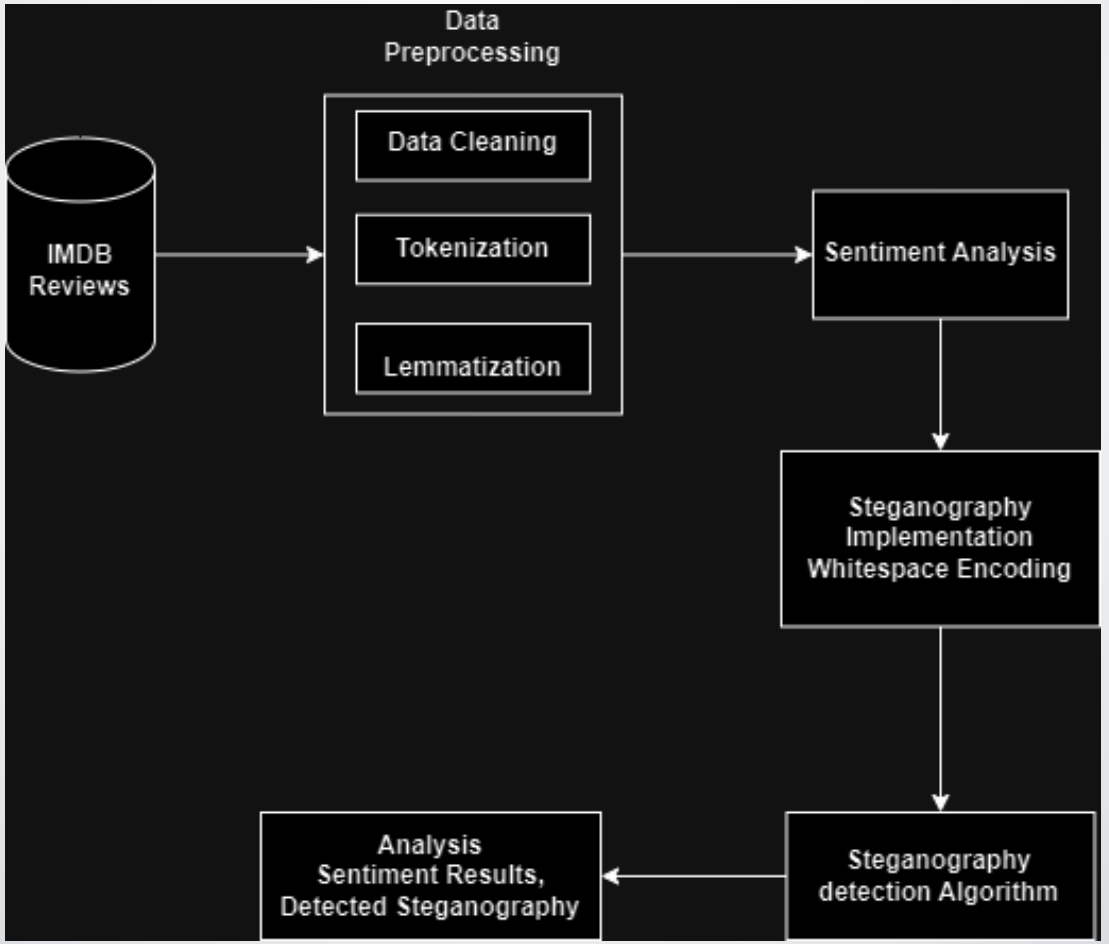
3. Steganography in Text

Definition: Steganography is the art of hiding information within other non-secret text or data. In the context of text, it involves embedding a secret message within a normal-looking text in a way that avoids detection.

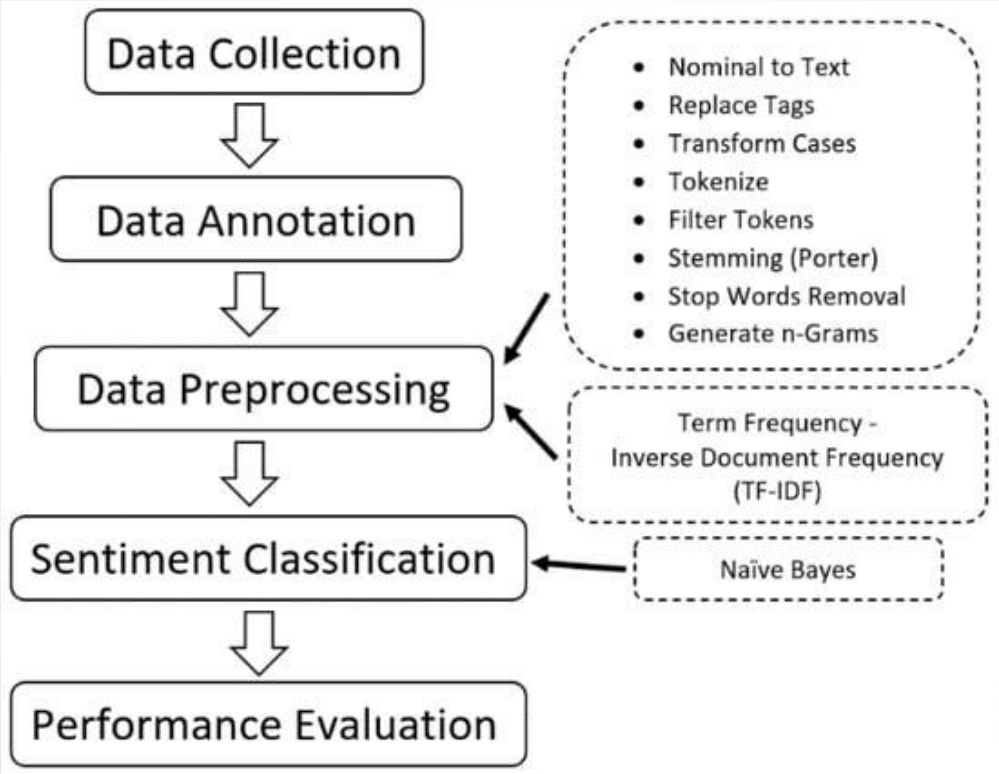
HYPOTHESIS

- In our project, we delve into the rich and complex world of IMDB movie reviews with a dual-purpose mission. Our primary objective is to analyze these reviews for sentiment, to understand whether viewers' opinions are positive or negative, and to gauge the overall emotional tone of their feedback.
- This aspect of our work seeks to provide a deeper understanding of public sentiment around movies, offering valuable insights for filmmakers and audiences alike. But our project takes an intriguing turn as we also employ steganography detection techniques.
- Here, our goal is to identify hidden messages embedded within the reviews. This unique approach not only challenges our data analysis skills but also opens up new possibilities in understanding how text can be used to convey more than meets the eye.
- By combining sentiment analysis with steganography detection, our project stands at the intersection of emotional analytics and covert communication detection, pioneering a common approach to text analysis.

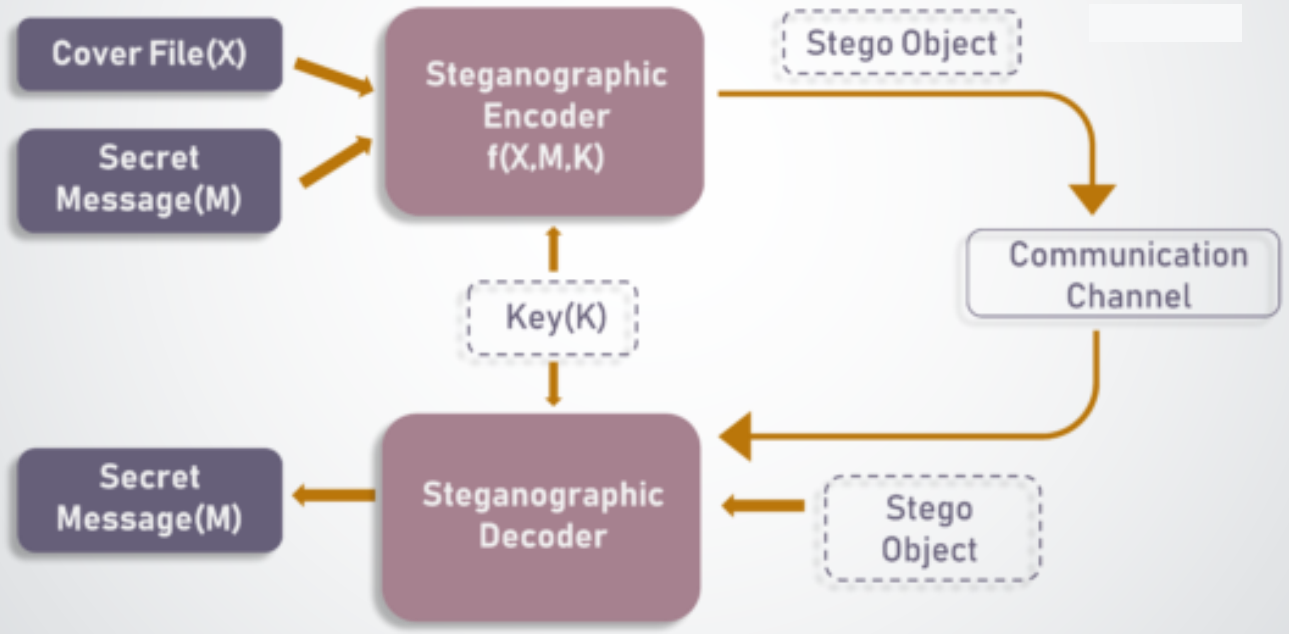
WORKFLOW DIAGRAM



ARCHITECTURE DIAGRAM OF SENTIMENT ANALYSIS



ARCHITECTURE DIAGRAM OF STEGANOGRAPHY DETECTION



DATASET OVERVEIW

- Our exploration is grounded in the IMDB dataset, a well-curated collection of 50,000 movie reviews. This dataset is a reflection of diverse cinematic tastes, encapsulating viewer responses to a broad spectrum of films.
- Sourced from the Internet Movie Database (IMDB), a leading authority in the movie industry, this dataset is more than just numbers and text. Each review is a blend of subjective opinion and objective description, offering a multifaceted view of the movie-watching experience.
- The dataset encompasses reviews from blockbusters to indie films, capturing the wide array of genres and styles that make up the film industry.

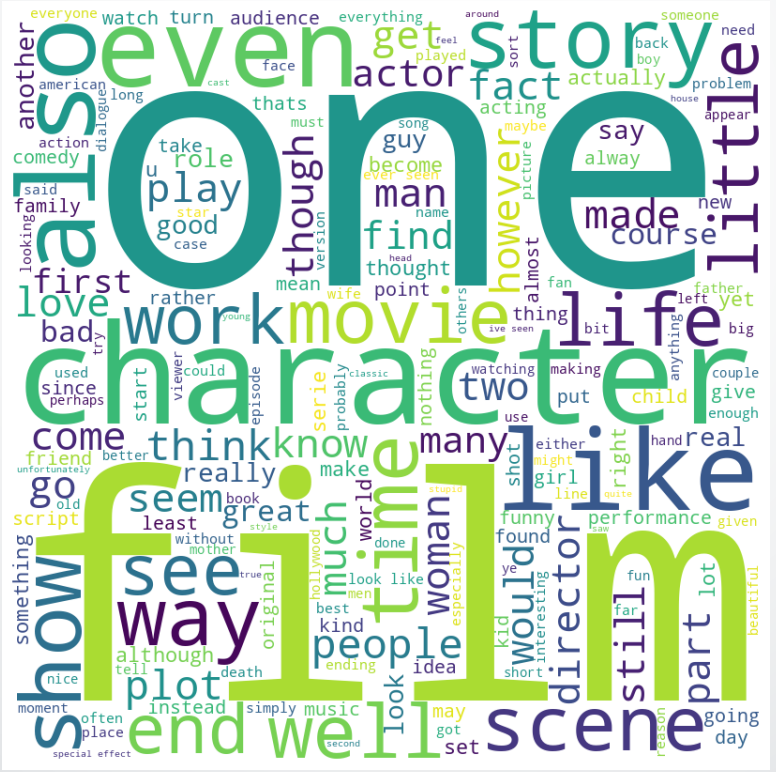
DATASET FEATURES

It's structured into two primary segments: 25,000 reviews are labeled as positive, and 25,000 as negative.

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
...	...	...
95	Daniel Day-Lewis is the most versatile actor a...	positive
96	My guess would be this was originally going to...	negative
97	Well, I like to watch bad horror B-Movies, cau...	negative
98	This IS the worst movie I have ever seen, as w...	negative
99	I have been a Mario fan for as long as I can r...	positive

DATA VISUALIZATION

VISUALIZATION OF ALL REVIEWS

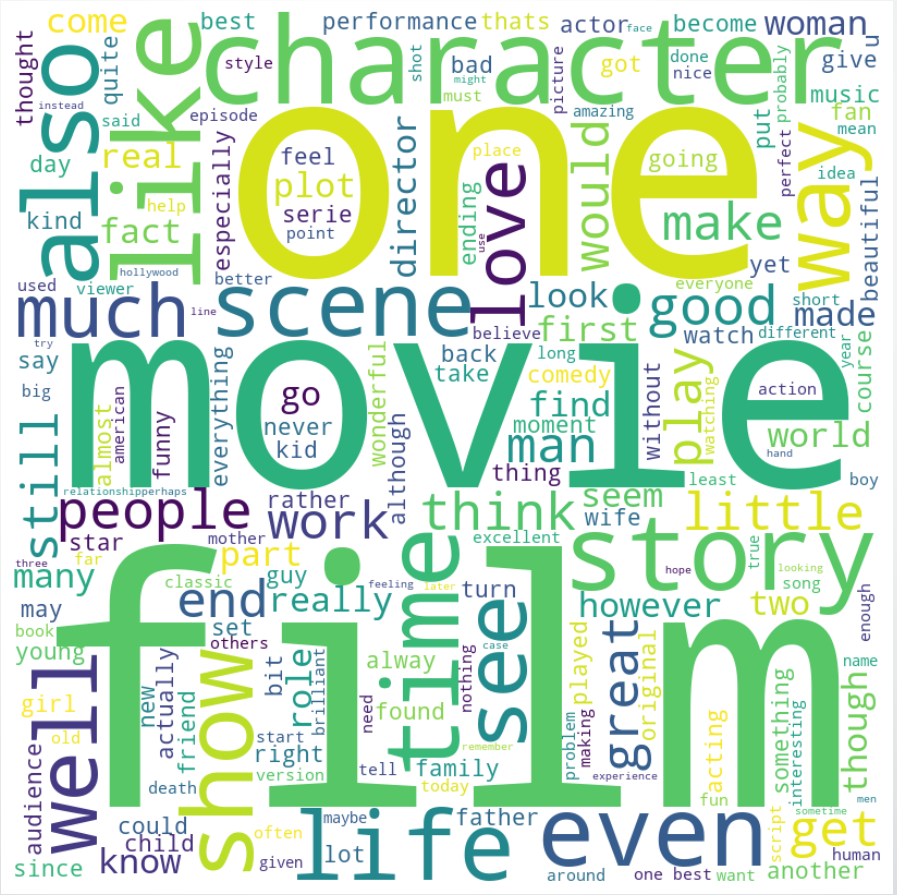


DATA VISUALIZATION

VISUALIZATION OF NEGATIVE REVIEWS

DATA VISUALIZATION

VISUALIZATION OF POSITIVE REVIEWS



IMPLEMENTATION LIBRARIES

These are all the libraries we've used in this project:

- Pandas
- Wordcloud
- Regex
- Numpy
- Matplotlip
  - Pyplot
- Nltk
  - Stopwords
- Sklearn
  - Train_test_split
  - Accuracy_score
  - Precision_score
  - Recall_score
  - F1_score
  - Confusion_matrix
- Random

DATA PREPROCESSING

- In our project, the first important step was to clean and prepare the movie reviews from IMDB. We started by removing HTML tags, which are not needed for our analysis. Then, we took out any special characters and numbers, leaving only the words. Next, we split these reviews into individual words. This process is called tokenization.

- We also removed common words like 'and', 'is', and 'the', which are not useful for understanding the feelings in the reviews or finding hidden messages. The last step was lemmatization, where we changed the words to their basic form. For example, 'running' becomes 'run’.

- This helps in making the analysis simpler and more accurate. By doing all this, we made sure our data was clean and ready for the next steps of our project - finding out the sentiments in the reviews and looking for any hidden messages.

SENTIMENTAL ANALYSIS MODEL

- In our project, we developed a tool for movie reviews for sentiment analysis using the Naive Bayes classifier. This is a straightforward yet powerful model that helps us figure out whether a review is positive or negative.
- We chose Naive Bayes because it's great with words – it looks at how often certain words show up in positive or negative reviews and uses this to make smart guesses about new reviews.
- It’s like teaching a computer to recognize happy or sad stories just by looking at the words used. This model is quite a star when it comes to sorting through lots of reviews quickly and effectively, which is exactly what we needed for our project.
- It’s not just about speed, it’s also really good at picking up on the different ways people talk about movies, making it a reliable choice for our analysis.

STEGANOGRAPHY IMPLEMENTATION

•In our exploration of IMDB reviews, we ventured into the realm of steganography, embedding hidden messages directly within the text. One of the key techniques we used is called whitespace encoding.

•Here, we cleverly manipulated spaces at the ends of lines in the reviews. These extra spaces, though seemingly innocuous, actually represent the bits of our concealed messages.

•Additionally, we employed synonym replacement, a more subtle approach where we switched certain words in the reviews with their synonyms. This technique allowed us to embed secret messages without altering the overall sentiment of the review.

•It's a delicate balance of maintaining the original tone while embedding our hidden data.

DETECTING THE UNDETECTED: STEGANOGRAPHY IN TEXT

- Detecting hidden messages in text presents its own set of challenges, which we tackled head-on. Our detection method was designed to identify unusual patterns that could suggest the presence of steganography.

- We specifically looked for anomalies like extra spaces at the end of lines – the hallmark of whitespace encoding. Additionally, we analyzed word choices for signs of synonym replacement, which could indicate an attempt to conceal a message.

- One of the significant challenges we faced was differentiating between normal variations in language and deliberate steganographic alterations. It required a keen eye for detail and sophisticated analysis techniques to distinguish genuine language usage from cleverly disguised hidden messages.

PERFORMANCE RESULTS

Metrics for Sentiment Analysis:

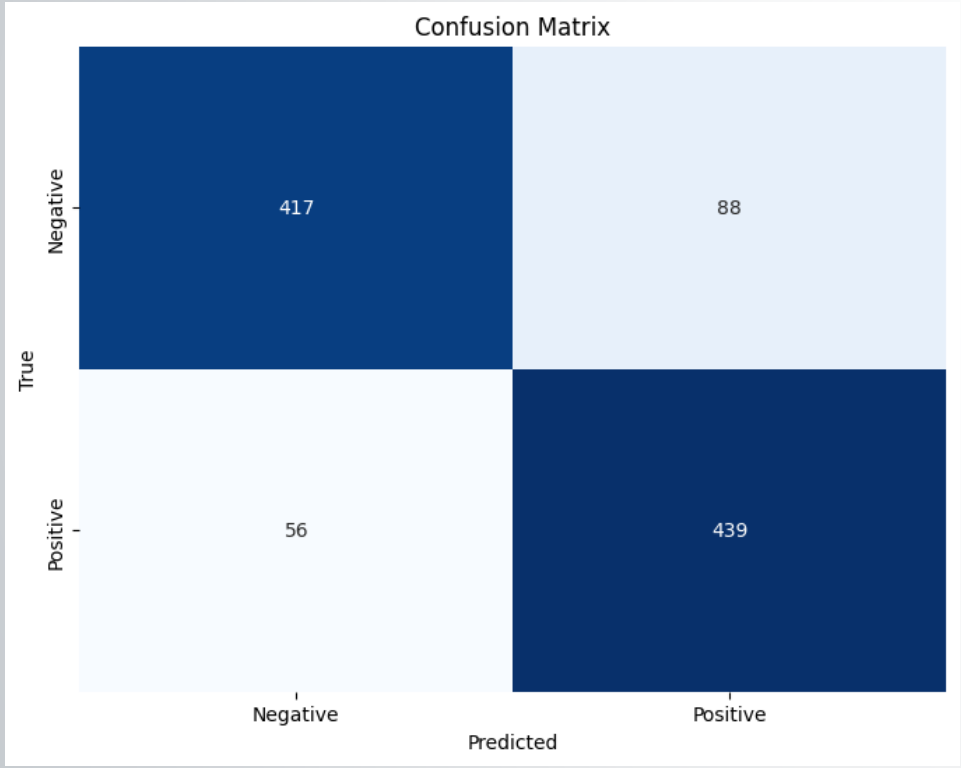
Sentiment Analysis Model Test:

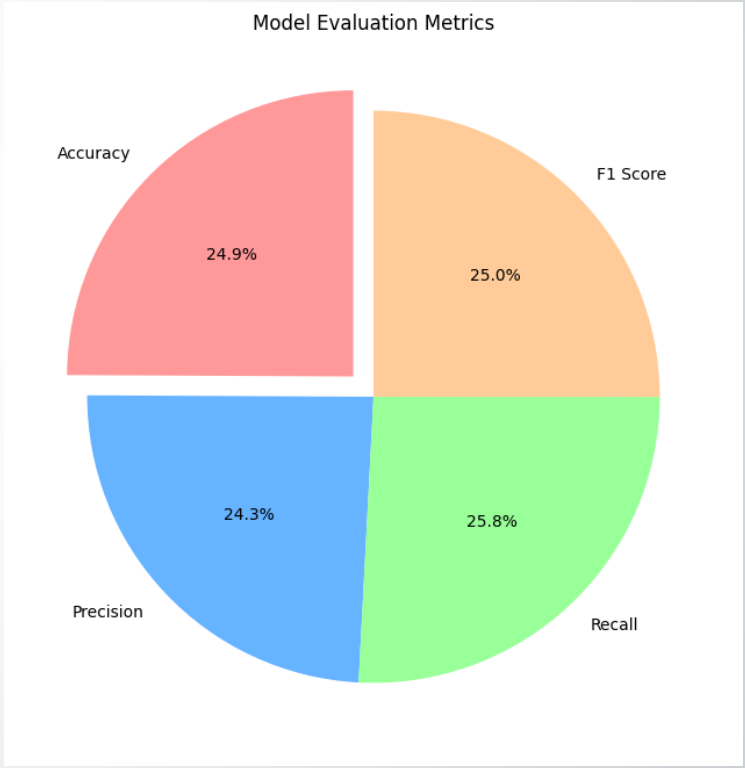
	precision	recall	f1-score	support
0	0.88	0.83	0.85	505
1	0.83	0.89	0.86	495
accuracy			0.86	1000
macro avg	0.86	0.86	0.86	1000
weighted avg	0.86	0.86	0.86	1000

Metrics for Created Pipeline:

	precision	recall	f1-score	support
0	0.90	1.00	0.95	45000
1	0.00	0.00	0.00	5000
accuracy			0.90	50000
macro avg	0.45	0.50	0.47	50000
weighted avg	0.81	0.90	0.85	50000

PERFORMANCE RESULTS





APPLICATIONS

- Application in Film Industry and Marketing
- Content Moderation and Online Safety
- Security and Privacy in Digital Communication
- Research and Academic Studies

ISSUES

- We’ve faced issue in choosing the appropriate dataset for the steganography and sentiment analysis.
- Detecting hidden texts from the dataset.
- Finding the synonym’s and replacing the words, without losing its sentiment.

FUTURE WORK

- We plan to explore more advanced models like LSTM (Long Short-Term Memory) networks for sentiment analysis, aiming for higher accuracy and contextual understanding.

- To ensure our models are robust and unbiased, we'll incorporate a wider range of text sources, including social media posts, news articles, and different languages.

REFERENCES

- Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.
- Aggarwal, C. C., & Zhai, C. (2012). Mining Text Data.
- Springer US.Feldman, R. (2013). Techniques and Applications for Sentiment Analysis. Communications of the ACM, 56(4), 82-89.
- Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113. doi:10.1016/j.asej.2014.04.011

CONCLUSION

To Conclude, our journey into the realms of sentiment analysis and steganography has been both enlightening and groundbreaking. We've delved into the depths of movie reviews, unraveling emotions through our Naive Bayes model, and ventured into the intricate world of hidden messages with our text-based steganography techniques. These steps forward are not just about technology; they're about understanding human expression and securing communication in innovative ways. The challenges we encountered along the way were steppingstones, leading us to new learning and growth.