COMP 370 – Final Project

Written by Max-Henri Chanut[1], Xiao Qi Li[2], Gabrielle Macinnes[3], Xinyi Wang[4]

[1]261036357
[2]261037638
[3]261036053
[4]261038442

Introduction

This research project delves into the media coverage of Taylor Swift within North American news outlets, aiming to provide insights into the sentiment and topics surrounding her portrayal. Leveraging a dataset of 500 articles sourced from NewsAPI.org, our analysis employs a three-fold approach.

First, we conducted an open coding on a subset of 200 articles to identify and categorize key topics. Each article was assigned to a single topic, resulting in a comprehensive understanding of the prevalent themes in the coverage. Subsequently, the entire dataset was manually annotated with a focus on 7 identified topics.

Then, we assessed the sentiment of the articles by conducting a coding process, categorizing them as positive, negative, or neutral. A careful and defensible interpretation framework was developed to ensure consistency in sentiment labelling.

Finally, to characterize these topics, we employed tf-idf scores, calculating the 10 words with the highest importance for each topic and sentiment based on their inverse document frequency across all 500 articles. This approach ensures a nuanced representation of the most significant terms associated with each topic and sentiment.

In our analysis, we found some interesting insights, such as the importance of her relationship with Travis Kelce, the influence of the fan incident in Rio de Janeiro, or the dominance of factual, neutral articles over opinion-based articles praising or criticizing Taylor Swift.

Our findings shed light on the nature of media coverage surrounding Taylor Swift, providing valuable insights for our client, a media company. This report contributes to a deeper understanding of the portrayal of Taylor Swift in North American media, enabling informed decision-making and strategic planning for our client.

Data

Our dataset contains exactly 500 English articles from North-American news sources. The raw JSON file contains all the default information included in the response of query to *newsapi.org* such as 'source', 'title', 'author', 'description', 'url' and more. Articles were queried if one or more of our selected keywords were found in the title or description of the article. The keywords we chose to search for are: 'Taylor Swift', 'Swifties', and 'Eras Tour'. When transferring data to the TSV file, we kept only the 'title' and 'description' fields.

Due to the 100 article per query limit set by NewsAPI, we queried day by day to maximize the total number of articles included after running our *newscover.collector* script. When the script runs, it sends multiples queries, going back one day at a time, and stops whenever the total number of articles fetched is greater or equal to 500. At each iteration of the loop, we filter out duplicate articles using the 'url' field. At the end, the first 500 articles are kept.

After running our *newscover.collector*, we printed out the unique sources from our list of articles. We then passed this list to ChatGPT-4 to identify which sources were not based in North America, and added them to an 'undesirables' list in our *newsapi.py* script to be filtered out from each API response. We would then run *newscover.collector* again. This was done repeatedly until all our sources were North-American.

This approach was chosen as an attempt to limit the introduction of bias. We purposely elected against curating a list of North-American news sources which we would include in our dataset . In fact, hand-picking which sources to sample from will inevitably bias the corpus against lesser known sources, something we wanted to avoid. By taking filtering out non North-American sources, we ensure that the news sources present have not been influenced by us.

On the topic of bias, it is important to mention that querying articles in English was only a practical decision that we made, not a necessary one. However, this decision introduces a bias against French and Spanish speaking regions in North-America such as Quebec and Mexico. We knowingly allowed this bias in our dataset to be able to reliably annotate the articles depending on their content.

Methods

Topic Identification

We started by conducting an open coding on 200 articles to develop a typology, before manually annotating the remaining articles. The open coding was conducted by first summarizing each article with one or two keywords and then making note of the keywords that appeared frequently. We then tried to find commonalities between the remaining articles that didn't fit into an established category. The open coding was done in four iterations until we found a typology that was well-suited to our data.

Preprocessing

- Shuffling of data: We noticed that, as our data scraping occurred on a daily basis, the articles were naturally sorted in chronological order. Consequently, articles with similar subjects tended to be grouped together since they were relevant around the same time. To ensure the development of a representative typology for the entire dataset, we shuffled the dataset before conducting the open coding on the first 200 articles.
- Removal of words: we removed common stop-words along with contextually irrelevant terms from our dataset. Stop-words are words that appear frequently in a language but don't add much meaning, such as ”the,” ”is,” and ”at.” By eliminating these words, we were able to concentrate our analysis on more significant terms, which helped us have a deeper understanding of how Taylor Swift is portrayed in the media. The list of stop-words we used comes from nltk.corpus, a python package for natural language tasks. Although TF-IDF usually takes care of stop-words by nullifying their 'idf' value, our article descriptions were often incomplete due to the maximum length cut off. This means that some stop-words could escape a zero 'idf' value and pollute our word freqency results, hence the removal of stop-words.
- Lemmatization: We used lemmatization. This process converts words into their base form. For instance, the word 'albums' gets converted into 'album'. This step helps us group different variations of a word together. This way, we could more accurately count how often each word appeared in our dataset, leading to a more precise analysis.

Calculation of TF-IDF

In our analysis, Term Frequency (TF) measures how often a word appears within a specific category. It is a count of the occurrences of each word in the documents belonging to that category. Inverse Document Frequency (IDF) is calculated by dividing the total number of articles in that category by the number of articles that include the specific word. We then apply the logarithm to this division to reduce the effect of high frequencies. The final TF-IDF score for each word we obtained is by multiplying its TF by its IDF. Normally the IDF term ensures that common words receive a low TF-IDF score and so the removal of stop words is not necessary. We decided to remove stop words anyways since we are only looking through the title and description, which is cut off, and so the TF-IDF score for some stop words may not be 0 even if that word appears in every article. The other advantage of removing stop words is that it reduces the computational complexity of the TF-IDF calculation.

Calculation of Percentage for Articles' Sentiment and Categories

- Percentage of each category
  - This is just a standard calculation, taking the count of each topic and diving by the total article count (500 in our case). Although simple, these percentages shows us a general view of the topic distribution in our dataset and provide some insights into Taylor Swift's recent news coverage trends. A similar calculation was done to gauge the general sentiment distribution.
- Distribution of sentiment within each category
  - This is calculated by taking the counts of each sentiment in a topic and dividing by the total count of that particular topic. This measure yields more specific insights into the sentiment trend for each topic.
- Distribution of sentiment across all topics
  - This is calculated by taking the counts of each sentiment in each topic and dividing by the total article count for that sentiment. This measure allows us to see which topics have high impact on the overall presence of each sentiment among all articles.

Results

Topics typology

The typology chosen for classifying articles into representative topics is the following :

1. Music Releases

- Articles related to Taylor Swift's music releases, including new albums, singles, collaborations, or any significant musical projects.
- Example: *”Drake shouts out Taylor Swift on surprise release”*.

2. Deceased Fan

- Articles related to incidents, stories, or discussions about the unfortunate occurrence of a fan passing away during Taylor Swift's concert in Rio de Janeiro.
- Example: *”Taylor Swift Mourns Fan Who Died Before Rio Concert”*.

3. Travis Kelce

- Articles involving or discussing Travis Kelce in relation to Taylor Swift, including personal connections, collaborations, or joint events.
- Example: *”Everything There Is to Know About Taylor Swift and Travis Kelce”*.

4. Musical Accomplishments

- Articles focusing on Taylor Swift’s broader musical achievements beyond specific releases, such as awards won or chart performances.
- Example: *”Taylor Swift Named Apple Music’s Artist of the Year”*.

5. Eras Tour

- Articles related to Taylor Swift’s Eras Tour, including announcements, updates, and reviews.
- Example: *”Taylor Swift Surprises Fans With 2 Songs Believed to Be About Harry Styles During ’Eras Tour’”*.

6. Global Influence

- Articles involving global influence and impact, including Swifties, T.S. Themed Parties, and other articles related to Taylor Swift’s international reach.
- Example: *”Why is Taylor Swift so popular?”*.

7. Miscellaneous Facts

- A category encompassing articles that present unique or unconventional facts and information about Taylor Swift.
- Example: *”Biden Confuses Taylor Swift With Britney Spears in Awkward Turkey Pardon Moment”*

Table 1: Top TF-IDF Words for Each Topic

Rank	Deceased Fan	Eras Tour	Global Influence
1	die	fan	dancing
2	heat	concert	star
3	friday	show	night
4	ana	brazil	swifties
5	tour	rio	era
6	era	night	tour
7	clara	de	episode
8	benevides	era	son
9	janeiro	janeiro	new
10	death	tour	pop

Rank	Miscellaneous Facts	Music Releases
1	rumor	drake
2	biden	new
3	britney	version
4	spears	hour
5	turkey	song
6	deadpool	reveal
7	film	red
8	ryan	button
9	renaissance	year
10	concert	release

Rank	Musical Accomplishments	Travis Kelce
1	year	chief
2	album	kansas
3	version	city
4	chart	fan
5	artist	end
6	grammy	tight
7	billboard	relationship
8	music	star
9	hot	romance
10	week	jason

Table 2: Topic Distribution

Topic	Count	Percentage (%)
Travis Kelce	150	30
Eras Tour	103	21
Deceased Fan	80	16
Musical Accomplishments	59	12
Global influence	53	11
Miscellaneous Facts	37	7
Music Releases	18	3

Sentiment typology

The typology chosen for classifying articles into positive, neutral, and negative sentiments was based on the need for a nuanced understanding of the media coverage of Taylor Swift. Each category serves a specific purpose and reflects different aspects of Taylor Swift’s portrayal in the media :

1. Positive

- Praise toward Taylor Swift.
- The title and description should contain some form of compliment or appreciation.
- Examples: Music on top of charts, award wins, tour success, etc.

2. Neutral

- Factual reports about Taylor Swift.
- The title and description should contain only general information without the presence of subjective comments.
- Examples: Tour announcements, music releases, personal life updates, etc.

3. Negative

- Criticism toward Taylor Swift.
- The title and description should contain some form of controversy or unflattering language.
- Examples: Incident during a performance, legal disputes, personal life drama, etc.

Table 3: Top TF-IDF Words for Each Sentiment

Rank	Positive	Neutral	Negative
1	year	travis	death
2	song	kelce	died
3	tour	tour	heat
4	album	era	tour
5	new	fan	era
6	fan	chief	janeiro
7	era	city	de
8	week	star	die
9	music	kansa	show
10	artist	concert	friday

Table 4: Sentiment Distributions

Topic	Count	Percentage (%)
Positive	123	24
Neutral	286	58
Negative	91	18

Table 5: Within Topic Sentiment Distributions

	Within topic percentages (%)		
Topics	**Positive**	**Neutral**	**Negative**
Deceased Fan	6	10	**84**
Eras Tour	22	**68**	10
Global influence	28	**70**	2
Music Releases	**50**	**50**	0
Musical Accomplishments	**85**	12	3
Travis Kelce	11	**84**	5
Miscellaneous Facts	11	**84**	5

Table 6: Percentage of Sentiment for Each Topic

	Percentage of Sentiment (%)		
Topics	**Positive**	**Neutral**	**Negative**
Deceased Fan	4	3	**74**
Eras Tour	19	24	12
Global influence	12	13	1
Music Releases	7	3	0
Musical Accomplishments	**41**	2	2
Travis Kelce	13	**44**	9
Miscellaneous Facts	4	11	2

Discussion

We have ensured that our typology was precise and comprehensive. This was particularly needed in this project as some articles we collected were ambiguous as to their coverage on Taylor Swift. Numerous articles referenced Taylor Swift but were not explicitly about her or covered events with a Taylor Swift theme, such as 'Taylor Swift Night' on Dancing With The Stars. Articles of this nature were categorized as 'Global Influence,' capturing their connection to the impact her music and career have had across various domains.

Common causes of confusion among categories.

- 'Eras Tour' vs. 'Music Releases': Some articles talked about Taylor Swift performing her new music at the Eras tour. This group of articles was put into the 'Eras Tour' category since they talked about Taylor Swift's performance rather than music releases.
- 'Eras Tour' vs. 'Travis Kelce': There were multiple articles about Travis Kelce visiting Taylor Swift while on tour. We had to use our judgement to determine whether each article was centered more on Taylor Swift and Travis Kelce's relationship or the Eras tour itself. For example, articles which focused on Taylor Swift running up to kiss Kelce after the show were put in the 'Travis Kelce' category and articles in which her performance was the focal point were categorized as 'Eras Tour'.
- 'Eras Tour' vs. 'Deceased Fan': Since the incident where one of Taylor Swift' fans passed away happened at one of the Eras tour shows, the two categories could have been combined. We chose to make 'Deceased Fan' its own category since the incident was significant enough to account for 16% of our dataset.
- 'Global Influence' vs. 'Miscellaneous Facts': Both categories contain articles where Taylor Swift is not the main focus. We make the distinction based on whether it relates to her impact on society or just mentions her name in passing.

The 'Eras Tour' category had the most overlap with other categories since it is a relatively broad topic. We chose to keep these more specific categories such as 'Travis Kelce' and 'Deceased Fan' since they make our typology more expressive and provide an accurate representation of how Taylor Swift is covered in the news.

For the media company client, this typology provides actionable insights. Understanding the balance between positive and negative coverage, as well as the volume of neutral reporting, aids in strategic decision-making for public relations, marketing, and overall media management.

Most frequent words analysis

From Table 1 above, we can see that the category 'Eras Tour' focuses on her concert experiences, discussing notable achievements and challenging events. This topic is the second most important based on the number of articles classified in this category (Table 2). This can be explained by the fact that the Eras Tour is still ongoing and thus updates, announcements and events are immediately covered by the media. The top TF-IDF words of this category show that the traumatic incident of the deceased fan in the Rio de Janeiro concert deeply influenced the articles written about the Eras tour recently. We also note that it is the category with the most even distribution between the three sentiments, indicating a wide range of subjects included in it.

The top TF-IDF words for the category ’Deceased Fan’ translate the surrounding environment of this traumatic incident, such as the day on which it happened (’friday’), the location in which it happened (’janeiro’) and the reason due to which it happened (’heat’).

For ’Global Influence’ articles, it is immediately apparent that the two main topics discussed are ’Taylor Swift Night’ on Dancing With The Stars (’dancing’, ’star’, ’night’, ’episode’) and her devoted fans (’swifties’).

In the ’Miscellaneous Facts’ articles, coverage was focused on rumours of a Taylor Swift appearance in the 3rd Deadpool movie (’rumour’, ’deadpool’, ’ryan’, ’film’) and the mishap involving American President Joe Biden mistaking Taylor Swift for Britney Spears at the annual turkey pardoning ceremony (’biden’, ’spears’, ’turkey’).

’Music Releases’ and ’Musical Accomplishments’ concentrate on her career in music, including new songs, collaborations, and success. We notice that ’drake’ was the top TF-IDF word in the ’Music Releases’ category, referencing how Drake praised Taylor Swift in his new song, ’Red Button’. However, only a few articles were classified as ’Music Releases’ as Taylor Swift has not recently released any new songs.

Lastly, the ’Travis Kelce’ category covers her personal life with the Football player, including his football team and the involvement of their fans in their relationship (’chief’, ’kansas’, ’fan’, ’relationship’). We notice that this is the topic with the highest number of articles, highlighting it’s importance and influence in the coverage of Taylor Swift in media.

Overall, this analysis allows us to clearly see the topics of the articles collected and how they relate to their classification. This is extremely useful to our client as they can immediately determine what the articles in each section focus on based on our classification and analysis.

Sentiment Analysis

- Positive: This sentiment is characterized by Taylor Swift’s success in the music industry. It tends to include articles about her new songs or albums and the awards she has received, which is why we see that articles within ’Music Releases’ and ’Music Accomplishments’ categories are largely positive. The frequent appearance of music-related terms in Table 1, which have high TF-IDF scores, supports this focus.
- Neutral: We see in Table 4 that the majority of articles are classified as ’Neutral’. This shows that media often describes events and news about Taylor Swift in a informative manner. We relate this to the importance of Taylor Swift’s fan community which warrants a large demand of news, updates and facts about their idol. In fact, we see in Table 5 and 6 that her relationship with Travis Kelce is often covered in a neutral way. We observe the same trend in ’Miscellaneous Facts’, which we had expected since facts are often neutral and informative.
- Negative: This sentiment centers on a tragic incident at one of her concerts where a fan passed away (74% of articles classified as ’Negative’ came from the ’Deceased Fan’ category). We made the choice of categorizing reports about incidents during a performance as negative, unless there was explicit complimentary comments about Taylor Swift’s reaction or response. We justify this choice by arguing that an article reporting such an incident generally has a negative impact on the artist’s image, as it portrays a bad concert-going experience. The most prominent terms in this category, highlighted by their high TF-IDF scores, reveal details related to the concert, such as the location and the circumstances of the incident.

Limitations

In our first attempt at data collection, we only searched for articles which had ’Taylor Swift’ in the title. With this method we were unable to collect 500 unique articles since NewsAPI imposes certain limits on queries unless we paid to upgrade our plan. To rectify this issue, we tried supplementing our dataset with articles from GNews, an API for articles on Google News. The articles from GNews did not have a description which was distinct from the article title so we found them very difficult to annotate. Because of this, we decided to try NewsAPI again with a different query that would return more articles. Instead of just searching for ’Taylor Swift’, we also searched for ’Swifties’, and ’Eras Tour’. We also changed the ’searchIn’ parameter to search for keywords in the title and description instead of just the title. We were able to get all 500 articles this way, however if we had been able to just search for ’Taylor Swift’ our data would have likely looked different. For example, we would have had fewer articles in the ’Eras Tour’ category.

Moreover, NewsAPI only allows users with the base plan to query articles from within the past 30 days, and as a result, our dataset may not be the best representation of how Taylor Swift is typically covered in the media. For example, 16% of our articles were about the Taylor Swift fan who passed away at one of her concerts. If we were to collect news data now, we would most likely not see any articles falling under the ’Deceased Fan’ category. Instead, we would find more coverage of recent events that are not in our dataset. This particular example also affects our sentiment analysis since the majority of the articles we labelled as ’Negative’ were in this category, so if we had gotten our data from a different date range, we would likely see Taylor Swift being reported in much more positive fashion.

Moreover, we only included articles written in English, a choice which introduces bias against North-American regions where English is not the primary language, such as Quebec and Mexico. This decision was made to allow for easier annotation, but definitely not a necessary one. If a similar study were to be conducted again, one could include North-American articles of any language and use a reliable translator tool to annotate articles they are unable to read fluently. This method would take more time than considering only English articles, but would yield more representative and less biased results.

Lastly, when conducting the open coding as well as the annotation, we had to categorize articles based on only the title and description of the article as we would not have been able to read all 500 articles. As such, we had to infer a topic and sentiment in some cases. The description was also often cut off and did not provide much more information than was already in the title.

Group Members Contributions

- Max-Henri[1]: Typology definition, annotation, most frequent word and sentiment analysis, introduction.
- Xiao Qi[2]: Wrote data collection script, developed the sentiment definitions, computed topic-sentiment percentages, conducted the sentiment coding, contributed to data and discussion sections.
- Gabrielle[3]: Collected data, conducted the open coding, contributed to the method and discussion sections.
- Xinyi[4]: Conducted TF-IDF coding, contributed to the result, method and discussion sections.