Fall 2023 – COMP 370 Project: Taylor Swift Media Coverage

Sissy Chen, Chelsea Chisholm, Donald Szeto, Yu Hao Tian

Introduction

From breaking concert sales records to influencing economies to dictating local traffic patterns in Philadelphia, Taylor Swift is at the forefront of many headlines. Such significant fame inevitably draws an immense amount of attention and scrutiny from the media. This study aims to analyze and categorize Taylor Swift's recent North American media coverage into specific topics based on public interest. Using the available data, a consensus is determined based on the overall positive or negative sentiment of Taylor Swift's recent media coverage, within the last 4 months, by reputable and popular news media across 8 categories.

More formally, this paper seeks to determine the top eight topics related to articles dating from August to November 2023 related to TS scraped from MediaStack API with the country filter set to Canada, US and language filter set to English and positive, neutral, and negative sentiment related to those articles by percentage.

This was investigated with a combination of computation and manual methods. For the top eight topics in the media about Taylor Swift, open coding was used to create typology, and a Naïve Bayes model was considered for automating the process of implementing said typology. For each topic, a TF-IDF score determined the top 10 words in each category and sentiment combination. Lastly, the general sentiment was determined using both manual and computational (SentiWordNet Lexicon model) annotation methods.

Results determined that across 500 sampled articles, the media coverage of Taylor Swift is dominantly positive (60.2% according to manual annotation, 51.8% according to computational annotation), although the dominant sentiment across topics varies. This is regardless of the annotation method.

This article investigates the prevalent topics related to Taylor Swift and the general sentiment surrounding these news articles.

Data

Budgetary constraints and the time-restrictive access (within the past month) of the free developer tier of NewsAPI prompted the search for an alternative news source that could provide a sufficient number of articles to analyze. A viable alternative was found in the MediaStack API, wherein the articles obtained were written no more than 4 months before the date of the request, the latter of which was confirmed via testing.

Parameters given to the query were relatively open with only the keywords "Taylor Swift", language (EN), and countries of origin (US/CA) specified. Spanish media publications were explicitly removed due to the unavailability of fluent Spanish speakers on the project team. This yielded raw data in the form of approximately 7,900 articles dated August 2023 to November 2023.

The collected articles were then filtered according to the mention of "Taylor Swift" in the article title as an agreed-upon convention that the article is then about her and for the removal of duplicates. The latter of which was defined as any article with a URL that is an exact match to any of the other articles' URLs. Filtering resulted in approximately 3000 valid articles.

Duplicates were defined as URL duplicates and not as identical articles published by different news outlets. This was done to ensure a representative sample of the true distribution of total articles. Removing title/content duplicates could compromise the proportionality of the sample and negatively impact the accuracy of the top eight topics and sentiment annotation and therefore the results.

Finally, the selection policy used to sample the dataset of 500 articles prioritizes articles whose sources belong to reputable and otherwise popular news outlets. Popular sites were determined from sources such as Top10 and Forbes. Articles from this list were automatically selected, and the remainder were randomly selected.

The selection policy prioritizes articles published by popular sources to produce a representative sample. If an article has multiple sources of readership, then the article is likely to be more influential. These decisions were to eliminate the common systematic sampling biases.

Methods

The manual annotation process consisted of 2 phases: devising the labels to use and assigning the labels. The first phase was executed during a group session, where each group member contributed to annotating the first 200 articles through 3 open coding attempts for each article. The typology was finalized at the end of the first phase, along with an annotation guide. The second annotation phase consisted of assigning a label to each of the remaining 300 articles. This phase was performed by 2 expert annotators based on the annotation guide devised by the group.

The resulting typology consists of 8 topics: Achievements, Dating, Discography, Filmography, Gossip, Social Impact, Tour, and Unrelated. Articles were labeled with each topic based on the following talking points in its content:

Achievements: Record-breaking, historical firsts, top rankings, awards

Dating: Everything related to Taylor Swift's current partner regarding their romantic relationship

Discography: Release of songs and albums, re-recordings, music track information

Filmography: The Eras Tour movie

Gossip: Speculations by the media and/or the public regarding Taylor Swift's life and friendships, excluding her romantic relationship

Social Impact: Taylor Swift's impact on her fanbase, on business sales, and on the general social landscape

Tour: The Eras Tour

Unrelated: Not revolving around Taylor Swift. (e.g., clickbait, ads that use Taylor Swift's name but are not about her actions, thoughts, or opinions)

Articles labeled as Unrelated are often about another celebrity, such as another music artist, sports player or team, or an individual in the film or journalism industry. In most of these articles, Taylor Swift is mentioned as an example connected to the subject matter discussed.

All 500 articles were also manually annotated once for sentiment analysis, with a value of 1 for positive sentiment, 0 for neutral, and –1 for negative. Once all 500 articles were labeled with a topic and with a sentiment integer value, the number of articles belonging to each topic was computed. Then, the word count was computed for each word appearing more than 5 times in total in a given document. A document was defined as the title, description, and contents of all articles with the same topic and sentiment label. Therefore, articles belonging to the same topic produce between 1 and 3 documents, depending on the total number of sentiment labels used to label these articles. Since at least one article on each topic was labeled to be Positive, Neutral, or Negative, the 500 articles categorized into 8 topics produced 24 documents. This step labels each word with a frequency count, a sentiment label, and a topic.

Finally, the TF-IDF score for each word was computed by multiplying the term frequency of the word by its inverse document frequency. While the term frequency is equivalent to the word count, the inverse document frequency is the log of the quotient resulting from dividing the total number of documents for the topic the word is associated with (between 1 and 3 inclusively) by the number of documents the word appears in (1 to 3 inclusively). Note that the total number of documents for the topic the word is associated with is always larger or equal to the number of documents the word appears in. The higher the TF-IDF score for a word is, the more likely an article with the same topic and containing the same word is of the same sentiment, as it means that the word is not used as frequently in articles with the same topic but of a different sentiment.

The computational approach for sentiment annotation used a similar method as in Caltech University's 'Sentiment Analysis of News Articles: A Lexicon-based Approach' written by Antony Samuels and John Mcgonical from Cornell's scholarly article online archive, arXiv [Samuels and Mcgonical, 2007]. This method was chosen since it works specifically for English news articles and does not produce highly neutral results. It is therefore a method that is good at determining polarity in connotation.

The applied method involves the tokenization, cleaning, lemmatization, and application of parts of speech tags to all words in each article. Next, smooth TF-IDF was computed for each of these words and selected the top ten for each article. Smooth TF-IDF was chosen for better information retrieval and subsequent better sentiment scoring since it prevents division by zero errors.

NLTK's SentiWordNet library was used to score the sentiment of a word. This library utilizes Princeton's English lexical database, WordNet. The library scores each word on a continuous sentiment range of [-1,1], where –1 is negative, 0 is neutral, and 1 is positive. To determine the sentiment of a whole article, the top 10 words as determined by the TF-IDF scores are multiplied by their respective sentiment score and then summed. Using only the top ten words and multiplying their TF-IDF score by their sentiment score was not a decision informed by the academic paper. However, this step is necessary since each word should contribute according to their importance in finding the sentiment score of the whole article. In fact, when multiplication was skipped, the results were overwhelmingly neutral since common words that score neutrally would be given the same weight as words with high TF-IDF scores when computing the sentiment of the article.

Therefore, an article is determined to be Positive, Neutral, or Negative based on their total sentiment sum in the range of $\mathbb{R}^+ \setminus \{0\}$, 0, and $\mathbb{R}^- \setminus \{0\}$ respectively.

Results

For manual annotation of the topics using the open-coding developed typology, of the 500 articles, 19.4% (97) were labeled as Dating, 16.6% (83) were labeled as Tour, 13.4%, (67) were labeled as Unrelated, 12.4% (62) were labeled as Social Impact, 12.0% (60) were labeled as Achievements, 11.6% (58) were labeled as Filmography, 7.6% (38) were labeled as Gossip, and 7.0% (35) were labeled as Discography.

Introducing a Naïve Bayes model was originally an idea posited that would help expedite the labeling process after the typology had been decided. The main draw was that after the manual labeling of a few hundred training examples in the dataset was done, it would be possible to train a machine learning model with a (hopefully) good enough test accuracy so that the rest of the process could be done automatically, and hopefully with only little human supervision needed as to the output results. To benefit the most from this method and for the training to have been worthwhile, initial expectations of test accuracy were projected to be around the 80% mark.

In actuality, the most tuned Naïve Bayes models (using multinomial likelihood and Lidstone smoothing, as well as stop words removal and lemmatization) only yielded 70.9% test accuracy on average. Those were the results obtained after accounting for the bias in the data sampling (where we opted for a uniform distribution over class categories on the cardinality of the training set – that is, the training set was specifically engineered so that articles of rarer classes were accounted for). Note: Although the latter methodology does have inherent bias, i.e., the training data not being representative anymore of the sample – since sampling was done randomly, it resulted in a much more stable model.

While Naive Bayes models are typically quite effective with smaller datasets, their feature independence assumption for posterior likelihood calculations becomes a liability in high-dimensional spaces, which is only exacerbated by the curse of dimensionality. That is to say: Since we couldn't get as good an accuracy as initially expected, and with the fact that better accuracy for Naïve Bayes only gets harder with the more training data there is, this might just point to the limitations it has training over complex texts.

Taking these limitations in the usefulness of our model into account, as well the various time-consuming defects in the data collected (noise in the training corpus that would have needed manual attention for every single training example) that were not fixed for time-constraints, it was ultimately decided that the best course of action was to leave this part of the project on hold, to maybe revisit the idea one day, implementing more powerful models for the task (pre-trained GPT, or BERT models for instance). The rest of the annotations would be done manually henceforth.

For manual annotation of sentiment analysis out of 500 articles, 60.2% (301) were labeled as positive, 28.4% (142) were labeled as neutral, and 11.4% (57) were labeled as negative.

This was then compared to the computational annotation using the SentiWordNet lexical approach. Using the computational model, of the 500 articles, 25% (125) were Negative, 23.2% (116) were Neutral, and 51.8% (259) were Positive. When comparing the manual annotation to the computational annotation, the Negative category increased by 13.6%, whereas Neutral and Positive decreased by 5.2% and 8.4% respectively.

Next, the percentages of Positive, Neutral, and Negative tagged articles for each Topic were compared for both the manual and computational annotation methods.

Topic	Total Number of Articles	Positive	Neutral	Negative
Dating	97	57.73%	25.77%	16.49%
Tour	83	51.81%	25.30%	22.89%
Unrelated	67	34.33%	43.28%	7.46%
Social Impact	62	88.71%	8.06%	3.23%
Achievements	60	83.33%	15.00%	1.67%
Filmography	58	74.14%	20.69%	5.17%
Gossip	38	28.95%	44.74%	26.32%
Discography	35	57.14%	40.00%	2.86%

Figure 1: Manual Annotation Percentages of Positive, Neutral, and Negative Articles by Topic

Topic	Total Number of Articles	Positive	Neutral	Negative
Dating	97	50.52%	19.59%	29.90%
Tour	83	45.78%	16.87%	37.35%
Unrelated	67	56.72%	25.37%	17.91%
Social Impact	62	33.87%	48.39%	17.74%
Achievements	60	68.33%	18.33%	13.33%
Filmography	58	43.10%	29.31%	27.59%
Gossip	38	57.89%	13.16%	28.95%
Discography	35	71.43%	8.57%	20.00%

Figure 2: Computational Annotation Percentages of Positive, Neutral, and Negative Articles by Topic

Most topics follow the same trends as in the manual method, with the notable exceptions of Social Impact, Gossip, and Unrelated. Gossip and Unrelated as annotated by SentiWordNet are more polar, whereas Social Impact is significantly more neutral.

Lastly, of all the articles only 42.6% (213 out of 500) were tagged in the same category by both methods. Furthermore, the percentages of articles that were tagged in the same sentiment category by both annotation methods. 158 mutually agree on Positive. 20 mutually agree on Neutral. 35 mutually agree on Negative.

Notably, both methods identified most articles as Positive. Further, the sentiment category with the greatest number of sentiment article annotation agreement is Positive.

Discussion

In retrospect, the data cleaning and filtering steps may have benefitted from clarifying the meaning of "North American media coverage". One refinement would have been to select popular news websites according to some quantifiable definition of popularity such as total readership or site visits. The original popularity rankings used in this study were strictly qualitative and might have been more prone to the personal biases of the ranking's author.

A feasibility attempt was conducted with one such ranking found on the internet. This ranking (PressGazette.co.uk) contained the top 50 biggest news websites according to US site visits within the past month (October 2023). Examples of news websites found therein include all the major North American outlets (e.g. CNN, NYTimes, Fox News, etc.), but also many international websites such as the Daily Mail, Al Jazeera, and the Times of Israel. While not necessarily North American media, their relevance to US readership still would've been of particular interest to the study at hand as it examines media coverage from the perspective of the (US) reader instead.

This change to our approach may have yielded a more faithful representation of "North American media" by aligning the batch of articles to be analyzed with what media is actually consumed by Americans. The question that we would then have answered would be specific to US coverage.

For articles categorized into Dating, the words "viewership" and "serious" turn out to be informative of positive sentiment. This suggests that an article about Taylor Swift's dating life containing either the word "serious" or the word "viewership" is more likely to speak about her positively than with a neutral or negative tone. This makes sense as a relationship becoming more serious is often an indication that things are going well in the dating aspect. Meanwhile, the word "billionaire" seems to infer negative sentiment. This suggests that an article about Taylor Swift's dating life containing this word is more likely to speak about her negatively than with a neutral or positive tone. This also suggests that relating her dating life to her net worth or her partner's net worth is more likely to be brought up with negative connotation than positive.

For articles categorized into Tour, the terms "heat", "Rio de Janeiro", and "Kounalakis" turn out to be informative of negative sentiment, while words like "Canada", "reopened", and "closure" turn out to be informative of neutral sentiment. This makes sense as a Taylor Swift fan recently passed away during one of her concerts in Rio de Janeiro due to a heat wave and suggests that an article about Taylor Swift's 2023 tour containing either the word "heat" or the term "Rio de Janeiro" is likely to be negative rather than neutral or positive. Furthermore, words like "Canada", "reopened", and "closure" seem to infer neutral sentiment makes sense as they were related to articles reporting facts about the Canadian leg of Taylor Swift's 2023 international tour. The words "tour" and "tickets" seem to infer positive sentiment. This suggests that an article containing either or a combination of those words is more likely to speak about Taylor Swift's 2023 tour positively rather than neutrally or negatively. This makes sense as the ticket sales of her tour have been breaking sales records.

For articles categorized into Social Impact, the words "cultural" and "younger" turn out to be informative of negative sentiment. This suggests that an article containing either or a combination of those words is likely to be more negative than neutral or positive. This also suggests that a certain part of the population does not perceive Taylor Swift's cultural impact, as well as her impact on younger fans, to be the most positive.

For articles categorized into Achievements, the words "year", "copies, "sold", "nominated", and "feat" turn out to be informative of positive sentiment. This makes sense as Taylor Swift has been receiving numerous accolades honoring her work ethic and record sales. The word "predictions" seems to infer neutral sentiment. This suggests that an article with the same topic and containing this word is more likely to have a neutral tone than veering toward positive or negative, which makes sense as predictions are yet to be confirmed and are often based on statistics, which are factual in nature and therefore exhibiting neutral sentiment.

For articles categorized into Filmography, the words "release" and "premiere" turn out to be informative of positive sentiment. This makes sense as Taylor Swift's The Eras Tour movie was immensely successful following its release and the movie premiere garnered a lot of positive attention from the public. Therefore, an article discussing The Eras Tour movie and containing either or a combination of the words "release" and "premiere" is likely to speak about Taylor Swift positively rather than neutrally or negatively. However, the word "cultural" seems to indicate negative sentiment, similarly to its sentiment association for articles categorized into Social Impact. This suggests that some people may criticize the movie regarding its cultural aspects.

For articles categorized into Gossip, the words "bff" and "friendship" turn out to be informative of positive sentiment, while names like "Olivia Rodrigo", "Kayla", and "Matty" seem to infer negative sentiment. On one hand, it suggests that an article containing either "bff" or "friendship" is more likely to speak of Taylor Swift positively, indicating that her friendships are viewed positively as well. On the other hand, an article mentioning Olivia Rodrigo, Kayla Nicole, or Matty Healy in relation to Taylor Swift is likely to have a more negative tone. This makes sense as there was a rumored feud between Taylor and Olivia a few years ago, Kayla Nicole is the ex-girlfriend of Taylor's current boyfriend, and Matty Healy is Taylor Swift's ex-boyfriend.

For articles categorized into Discography, the words "streaming", "spotify", "hot", "chart", "top", and "single" turn out to be informative of positive sentiment. This suggests that Taylor has encountered great success when it comes to her music career. Meanwhile, the word "labels" seems to infer negative sentiment. This makes sense as Taylor Swift was in dispute with her former record label, leading her to re-record her songs. Therefore, an article discussing her discography containing the word "labels" is likely to have a more negative tone rather than a neutral or positive tone.

A difficulty in using TF-IDF to define each category and finding context within these top words is that TF-IDF gives high scores to words that are frequent in a specific article, but uncommon in all other articles. The issue is that most of the articles mentioned multiple topics in some form. This makes the expected significant words of each topic less significant since they are mentioned throughout most of the articles. Therefore, since the human-based approach can identify nuances in typology if the typology overlaps in vocabulary for different categories, the TF-IDF is affected.

To evaluate the quality of the manual sentiment annotation and provide other annotation references, a computational method, which makes use of SentiWordNet, was used as a comparison. There are some significant discrepancies between the two methods in the results, particularly when defining neutrality and negativity.

One notable difference is how the polarity of sentiment increases in the computational method. For example, the Negative category is more than doubled when done manually. This increase in polarity is congruent with the SentiWordNet research paper, which states that Neutral is generally the smallest category as the model is good at distinguishing polarity in sentiment.

Such differences are likely caused by the difference in contextual evaluation between the two methods: human annotation takes the entire article content into context, whereas SentiWordNet sums significant words to determine sentiment. SentiWordNet is a lexical, dictionary approach. In the SentiWordNet method, the article's sentiment depends on the sentiment of the most significant words without context.

For example, a sentence such as "Taylor Swift fans were very satisfied with their gift baskets" would most likely be assigned a positive sentiment score by SentiWordNet. This is because the words "fans", "satisfied", and "gifts" have by default a positive connotation. However, according to the typology developed for manual annotation, this sentence would be neutral due to its factual reporting style. The same would occur for Negative annotation.

Furthermore, this method cannot account for slang, metaphorical language, or sarcasm, especially not in cases where the context is fully understood only after reading multiple words.

Another difference can be observed when comparing the annotation percentages of positive, neutral, and negative articles by topic produced by both methods. The increase in positive annotation for Gossip from the manual method to the computational annotation is likely due to the reason mentioned previously. Most articles categorized as Gossip revolved around Taylor Swift's outings and dinners with friends. The vocabulary is generally positively defined. This could account for the large increase in positive Gossip articles while the percentage of negative articles barely increased.

Unrelated sees an increase in both negativity and positivity. This is because the Unrelated topic is mainly clickbait or ads. Therefore, the variety of words with highly polar definitions is likely to increase since clickbait and ads use flashy vocabulary.

Social Impact becomes more neutral because the words in those articles are neutrally defined; Social Impact articles centered on Taylor Swift's effect on the economy, and changes in fan behavior, such as causing traffic jams, are neutral. The diverse yet neutral language is most likely what caused Social Impact to be neutrally defined in the computational method since it is based on word definitions.

This comparison demonstrates that the manual annotation using human open-coded typology is better suited for sentiment article analysis involving Taylor Swift. The articles reference popular culture and are meant to attract readers due to her celebrity status. Therefore, the writing style is often informal, which makes it difficult for a lexical dictionary approach to contextualize the text.

Most significantly, both approaches strongly agree that the sentiment of most articles is positive. Moreover, the Positive category has the most mutual agreement for both methods. Assuming our sample is representative, the results strongly indicate that news coverage of Taylor Swift according to the project parameters is mostly positive in sentiment.

Group Member Contributions

- Sissy Chen (260601294): Open coding group session, main annotator, annotation guide, TF-IDF implementation and analysis.
- Chelsea Chisholm (260799832): Open coding group session, data collection scraping, filtering, and sampling, computational sentiment annotation implementation and analysis.
- Donald Szeto (260403046): Open coding group session, MediaStack API data collection, and JSON raw article data formatting.
- Yu Hao Tian (260379200): Open coding group session, and Naïve Baye's implementation and analysis.

References

FeedSpot. 2018. Best 40 Celebrity Gossip Magazines & Publications. https://magazines.feedspot.com/celebrity_gossip_magazines/. Accessed: 2023-12-11.

Glader, P. 2017. 10 Journalism Brands Where You Find Real Facts Rather Than Alternative Facts. https://www.forbes.com/sites/berlinschoolofcreativeleadership/2017/02/01/10-journalism-brands-where-you-will-find-real-facts-rather-than-alternative-facts/?sh=1f127c78e9b5. Accessed: 2023-12-11.

MediaStack. 2022. Free, Simple REST API for Live News & Blog Articles. https://mediastack.com/. Accessed: 2023-12-11.

Press Gazette. November 17, 2023. Top 50 news websites in the US. https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/most-popular-websites-news-us-monthly-3/. Accessed: 2023-12-11.

Samuels, A.; Mcgonical, J. 2007. News Sentiment Analysis. University of Southern California/Caltech. https://arxiv.org/ftp/arxiv/papers/2007/2007.02238.pdf. Accessed: 2023-12-11.