

Final project Group 38 Project 2

**Arpan Saha¹(261022455), Danning Li¹(261019485),
Tahsin Kabir Etul(261020215)¹, Philippe Latour¹(261022386)**

¹McGill University

arpan.saha@mail.mcgill.ca, danning.li2@mail.mcgill.ca, tahsin.etul@mail.mcgill.ca, philippe.latour2@mail.mcgill.ca

Introduction

In an age dominated by digital discourse, our team conducted a comprehensive exploration of the news media landscape for the latest cinematic release, **Napoleon**. Commissioned by a discerning media company, our primary focus is to unravel the nuances of visibility and reception of the cinematic release of this movie, particularly in comparison to four other contemporaneously released acclaimed movies.

As an integral part of our methodology, we curated a robust dataset comprising over 500 news articles, meticulously sourced from NewsAPI.org. To ensure an unbiased representation of the movies we are focusing on, we implemented stringent filters that would yield articles with a high likelihood of relevance to the selected movies, and all content was strictly in English.

This meticulously curated dataset contains articles from a variety of media outlets, providing valuable insights into each film's thematic focus, visibility, and reception. Our approach involves a complex open coding process in which we dissect the titles and openings of 200 articles to identify the primary focus and distinct themes of each. Each article is carefully assigned to a topic, providing the basis for an in-depth understanding of the media discourse surrounding these films. Through manual annotation, we categorize the entire dataset into 6 different themes, enabling a nuanced exploration of the public's stance on the release of these films.

To further enrich our analysis, we employed advanced numerical statistics, specifically leveraging Term Frequency-Inverse Document Frequency (TF-IDF). By computing TF-IDF scores across all 500 posts, we aim to characterize each identified category by extracting the top 10 words with the highest TF-IDF scores, thus offering a quantitative perspective on the salient themes in media coverage.

As we delve into the intricacies of this analysis, our overarching goal is to provide our client with actionable insights into the public perception, thematic focus, and relative visibility of "Napoleon" in comparison to its contemporaries. By dissecting the topics

that resonate most with the public, measuring relative engagement, and assessing sentiment, we aim to contribute valuable insights to the broader discussion of contemporary film.

Data

We collected news articles data from November 7th to December 4th, 2023, encompassing a critical period for media coverage. In total, 600 articles were gathered for analysis. The movies selected for the study were:

1. The Marvels - (Genre: Action)
2. The Hunger Games: Ballad of Songbirds and Snakes - Action
3. Five Nights at Freddy's - Horror
4. Thanksgiving - Horror
5. Napoleon (Main Movie) - Historical Drama

To ensure a fair and representative sample, we utilized a filter in the form of the search query: "Marvels OR Napoleon OR (Five AND nights AND at AND Freddy's) OR (Hunger AND Games)." During the data collection procedure we hit a roadblock as the NewsAPI had a limit of returning only 100 articles. To overcome this limitation, we then ran the query six times over 5-day periods throughout the month to reach a sample size of 600 since each queries retrieved 100 articles. This filtering process aimed to reduce sampling bias in the selection of articles and ensure that the sample was genuinely reflective of the population of news related to the selected movies throughout the month.

We standardized the data by removing any results that are null or copies of existing articles in our dataset and structured the data from json to csv format so that it is easier for us to open code and annotate. Furthermore, we also shuffled the articles in our document so as to ensure a complete and effective open coding as the frequencies vary during certain days and weeks within that month.

Methods

The selection of these movies for the analysis involved a careful consideration of competing production houses and their movie releases during the specified timeframe. We chose to list the movies based on information obtained from sources external to NewsAPI such as IMDB and IGN. This decision aimed to maintain objectivity in the selection process and avoid unintentional skewing of the data.

The search query was deliberately constructed to be specific, including the use of logical operators to focus on the desired movies. By specifying the movies directly and considering competing productions, we aimed to strike a balance between inclusivity and relevance, extracting pertinent information while avoiding an overly broad search that might inundate the dataset with extraneous details.

In terms of data collection, the timeframe was chosen to capture a comprehensive snapshot of media coverage during the movie release period of our movie of interest, “Napoleon”. The choice of 600 articles ensured a robust dataset for analysis, providing sufficient depth for meaningful insights into the topics and coverage of the selected movies.

Results

To assess the visibility and reception of Napoleon, four key findings are required: a typology to classify each article, the characterization of each topic, the engagement by each movie with each topic, and the coverage of each movie

Selected Topics and their Definition

The following is an overview of the typology document created for annotating the data.

Box Office

The movie’s commercial performance. This topic is assigned to any article which contains more than one mention of any of the following:

- a. The words “box office” or similar derivatives
- b. Budget of the movie.
- c. Profit or revenue of the movie.

Cast

Information about cast members not related to their performance in the movie. This topic is assigned to any article which contains one or more mention of any of the following:

- a. An interview with a cast member
- b. A cast member’s future career

Interpersonal relations between a cast member and others.

LoreA

Represents contextual information about the movie’s plot or production. This topic is assigned to any article containing one or more mention of any of the following:

- a. Real historical information related to the movie’s characters, plot or setting.
- b. Background information on movie’s plot
- c. Conjectures about the movie’s fictional world

Production

Represents the technical elements of the movie’s making. This topic is assigned to any article containing two or more mentions of any of the following:

- a. Cinematography
- b. Directing
- c. Choice of soundtrack
- d. Editing, or visual effects

Writing

Represents the movie’s characters, plot, and dialog. This topic is assigned to any article containing one or more mentions of the following:

- a. A character’s physical or intellectual features
- b. The movie’s storyline
- c. Dialog between characters if the delivery of said dialog is not mentioned.

Delivery

Represents any performance work in the movie. This topic is assigned to any article containing one or more mentions of any of the following:

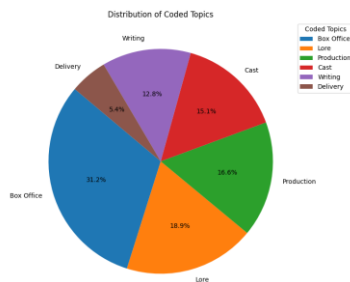
- a. The delivery of dialog by an actor.
- b. The portrayal of a character by an actor.
- c. The musical performance of the soundtrack.

Characterization of Topics

Each topic is characterized by taking the top 10 words with the highest TF-IDF score across the text from articles labelled with the topic. While this did yield interesting results, not all 10 words were useful for describing the topic. Therefore, only the words considered in the discussion will be presented. The meaningful characteristics are listed in decreasing order of TF-IDF score:

- “Box Office”: “weekend, hunger, thanksgiving”
- “Lore”: “marvel, napoleon, hunger”
- “Production”: “hunger, napoleon, ridley”
- “Cast”: “hunger, ridley, director, star, napoleon,”

- “Writing”: “hunger, napoleon, marvels”



Engagement with Topics

In the following figure, each movie is represented with its own pie chart. This pie chart is divided according to how often an article about the movie is labeled with the topic.

Figure 1: Topic Engagement

Coverage

Figure 2: Distribution the articles

Here, the movie “Thanksgiving” is a clear outlier. There are two plausible reason for this, which will be outline in the discussion.

Discussions

Based on the results obtained from the analysis of the visibility and reception of the movie "Napoleon" compared to other movies released at the same time, several key findings emerged. The study focused on four primary aspects: the classification of articles into specific typologies, the characterization of topics within these articles, the level of engagement of each movie with these topics, and the coverage of each movie in the articles.

Typology

The typology developed for annotating the data provided a comprehensive framework for classifying articles into distinct categories. These categories included "Box Office," "Cast," "Lore," "Production," "Writing," and "Delivery,"

each representing different facets of the movies' visibility and reception.

Topic Characterization

The characterization of topics was based on the top 10 words with the highest TF-IDF scores across articles labeled with each respective topic. However, the discussion narrowed down the relevant words that best described each topic. Notably, "Box Office" was associated with terms like "week-end," "hunger," and "Thanksgiving," indicating a possible correlation between the movie's commercial performance and the Thanksgiving holiday. Similarly, "Lore," "Production," "Cast," and "Writing" were characterized by specific terms related to elements like historical context, technical aspects, cast members, and narrative aspects, respectively, further delineating the distinct areas of focus within the articles.

Across the various movies analyzed, "The Hunger Games" exhibited notably higher engagement with all identified topics compared to its contemporaries. The pie charts representing each movie demonstrate that "The Hunger Games" consistently garnered a larger share in every topic category, indicating a pervasive and robust presence across diverse aspects.

Regardless of the category—be it "Box Office," "Lore," "Production," "Cast," "Writing," or "Delivery"—"The Hunger Games" emerged as a prominent subject within articles, showcasing a comprehensive engagement with multiple facets. This heightened engagement suggests a significant level of interest and attention from media sources and audiences alike towards the movie.

In contrast, while other movies like "Napoleon" and others were present in these topics, their engagement levels were comparatively lower, signifying a relatively narrower scope of coverage or attention within these specific aspects.

Overall, the findings from this analysis offer a nuanced understanding of how "Napoleon" was perceived and covered compared to other contemporaneous movies. The discussion section will further explore these findings, considering potential implications, limitations of the study, and avenues for future research.

Group Member contributions

Danning Li

- Annotated 150 Articles
- Wrote Python script for Preprocessing and TF-IDF Calculations
- Wrote JSON to CSV conversion Python script
- Wrote Introduction and Method for Annotation for Report

Tahsin Kabir

- Brainstormed and collected the 600 articles
- Performed open coding and brainstormed relevant topics
- Wrote Python script to remove duplicates
- Wrote the Data and Method Sections

Arpan Saha

- Brainstormed and assisted during data collection
- Annotated $\frac{1}{3}$ of total articles collected
- Participated in writing and reviewing the report

Philippe Latour

- Brainstormed and assisted data processing
- Wrote Results section
- Annotated half of the articles
- Brainstormed the topics that were selected