

# Diving into Darkness: A Comparative Analysis of Batman Movies Through NLP and Sentiment Analysis

Charity Smith  
MSDS 692 Data Science Practicum  
Regis Summer 2024

# Which Batman “Wins”?



*Batman*  
Screenplay by Sam Hamm  
1989

- Explore the themes of darkness and the complexity of relationships in Batman movies.
- Use NLP and sentiment analysis to identify the darkest version of Batman and the most subjective dialogue activity.
- Analyze movie reviews to determine the most popular Batman movie.
- Provide recommendations for future movie productions based on the analysis.



*The Dark Knight*  
Screenplay by Jonathan Nolan  
and Christopher Nolan  
2008

# Data: Movie Scripts

- **Movie Scripts:** PDFs for *Batman* (1986) and *The Dark Knight* (2008)
- **Libraries Used:**
  - pdfplumber: To extract text from PDF files.
  - re: For regular expressions to clean and preprocess text.
  - nltk: For tokenizing and removing stopwords.
- **Steps:**
  1. Extract text from PDFs using pdfplumber.
  2. Remove non-dialogue elements such as scene descriptions using regex patterns.
  3. Tokenize text and remove stopwords for further analysis.

```
# Function to extract text from PDF
def extract_text_from_pdf(pdf_path):
    with pdfplumber.open(pdf_path) as pdf:
        text = ''
        for page in pdf.pages:
            text += page.extract_text()
    return text

# Extracting text from each PDF and storing it in a dictionary
scripts = {}
for movie, pdf_path in pdf_files.items():
    scripts[movie] = extract_text_from_pdf(pdf_path)
```

--- Batman ---  
BATMAN  
Screenplay by  
Sam Hamm  
Based on the Character Created by  
Bob Kane  
FIRST DRAFT  
October 20, 1986  
FADE IN:  
EXT. CITYSCAPE - NIGHT  
The place is Gotham City. The time, 1987 -- once removed.  
The city of Tomorrow: stark angles, creeping shadows,  
dense, crowded, airless, a random tangle of steel and  
concrete, self-generating, almost subterranean in its  
aspect... as if hell had erupted through the sidewalks and  
kept on growing. A dangling fat moon shines overhead, ready  
to burst.  
EXT. CATHEDRAL -

--- The Dark Knight ---  
THE DARK KNIGHT  
by  
Jonathan Nolan and Christopher Nolan  
Story by  
Christopher Nolan & David S. Goyer  
Batman created by  
Bob Kane T H E D A R K K N I G H T  
BURNING. Massive flames. A dark shape emerg  
SYMBOL. Growing. Filling the screen with BL

## EXT. GOTHAM SQUARE - NIGHT

The pulsing heart of downtown Gotham, a neon nightmare of big-city corruption, almost surreal in its oppressiveness. Hookers wave to drug dealers. Street hustlers slap high-fives with three-card monte dealers. They all seem to know each other... with one conspicuous exception:

A TOURIST FAMILY, Mom, Dad, and little Jimmy, staring straight ahead as they march in perfect lockstep down the main drag. They've just come out of a bit show two blocks over; the respectable theatre crowd has thinned out, and now -- Playbills in hand -- they find themselves adrift in the predatory traffic of Gotham's meanest street.

MOM  
For God's sake, Harold, can we  
please just get a taxi??

# Batman Movie Scripts

## Extracting Dialogue for Different Characters:

- Used patterns to identify and remove scene descriptions (e.g., INT., EXT., FADE IN).
- Identified character names by detecting uppercase lines with limited words.

DEALER No, man. I don't like it tonight.

BUYER What're you, superstitious? You got more chance of winning the powerball than running into him...

INT. MAJOR CRIMES UNIT, GOTHAM CENTRAL -- NIGHT  
DETECTIVE RAMIREZ, female, rookie detective, 30's, makes coffee, watching a news show on the television.

ON SCREEN: The host, MIKE ENGEL, lays into the MAYOR.

ENGEL Mr Mayor, you were elected on a campaign to clean up the city... when are you going to start?

MAYOR Well, Mike-

Dialogue extracted:

```
DEALER: ["No, man. I don't like it tonight."]
BUYER: ["What're you, superstitious? You got", "more chance of winning the powerball"]
ENGEL: ['Mr Mayor, you were elected on a', 'campaign to clean up the city... when']
MAYOR: ['Well, Mike-', "I'm told our men in the Major Crimes"]
```

# Data: Movie Reviews

- **Movie Reviews:** Scraped from Rotten Tomatoes
- **Libraries Used:**
  - selenium: For web scraping.
  - pandas: To store and manipulate the scraped reviews.
- **Steps:**
  1. Use selenium to automate the browser and scrape reviews from Rotten Tomatoes.
  2. Click the "Load More" button to load all reviews.
  3. Extract and store reviews in a pandas DataFrame.

The screenshot shows the Rotten Tomatoes homepage with a search bar at the top. Below it, a banner for 'TRENDING ON RT' features '2024's Best Movies', 'Most Popular Shows', 'July's Anticipated Movies', and 'Horizon: An American Saga'. The main content area is titled 'BATMAN REVIEWS' and includes tabs for 'All Critics', 'Top Critics', 'All Audience', and 'Verified Audience'. Three reviews are listed:

- Casey Chong** (Casey's Movie Mania) - A review by Casey Chong from Casey's Movie Mania, rating 3.5/5. It describes Jack Nicholson as 'steals most of the show here with his gleefully over-the-top sociopathic performance.'  
[Full Review](#) | Original Score: 3.5/5 | Jun 17, 2024
- Joshua M. Patton** (CBR) - A review by Joshua M. Patton from CBR, rating 9/10. It describes Tim Burton's Batman as a 'Gothic superhero movie that embraced artifice and shunned grounded reality, and it was better for it.'  
[Full Review](#) | Original Score: 9/10 | Apr 30, 2024
- Cory Woodroof** (For the Win (USA Today)) - A review by Cory Woodroof from USA Today, rating 10/10. It describes the movie as 'Set to Danny Elfman's triumphant score, Burton's first Batman movie will live eternal in how it irreversibly changed movies and turned Keaton into a movie star overnight.'  
[Full Review](#) | Feb 28, 2024

# Batman Movie Reviews

## Challenges with the "Load More" button:

- The web pages required repeatedly clicking the "Load More" button to retrieve all reviews.
- Implemented a loop to click the "Load More" button until all reviews were loaded.

Resulting CSV:  
Scraped reviews were stored in a Pandas DataFrame and then saved as a CSV file.

The screenshot shows the Rotten Tomatoes homepage with a search bar and navigation links. Below, a section for 'THE DARK KNIGHT REVIEWS' is displayed. It includes reviews from 'All Critics', 'Top Critics', 'All Audience', and 'Verified Audience'. Three reviews are visible: one from Juan Pablo Russo, one from Akhil Arora, and one from Zaki Hasan. At the bottom right of the review list, there is a blue 'LOAD MORE' button with a yellow arrow pointing to it.

```
# Load all reviews by clicking the "Load More" button
while True:
    try:
        load_more_button = driver.find_element(By.CSS_SELECTOR, 'div.load-more-content')
        if 'hide' in load_more_button.get_attribute('class'):
            print("No more reviews to load. All reviews loaded.")
            break
        else:
            driver.execute_script("arguments[0].click();", load_more_button)
            print("Clicked 'Load More' button")
            time.sleep(2) # Wait for the reviews to load
    except NoSuchElementException:
        print("No more 'Load More' button found. All reviews loaded.")
        break
```

```
# Scrape reviews for each movie
for movie, url in urls.items():
    print(f"Scraping reviews for {movie}...")
    reviews_data[movie] = scrape_reviews(url)
    # Adding a delay to avoid hitting the server too frequently
    time.sleep(1)
```

```
Scraping reviews for Batman...
Clicked 'Load More' button
No more reviews to load. All reviews loaded.
Found 140 reviews for URL: https://www.rottentomatoes.com/m/1001781-batman/reviews
Scraping reviews for The Dark Knight...
Clicked 'Load More' button
Clicked 'Load More' button
```

| Movie  | Review   |
|--------|--|
| Batman | Tim Burton's Batman is a Gothic superhero movie that embraced artifice and shunned grounded reality, and it was better for it.                                     |
| Batman | Set to Danny Elfman's triumphant score, Burton's first Batman movie will live eternal in how it irreversibly changed movies and turned Keaton into a movie star.   |
| Batman | The result is a film that may please adults more than youngsters, and isn't that refreshing?   |
| Batman | The Gotham City created in Batman is one of the most distinctive and atmospheric places I've seen in the movies. It's a shame something more memorable came along. |
| Batman | It is both a technical and aesthetic triumph, a fact made all the more impressive since the movie was almost buried in its own hype before anyone had seen it.     |
| Batman | One of the most interesting comic-book movies ever made, a movie that vividly illustrates where most of the others have gone wrong.                                |

# EDA in Movie Scripts

Most common adjectives:

*Batman*

Black, little, dead

*The Dark Knight*

Good, little, new

## Number of Unique Characters in Each Movie:

*Batman*: 149, *The Dark Knight*: 143

### Word Count

*Batman*

- Mean: 7.60
- Median: 8.0
- Standard Deviation: 3.40

*The Dark Knight*

- Mean: 6.83
- Median: 7.0
- Standard Deviation: 2.95

### Average Word Length

*Batman*

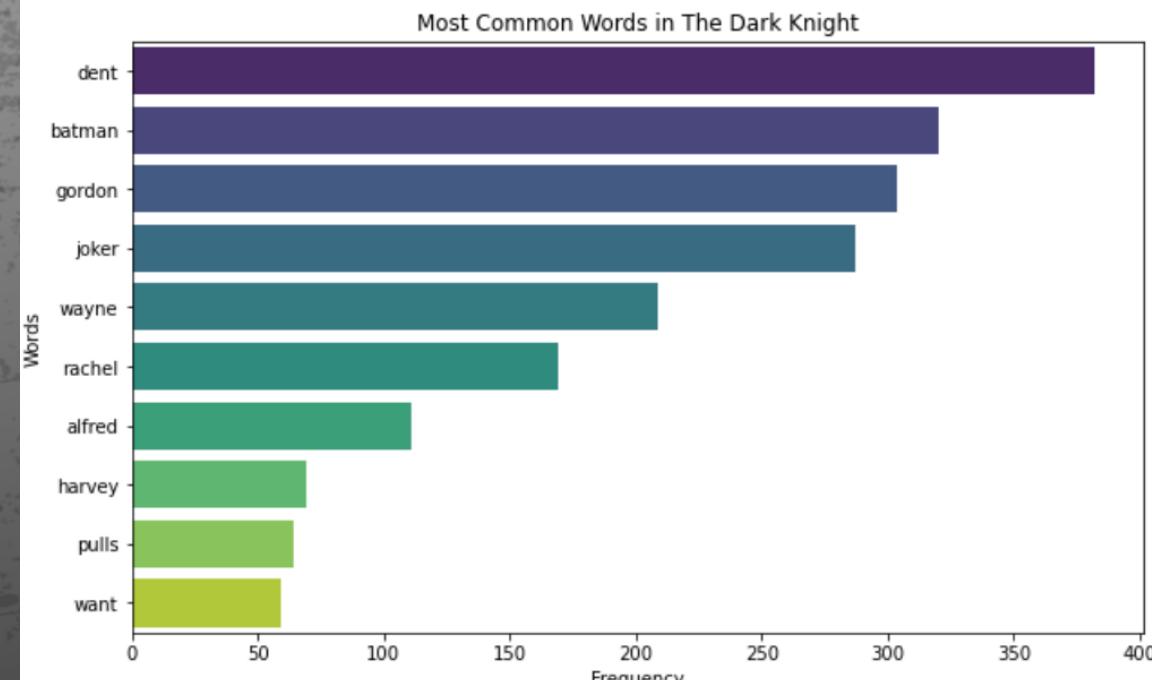
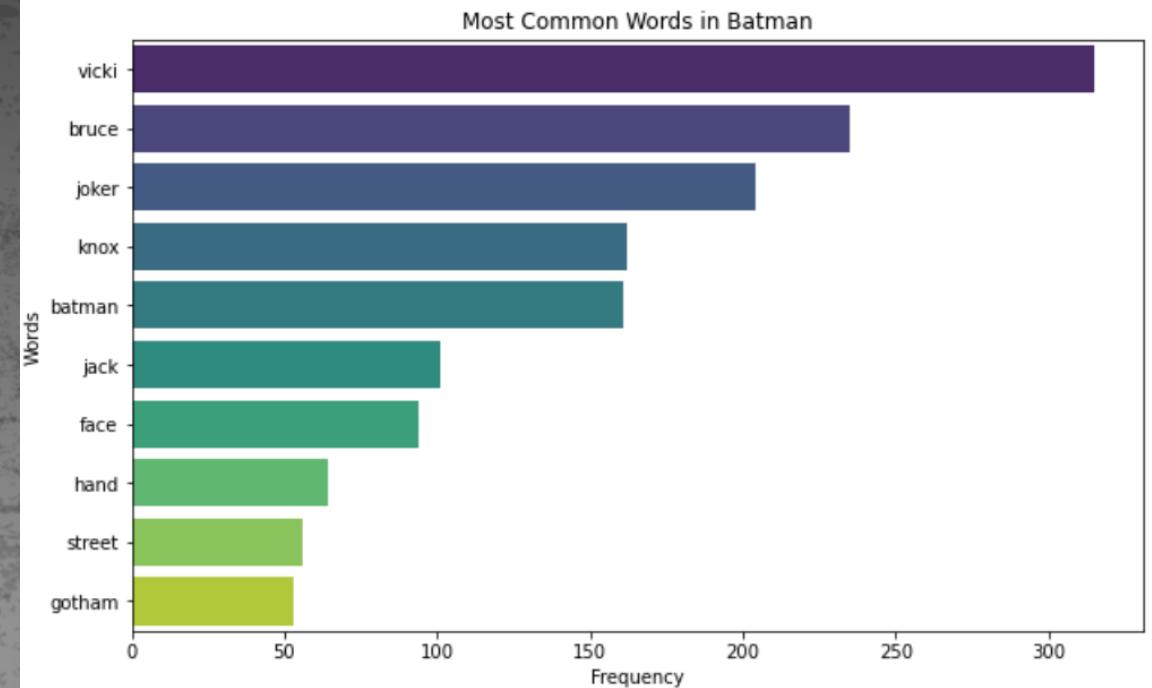
- Mean: 4.41
- Median: 4.25
- Standard Deviation: 1.15

*The Dark Knight*

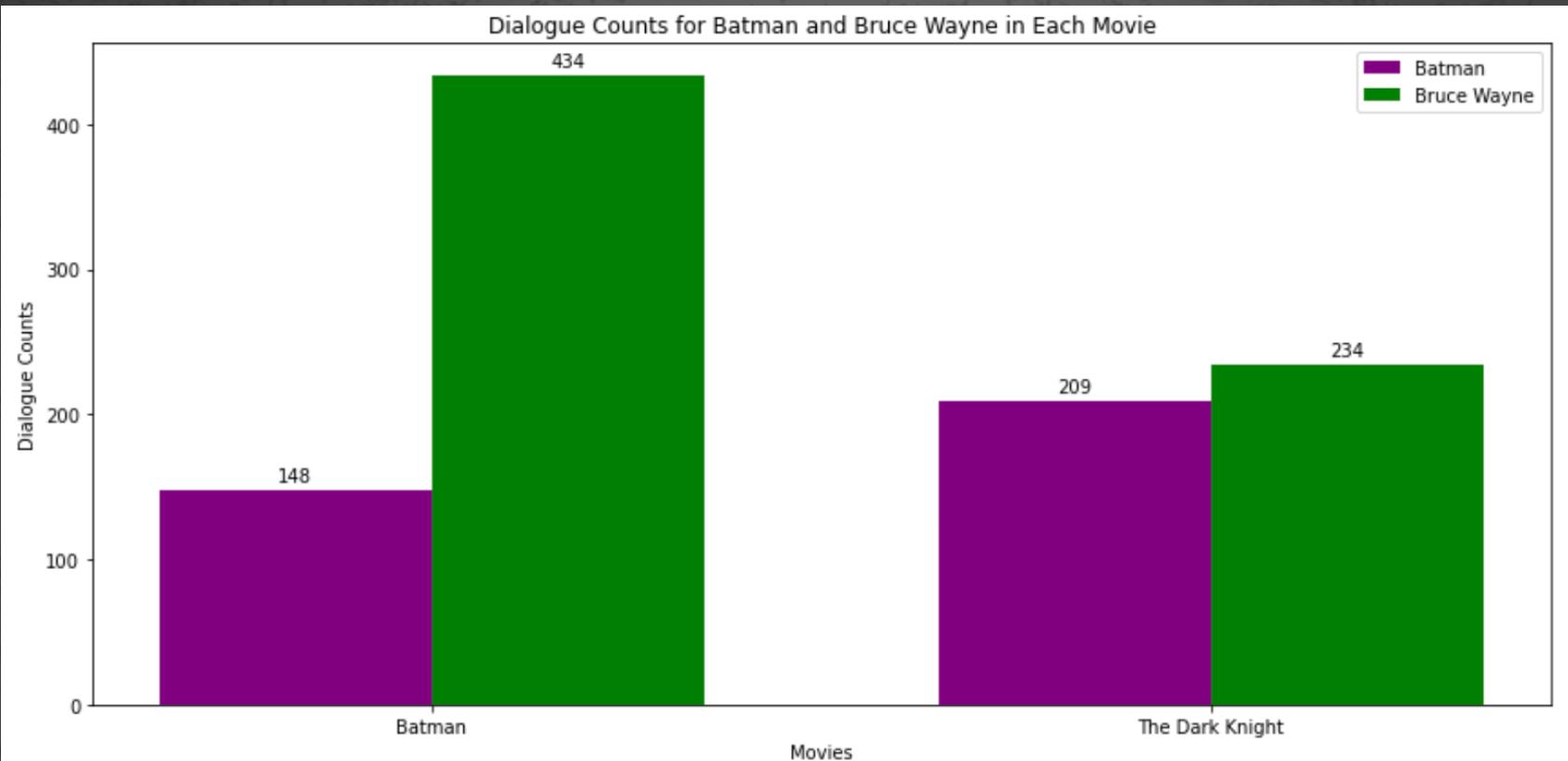
- Mean: 4.24
- Median: 4.09
- Standard Deviation: 1.10

*Batman* generally features longer words and dialogues with more variability in word count.

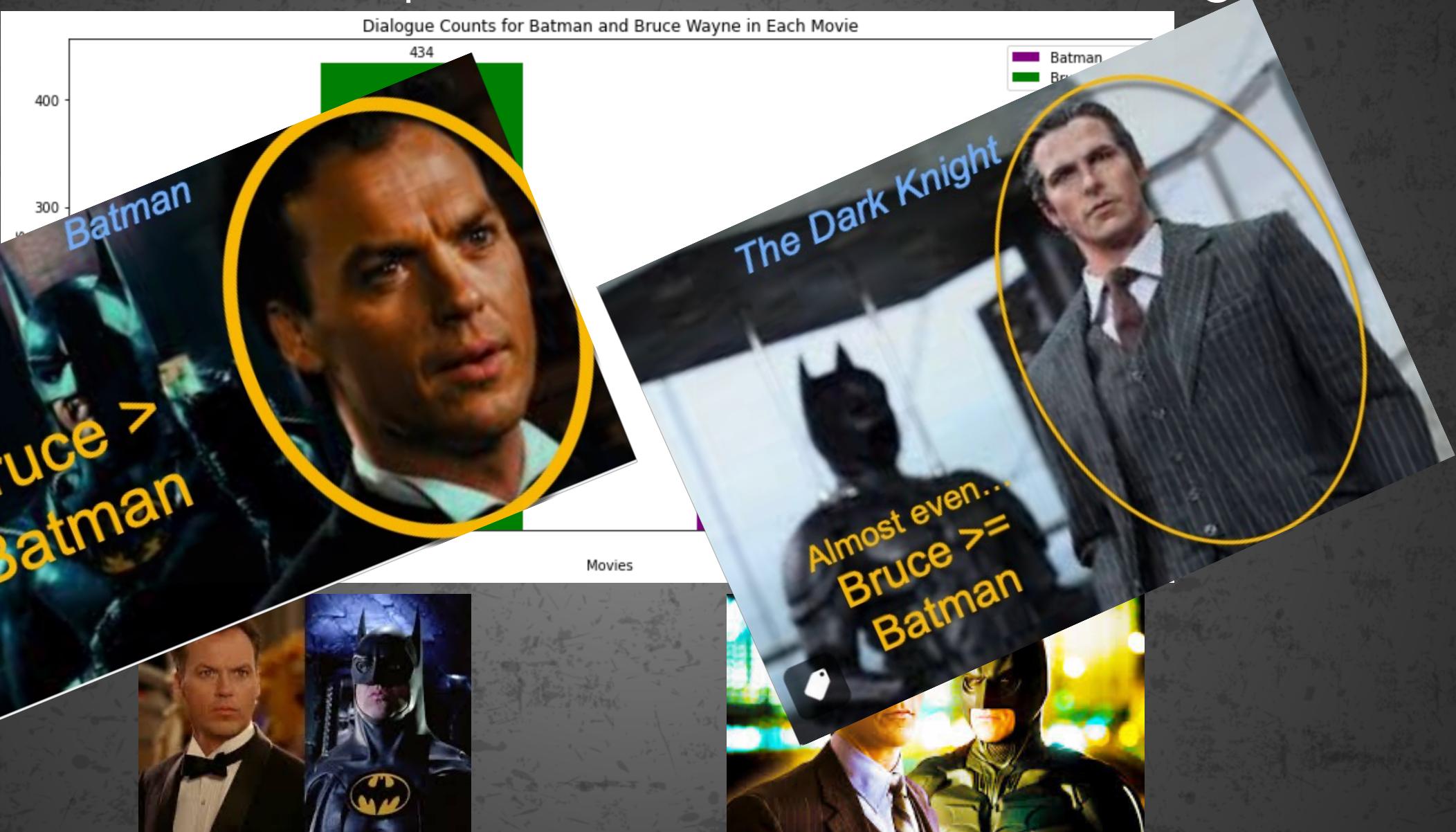
*The Dark Knight* has slightly more sentences per dialogue, with greater variability in sentence length.



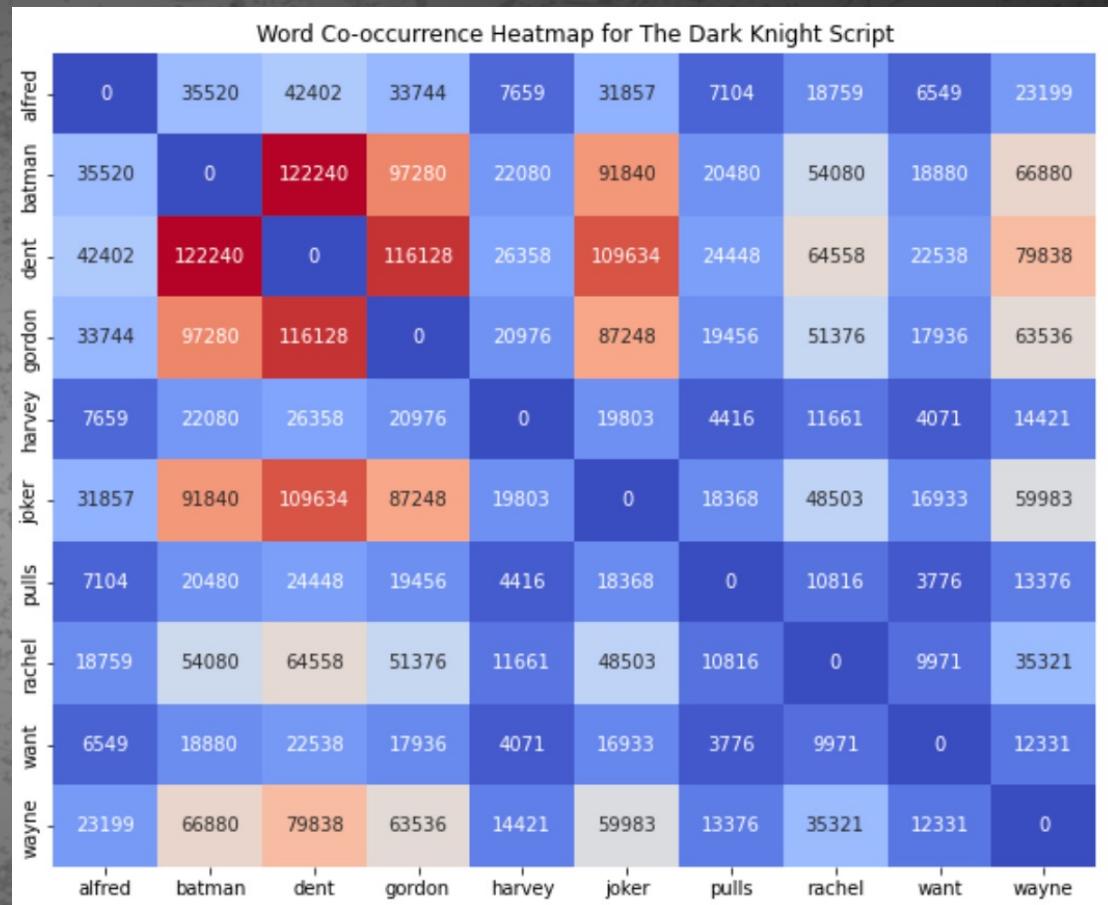
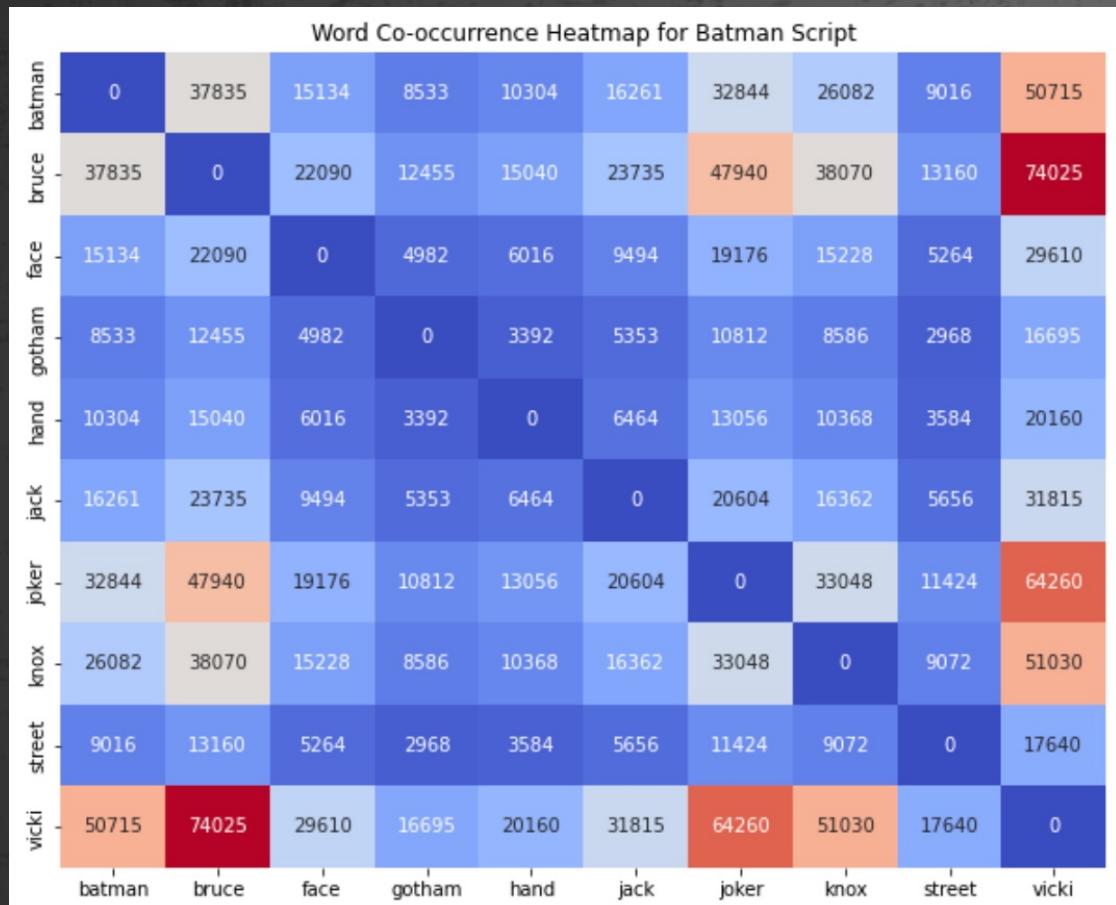
# EDA in Movie Scripts - Batman or Bruce Wayne?



# EDA in Movie Scripts - Batman or Bruce Wayne?



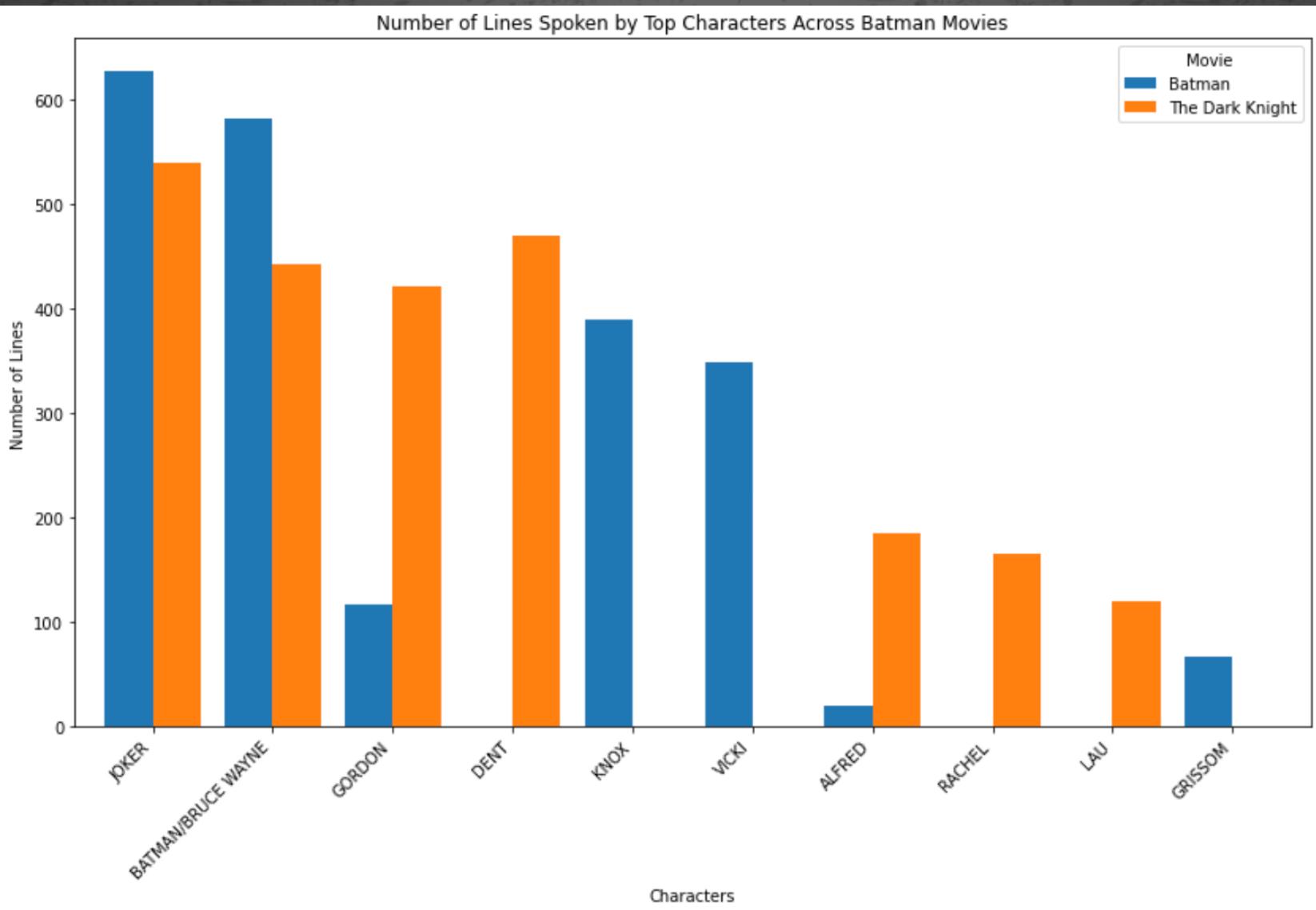
# EDA in Movie Scripts – Word Co-occurrence Heatmaps



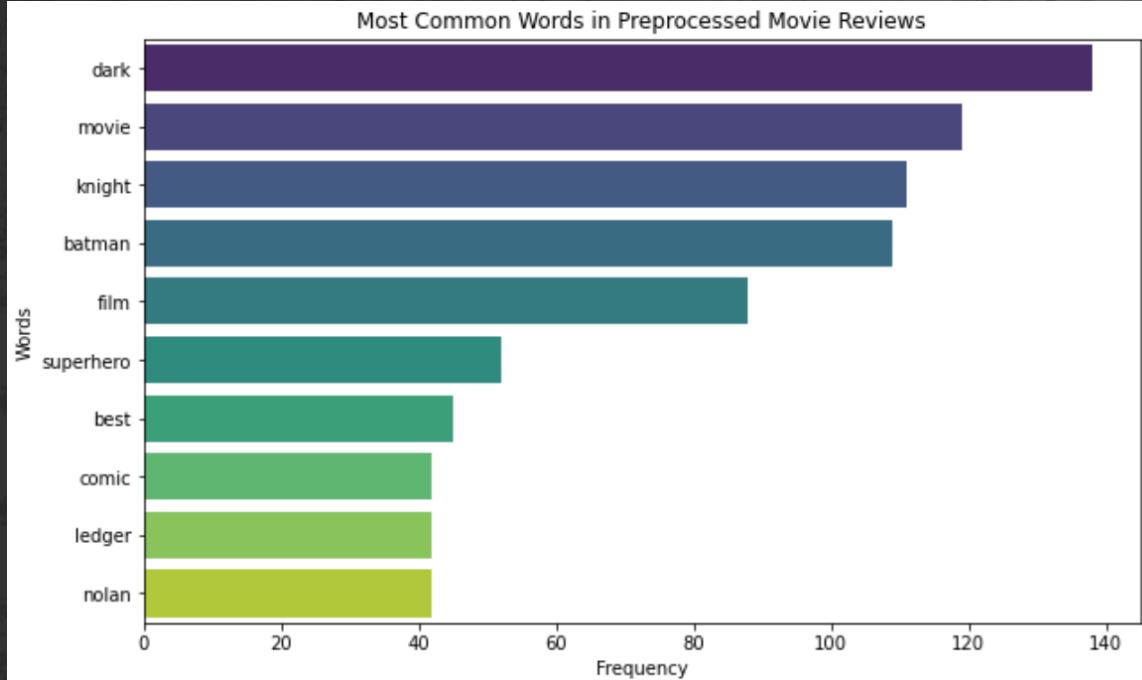
Characters like Bruce, Vicki, Dent, and Joker have strong interactions, as shown by the high co-occurrence frequencies.

The presence of frequent co-occurrences between key characters (e.g., Bruce and Vicki, Bruce and Joker) highlights pivotal relationships and interactions that drive the storyline.

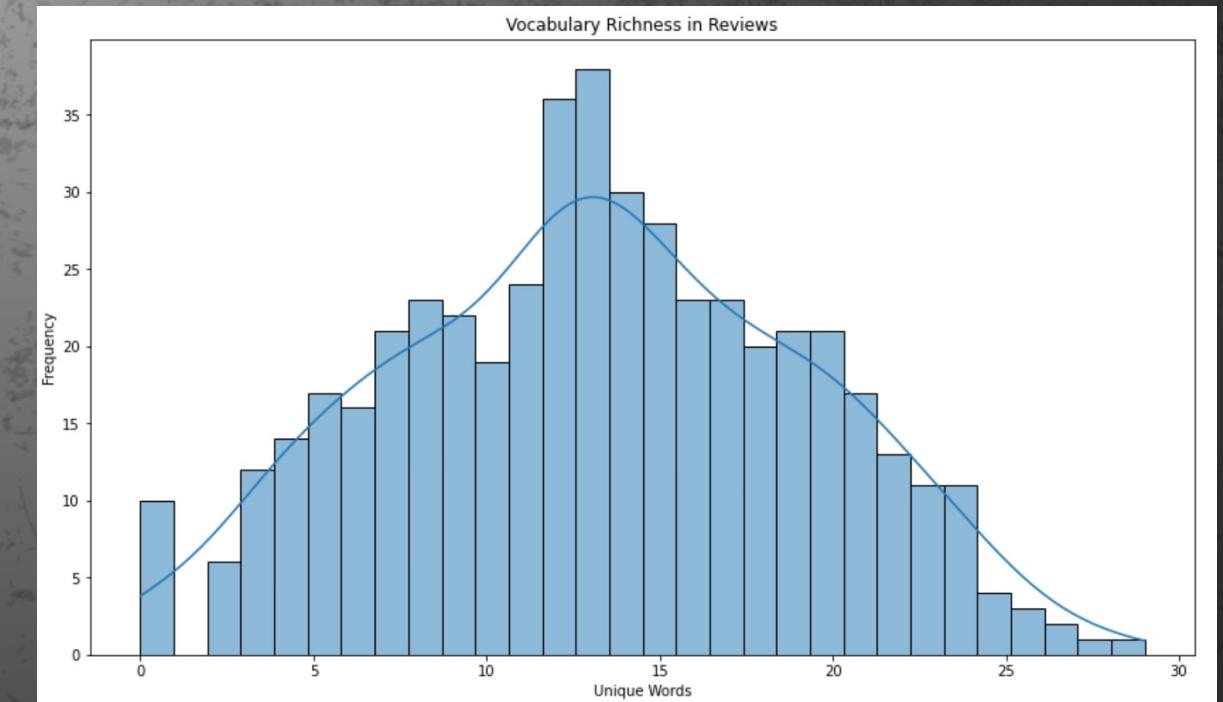
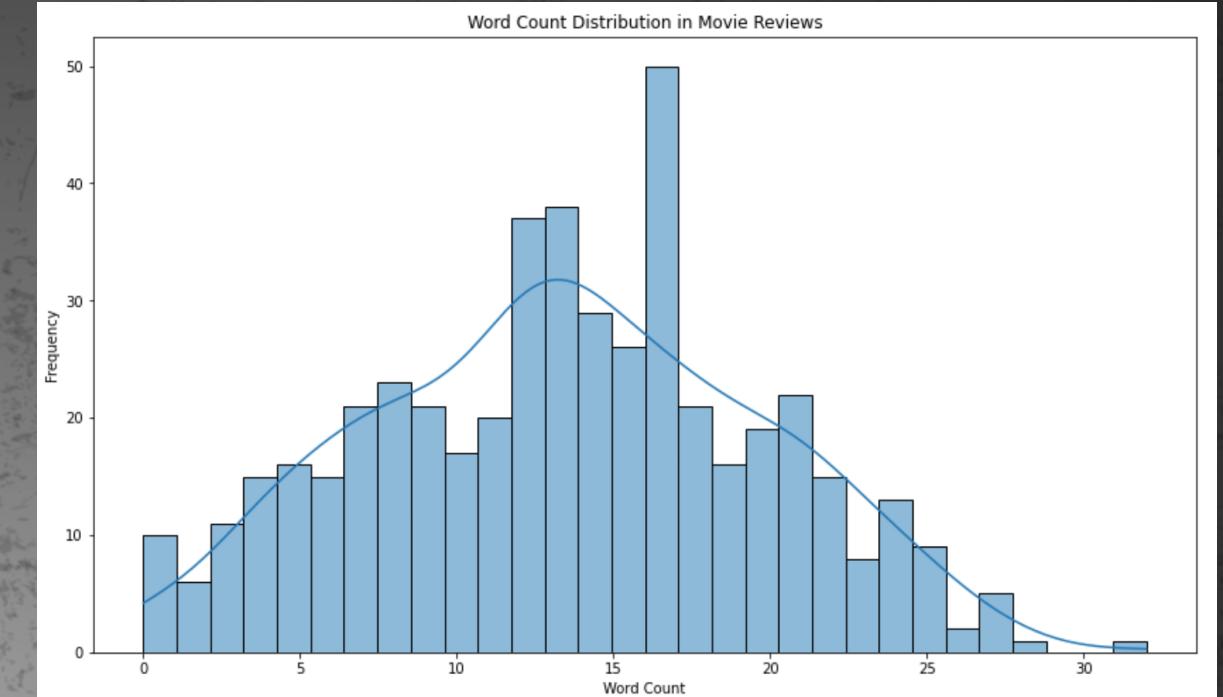
# EDA in Movie Scripts – focusing on dialogue



# EDA in Movie Reviews



Most common adjectives:  
Dark, best, good, great



# Sentiment Analysis in Movie Scripts

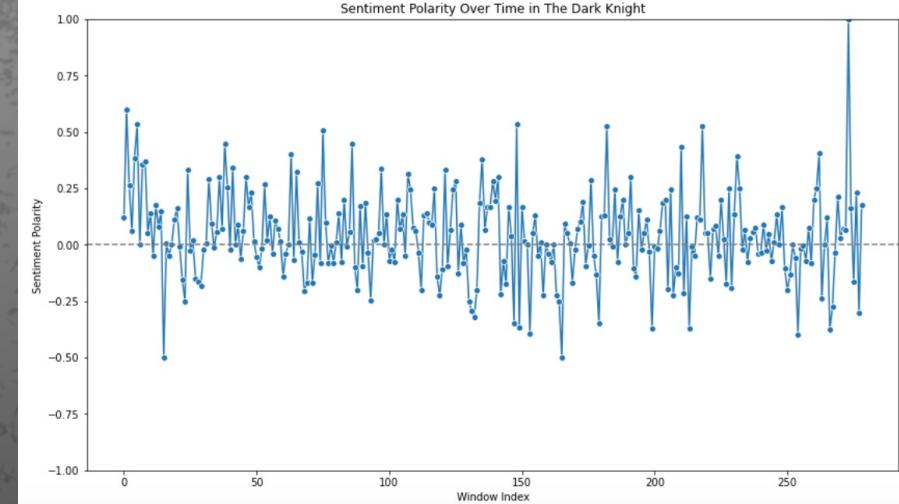
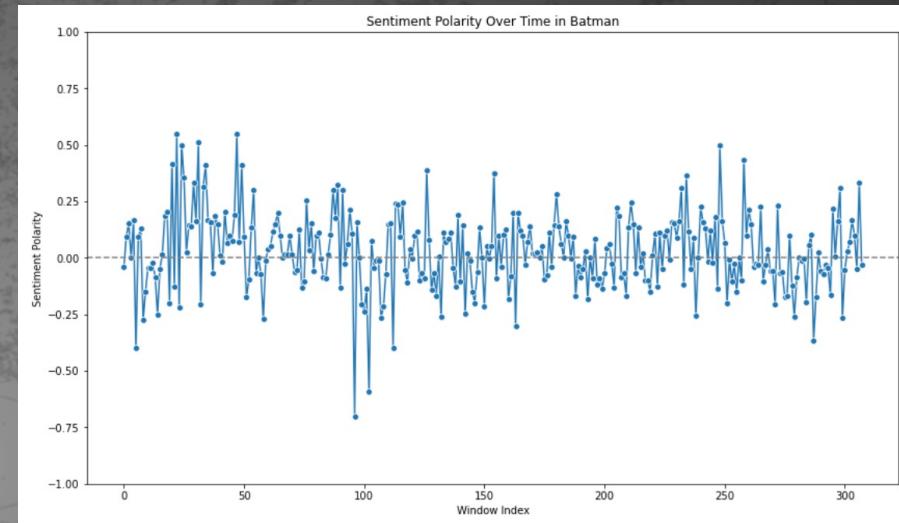
## Methodology:

### Using TextBlob for Sentiment Analysis:

- **TextBlob**: A Python library for processing textual data, providing simple APIs for NLP tasks.
- **Sentiment Analysis**: The process of computationally identifying and categorizing opinions expressed in text to determine whether the writer's attitude towards a particular topic is positive, negative, or neutral.

## Polarity and Subjectivity Scores for Preprocessed Text

- **Polarity Score**: A float within the range [-1.0, 1.0], where -1.0 indicates a very negative sentiment, 0 indicates a neutral sentiment, and 1.0 indicates a very positive sentiment.
- **Subjectivity Score**: A float within the range [0.0, 1.0], where 0.0 indicates very objective text and 1.0 indicates very subjective text.



# Sentiment Analysis in Movie Scripts

## Results:

### *Batman*

- Polarity: 0.027875
- Subjectivity: 0.503305

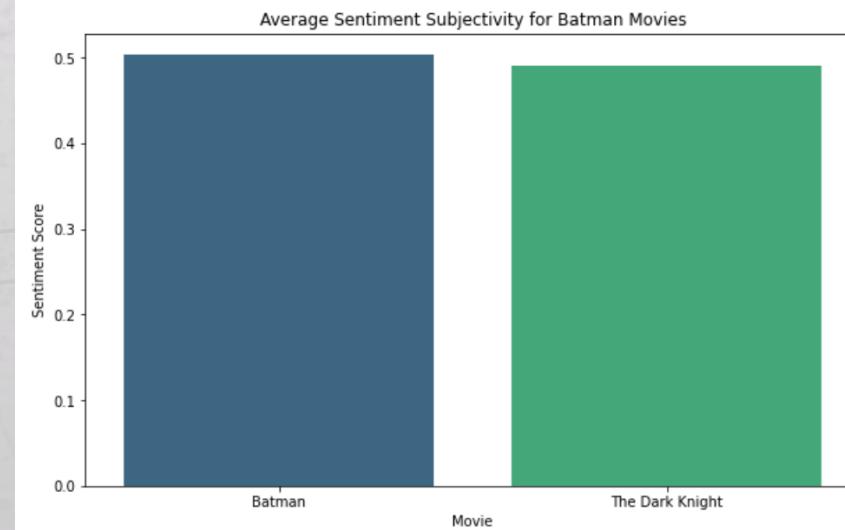
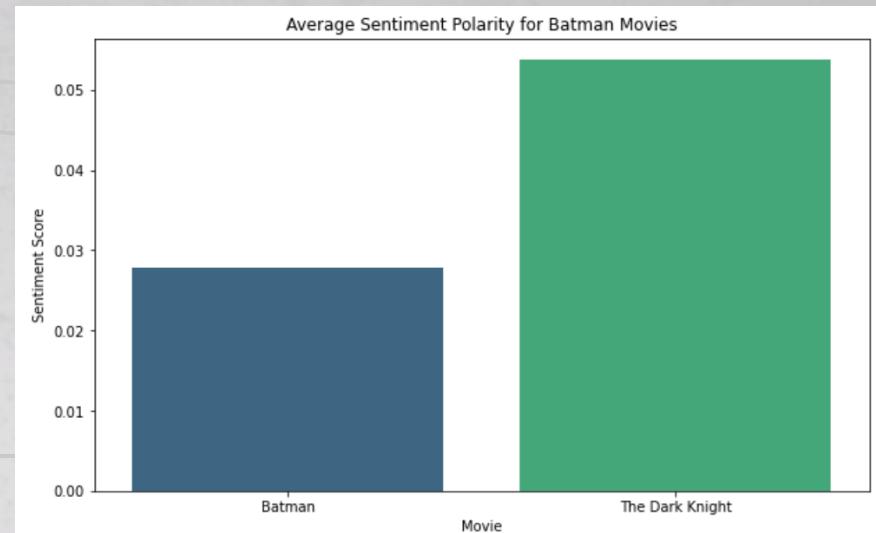
### *The Dark Knight*

- Polarity: 0.053719
- Subjectivity: 0.491237

## Summary:

*Batman*: Exhibits a slightly positive polarity and a moderately high subjectivity, indicating a balanced tone with some personal and opinionated elements.

*The Dark Knight*: Shows a more positive polarity and slightly lower subjectivity, indicating a more factual and objective tone but still with some positive sentiment.



# Sentiment Analysis in Movie Reviews

## Results – Average Sentiment Scores:

*Batman*

Polarity: 0.132775, Subjectivity: 0.503305

*The Dark Knight*

Polarity: 0.108812, Subjectivity: 0.491237

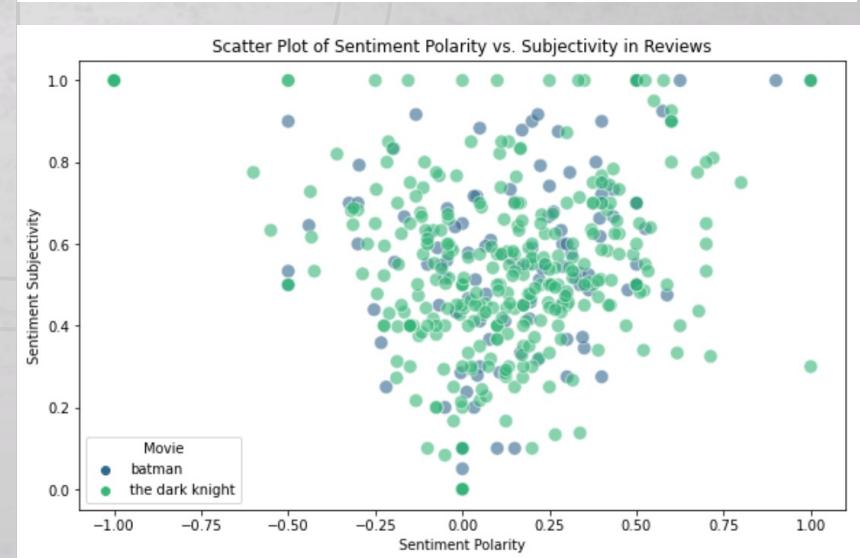
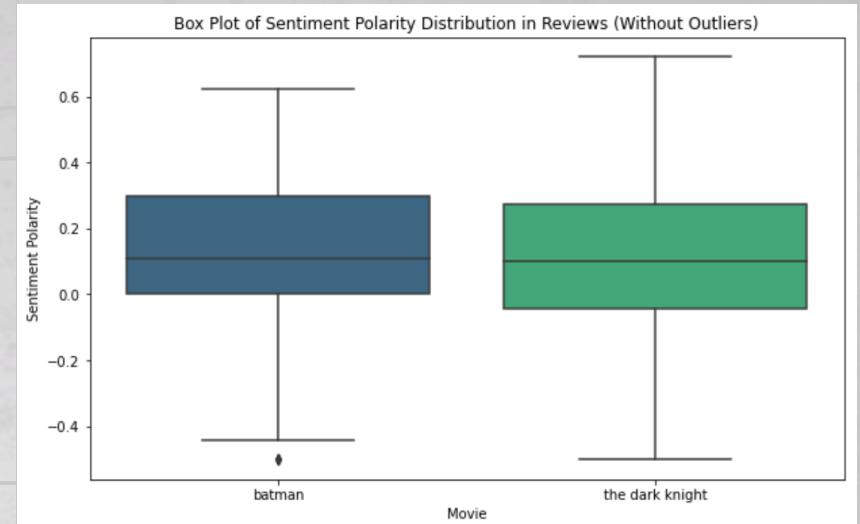
## Summary:

*Batman*

- Polarity: Slightly higher positive sentiment.
- Subjectivity: Slightly more subjective, indicating more personal and opinionated reviews.

*The Dark Knight*

- Polarity: Slightly lower positive sentiment.
- Subjectivity: Slightly more objective, indicating more factual reviews.



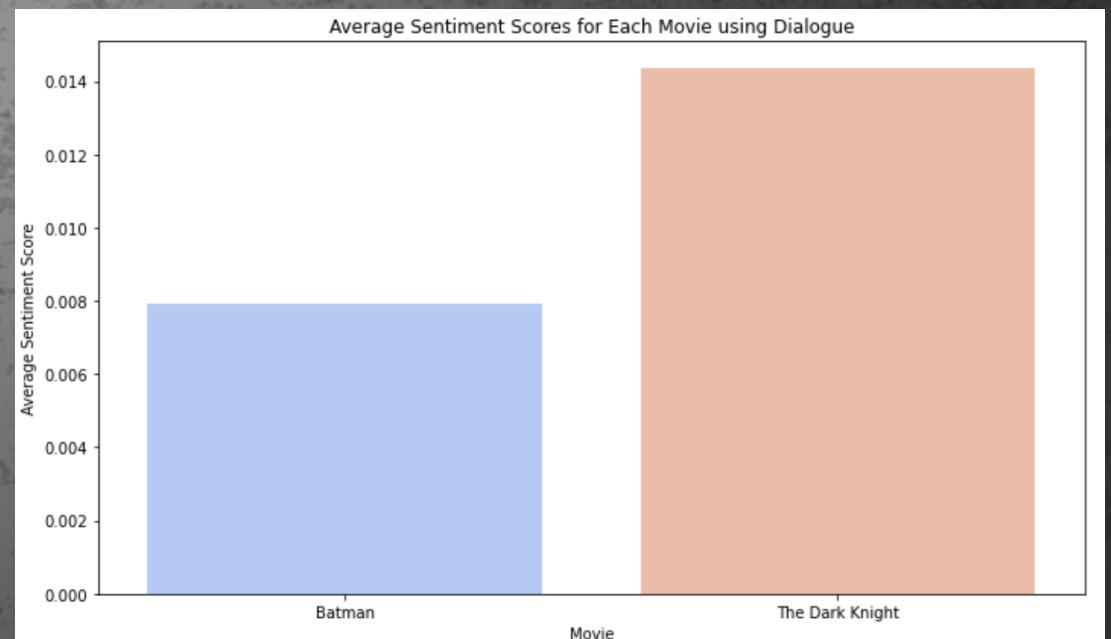
# Sentiment Analysis in Dialogue

## TextBlob and VADER Sentiment Analysis:

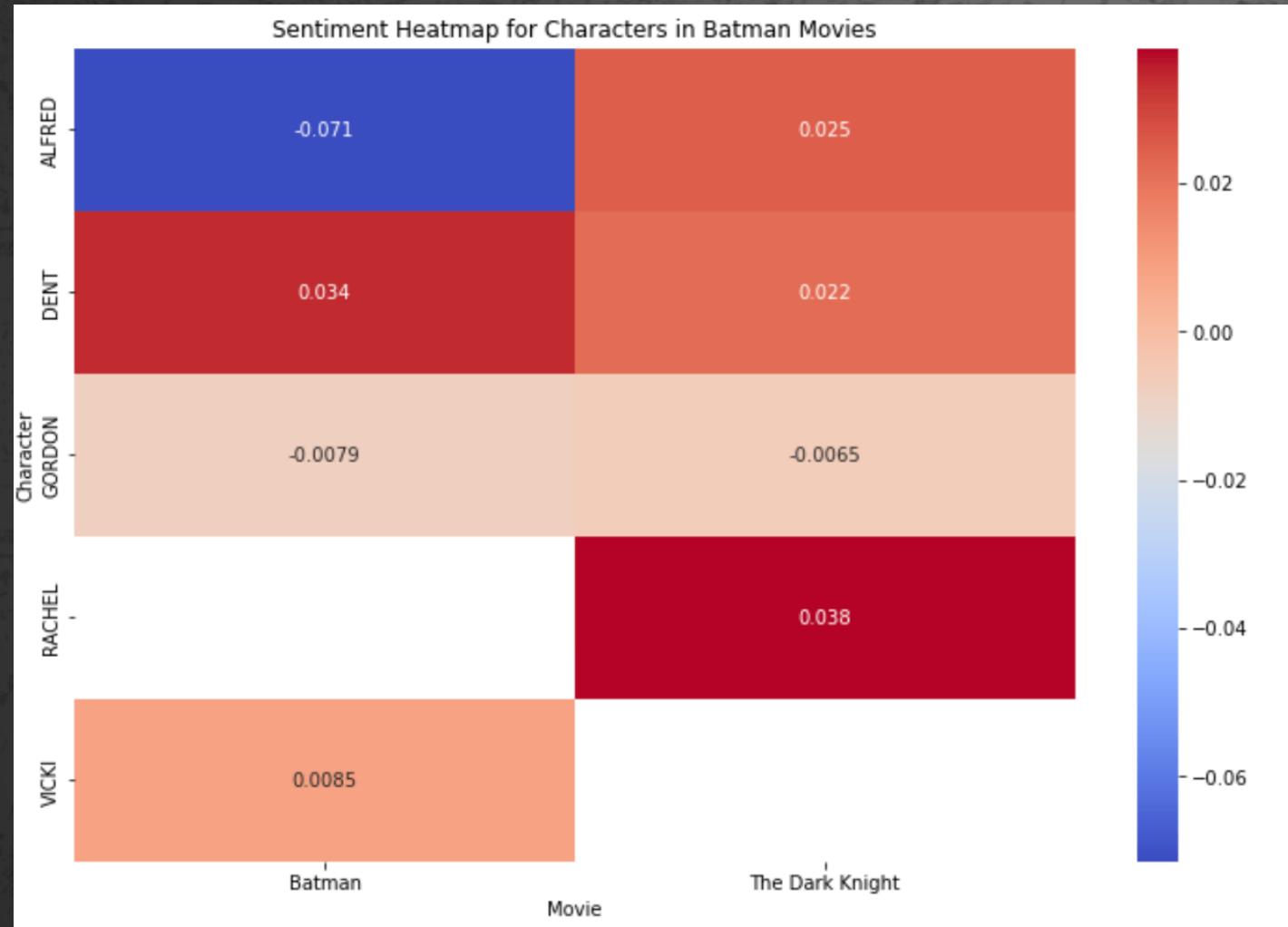
- TextBlob: Provides polarity and subjectivity scores for dialogues. Uses a predefined dictionary of words associated with positive and negative sentiments.
- VADER (Valence Aware Dictionary and sEntiment Reasoner): A lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media and informal text. Provides detailed sentiment scores including positive, neutral, negative, and compound scores.

## Comparison of Sentiment Scores for Characters:

- Analyzed sentiment for key characters in *Batman* and *The Dark Knight* using both TextBlob and VADER.
- Compared the results to understand differences in sentiment perception.



# Sentiment Analysis with Supporting Characters



Characters in *The Dark Knight* generally have more positive sentiment scores compared to those in *Batman*.

Alfred's sentiment is negative in *Batman* and positive in *The Dark Knight*.

Rachel in *The Dark Knight* has the highest positive sentiment score, indicating a very favorable sentiment in her dialogues.

# Sentiment Analysis in Dialogue – Results

TextBlob Sentiment Scores for Main Characters:

*Batman*

BATMAN: 0.0189, JACK: 0.0292, BRUCE WAYNE: 0.1703,  
JOKER: 0.0136

*The Dark Knight*

JOKER: 0.0047, BATMAN: 0.0241, BRUCE WAYNE: 0.0295

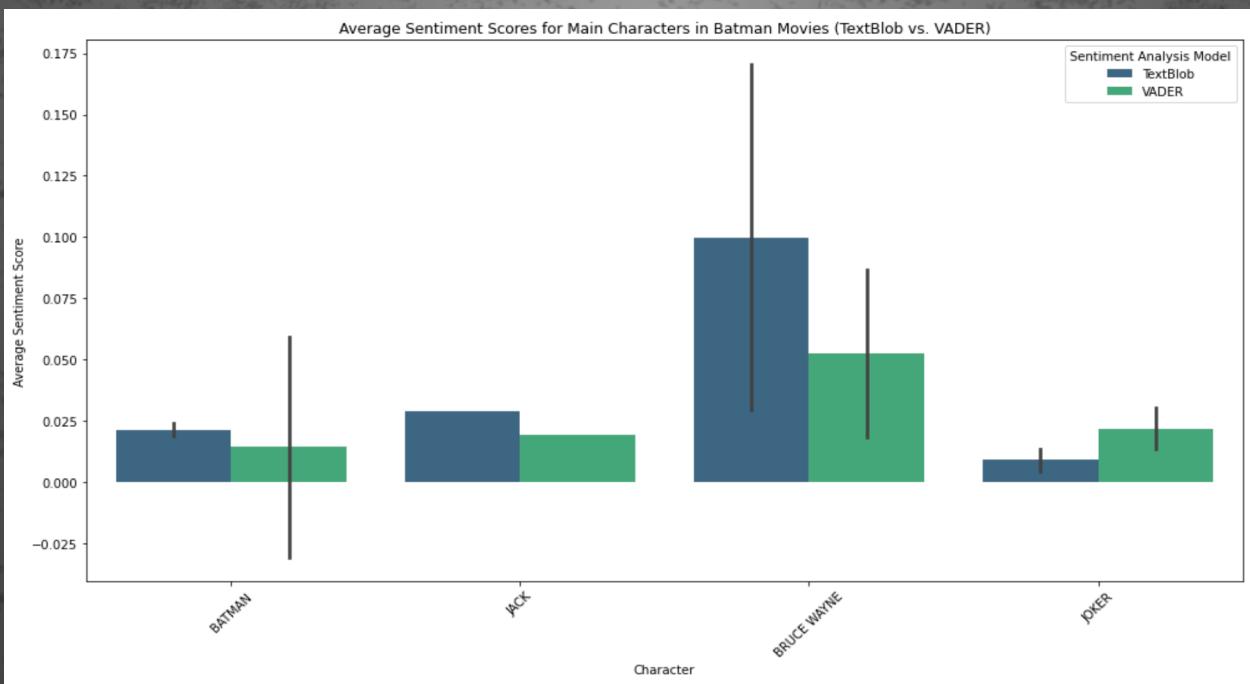
VADER Sentiment Scores for Main Characters:

*Batman*

BATMAN: 0.0593, JACK: 0.0196, BRUCE WAYNE: 0.0868,  
JOKER: 0.0307

*The Dark Knight*

JOKER: 0.0134, BATMAN: -0.0303, BRUCE WAYNE: 0.0184

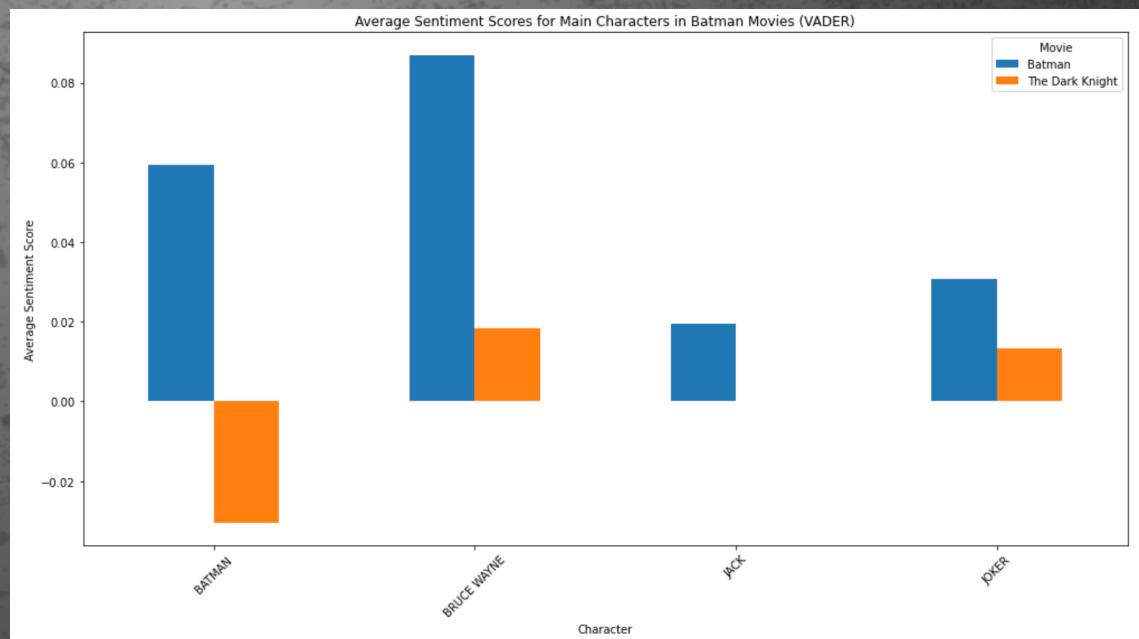
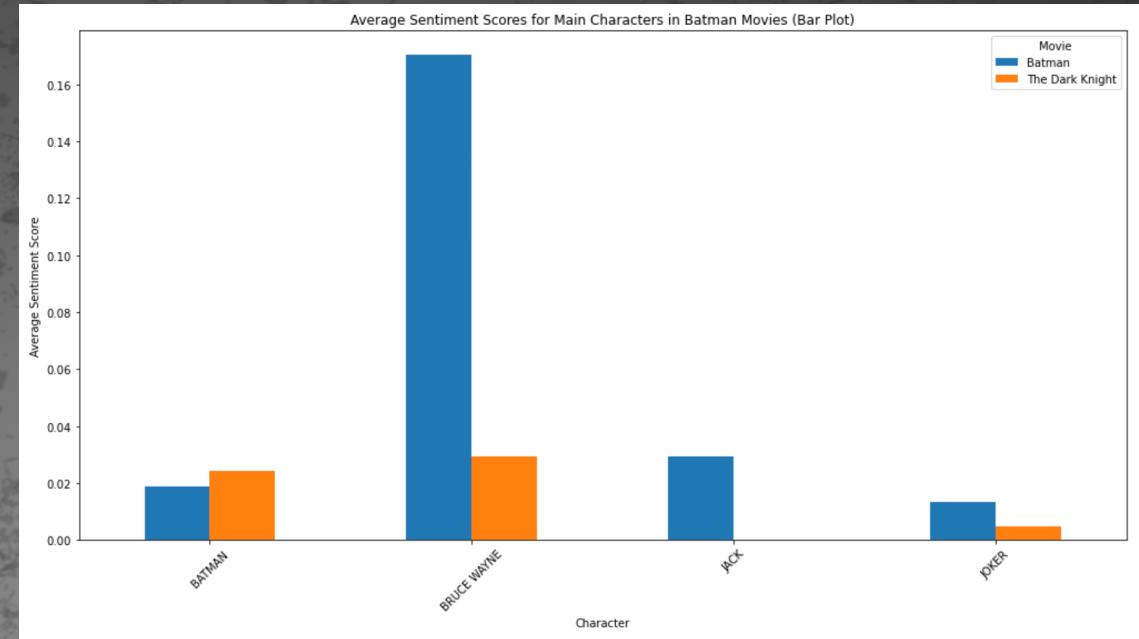


## TextBlob

- Bruce Wayne in *Batman* has the highest positive sentiment score (0.1703), indicating a more positive tone in his dialogues
- The Joker has neutral sentiment scores in both movies, with slightly positive scores but generally close to neutral.
- Batman has neutral to slightly positive sentiment scores in both movies, indicating a balanced tone in his dialogues.

## VADER:

- Bruce Wayne in *Batman* with a positive sentiment score (0.0868), but lower than TextBlob's score.
- The Joker has neutral sentiment scores in both movies
- Batman has a slightly negative sentiment score (-0.0303) in *The Dark Knight* contrasting with TextBlob's slightly positive score.



# The Joker is... neutral?

The Joker's dialogues, as analyzed through TextBlob and VADER, show a neutral or slightly positive sentiment score... but how?

The Joker is widely recognized as a dark and malevolent villain!

- **Complexity of Language:** The Joker often uses complex and nuanced language that can mask the underlying dark intent. His dialogues are laced with sarcasm, irony, and a playful tone that may be interpreted by sentiment analysis models as neutral or slightly positive.
- **Context and Delivery:** The emotional impact of the Joker's words is heavily influenced by the context in which they are delivered and his performance, including tone, facial expressions, and body language. These aspects are not captured by text-based sentiment analysis models.
- **Dark Humor:** The Joker's dialogues often contain dark humor, which can be perceived as humorous or playful. This can skew the sentiment analysis towards neutral or positive, even though the content is dark and disturbing.

Given these factors, the neutral or slightly positive sentiment scores for the Joker's dialogues can be understood as a limitation of the sentiment analysis models in capturing the full context and nuance of his speech.



# The Reviews say... which movie is the best?

## Machine Learning and Sentiment Polarity

Using Naive Bayes Classifier:

- Classified reviews into positive, negative, and neutral sentiments based on polarity scores.

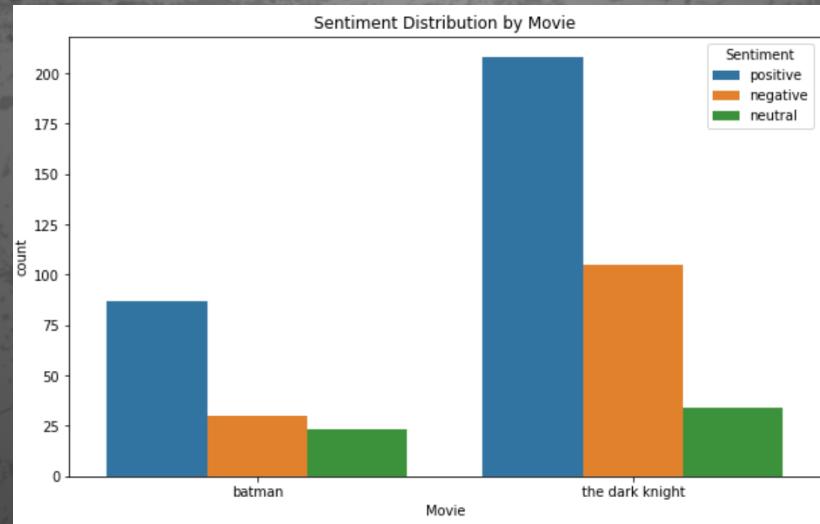
Sentiment Polarity Analysis Based on Reviews:

- Calculated the average sentiment polarity for each movie to determine the overall sentiment expressed in the reviews.
- Compared sentiment polarity between *Batman* and *The Dark Knight*.

## Initial Sentiment Polarity Analysis

*Batman*: Average sentiment polarity of 0.132296, suggesting generally positive reviews.

*The Dark Knight*: Average sentiment polarity of 0.108523, also suggesting positive reviews, but slightly less favorable.

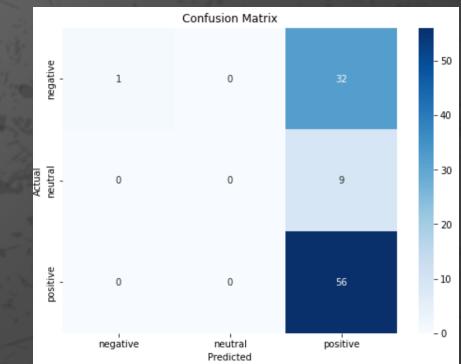


## Initial Model Performance

Accuracy: 58.16%

Confusion Matrix:

- Negative: 1 correctly predicted, 32 misclassified as positive.
- Neutral: None correctly predicted, all misclassified as positive.
- Positive: All correctly predicted.



# The Reviews say... which movie is the best?

## Machine Learning and Sentiment Analysis

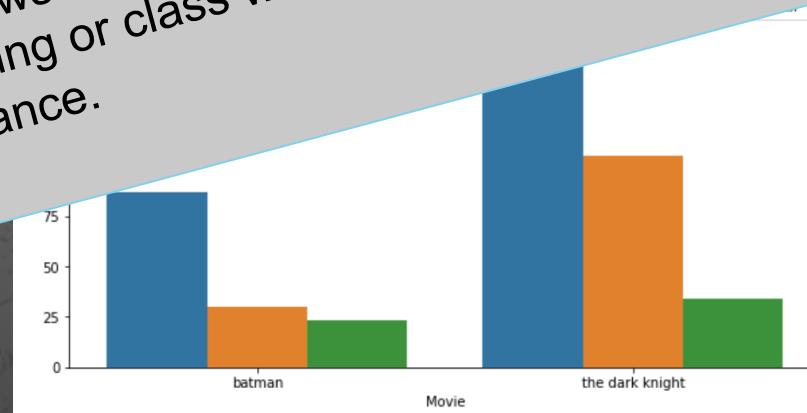
### Initial Model Performance: 58.16%

- High accuracy for positive class.
- Poor performance for negative and neutral classes.
- Bias towards positive class due to higher support (56 out of 98 reviews).
- Struggles to differentiate negative and neutral sentiments from positive ones.

### Class Imbalance:

- Many more positive reviews than negative or neutral ones.
- Techniques like resampling or class weighting needed to improve model performance.

Shared sentiment polarity between *Batman* and *The Dark Knight*.

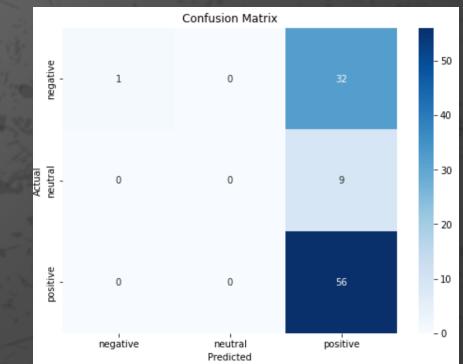


### Initial Model Performance

Accuracy: 58.16%

Confusion Matrix:

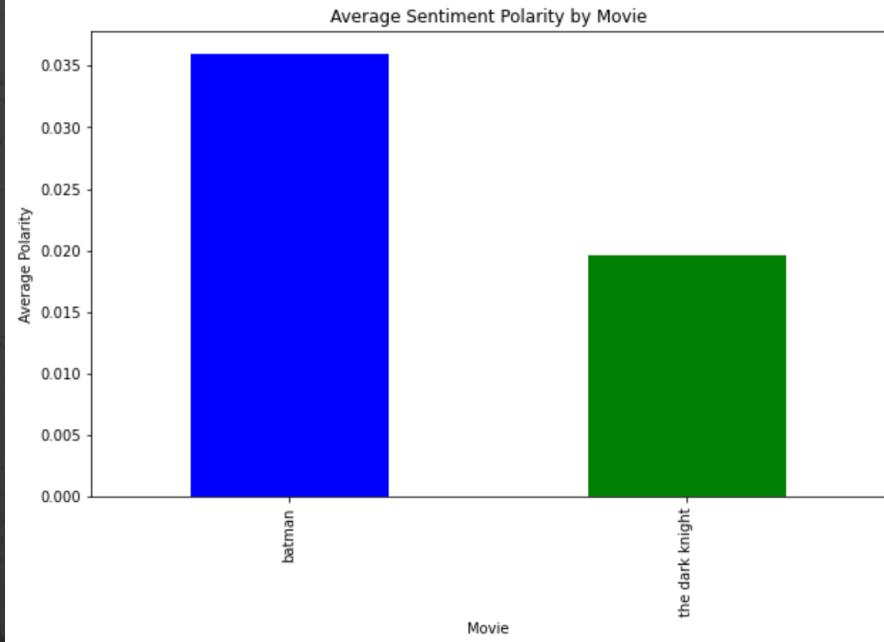
- Negative: 1 correctly predicted, **32 misclassified** as positive.
- Neutral: None correctly predicted, **all misclassified** as positive.
- Positive: All correctly predicted.



# The Reviews say... which movie is the best?

## Improved Machine Learning Approach

Addressing Class Imbalance:  
Resampled minority classes  
(negative and neutral) to  
balance the dataset.



## Best Movie Based on Overall Review Sentiment Analysis

*Batman* with a slightly higher average sentiment polarity.

The improved model significantly outperforms the initial model.

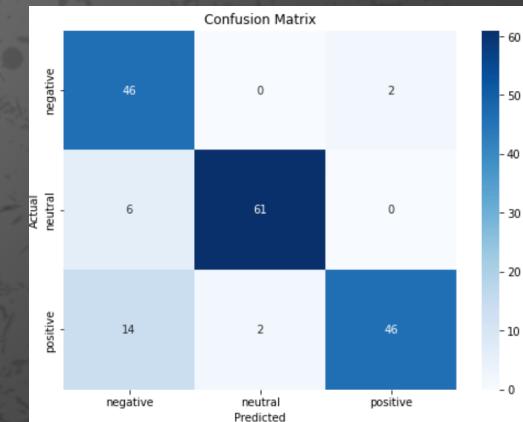
Addressing class imbalance through resampling greatly enhances model performance.

## Improved Model Performance

Accuracy: 86.44%

Confusion Matrix:

- Negative: 46 correctly predicted, 2 misclassified as positive.
- Neutral: 61 correctly predicted, 6 misclassified as negative.
- Positive: 46 correctly predicted, 14 misclassified as negative.



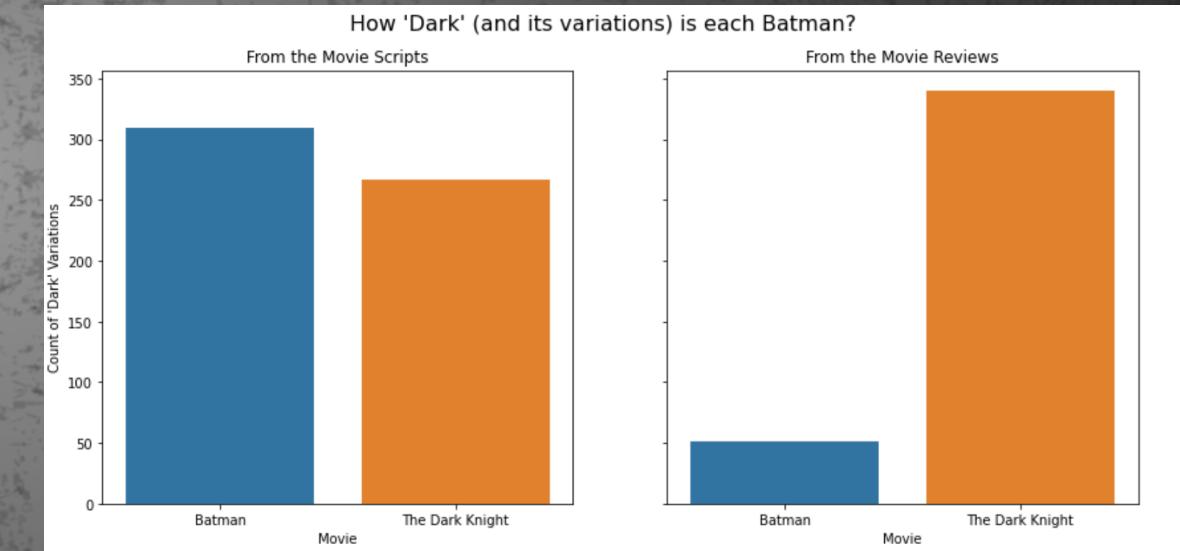
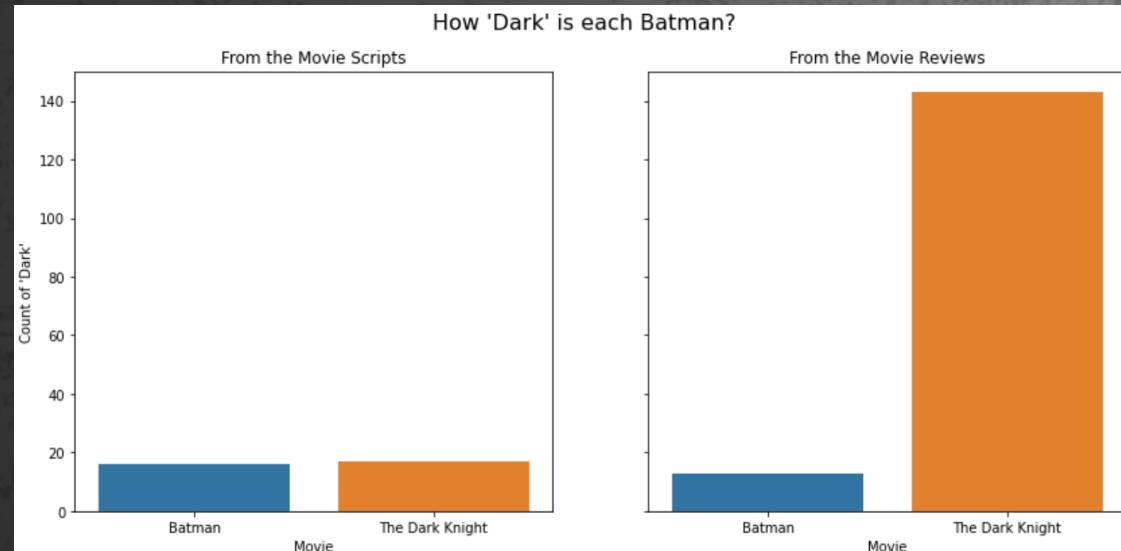
# Let's get literal...

# Which Movie is the "Darkest"?

## Method

Counting occurrences of the word "dark" and its variations in movie scripts and reviews.

Included a second analysis with terms like gloomy, shadow, grim, sinister, eerie, etc.



## Conclusion

The Dark Knight is the darkest movie based on the number of dark-related terms in reviews.  
Batman shows a higher count of dark variations in scripts.

# Conclusions

**Best Movie Based on Overall Review Sentiment Analysis:**

*Batman* - slightly higher average sentiment polarity.

**Darkest Movie Based on Sentiment Polarity in Reviews:**

*The Dark Knight* - higher occurrences of dark-related terms in reviews.

**Darkest Movie Based on Script Analysis:**

*Batman* - higher count of dark variations in scripts.

**Most Subjective Movie Based on Sentiment in Reviews:**

*The Dark Knight* - higher subjectivity scores indicating more personal and opinionated reviews.

**Most Subjective Movie Based on Sentiment in Scripts:**

*Batman* - higher subjectivity scores indicating more personal and opinionated dialogues.



5

# Final Recommendation

## Popularity:

Draw inspiration from *Batman* to maintain broad appeal.



## Character Complexity and Darkness:

Emphasize elements from *The Dark Knight* for a darker tone and more intricate character dynamics.



## Balance!

Combine the strengths of both movies to create a balanced and compelling portrayal in future Batman productions.



# Future Steps and Improvements

1. **Explore Character Interactions:** Create conversation maps to visualize character dynamics and interactions. Analyze dialogue to understand the flow and intensity of character relationships.
2. **Evaluate Sentiment Analysis:** Improve sentiment analysis models to better detect sarcasm and tone. Address the neutral sentiment results of characters like The Joker, which may be due to sarcasm and jokes.
3. **Custom Sentiment Lexicon:** Develop and use a custom sentiment lexicon tailored to the unique language and tone of Batman movies. Train and evaluate models using this custom lexicon to improve accuracy.
4. **Character-Specific Sentiment Analysis:** Perform sentiment analysis on a per-character basis to capture individual character tones and emotions. Compare and contrast sentiment dynamics between different characters.
5. **Temporal Analysis:** Conduct temporal analysis of sentiment to observe how character emotions and interactions evolve over the course of the movie.
6. **Additional Movie Analysis:** Extend the analysis to include additional Batman movies for comparative insights on character development and narrative style.

# CREDITS

Thomas, E., Nolan, C., & Goyer, D. (2009). *The dark knight directed by Christopher Nolan.* France; Warner Brothers.

Peters, J., Guber, P., & Hamm, S. (1989). *Batman.* United States; Warner Bros.

Presentation Template: [SlidesMania](#)