Twitter Sentiment Analysis: Comparing Classification Tasks to Predict Netflix's Trending US Movies

Patricia Sarmiento Gonzalez

pat.sarmient@gmail.com

Northwestern University, Evanston, IL

Abstract

Can tweets predict the popularity of Netflix movies? This study explores predictions of movie popularity by using the sentiment analysis model VADER (Valence Aware Dictionary for Sentiment Reasoning) to assign sentiment scores to tweets referencing movies streaming on Netflix. It then compares the performance of machine learning classification models Logistic Regression and Bernoulli Naïve Bayes in predicting which movies are ranked top ten at least once in Netflix's weekly top ten streamed movies list. Results show that Bernoulli Naïve Bayes slightly outperforms Logistic Regression with F1-scores of 53% and 49%, respectively, indicating that further work is required to optimize the model.

Keywords:

Twitter Sentiment Analysis, Pre-Trained Language Model, VADER, Machine Learning, Binary Classification, Logistic Regression, Bernoulli Naive Bayes

1 Introduction

Netflix took the 2021 award shows by storm with seven Oscar, and forty-four Emmy wins, leading all studios in both events, showcasing its media production competency, particularly in the television series space (Spangler 2021; Baysinger 2021). However, a recent report listing Netflix's historically top ten movies and television series proves that movies, totaling 818 million views, are still a strong contender compared to television series, totaling 761 million views (Tapp 2021). Given movies' high licensing and original content production costs, identifying early positive responses can help Netflix gauge future movie preferences.

This study predicts the popularity of Netflix movies by leveraging Twitter's sentiment analysis capability using the unsupervised lexicon-based model VADER (Valence Aware Dictionary for Sentiment Reasoning) to assign sentiment scores to tweets that mention Netflix movies streamed between January and March of 2022. The mean sentiment score per movie, total tweets per movie, and Netflix's weekly published top movies list helps train supervised binary classification models Bernoulli Naïve Bayes and Logistic Regression to predict which movies are trending or not.

Twitter's condensed data structure, allowing users a maximum of 140 characters per post or tweet, is a practical fit for sentiment analysis which relies on compact word content for its quantitative scoring method (Philander and Zhong 2016, 17). Furthermore, users on Twitter tend to post in real-time, increasing the strength of sentiments expressed beyond a binary polarity (Hutto and Gilbert 2014, 218). VADER can evaluate the sentiment intensity of words, attributing a higher positivity score to the post "amazing movie" than "good movie".

The comparative analysis of machine learning classifiers Bernoulli Naïve Bayes and Logistic Regression suits the probabilistic classification task at hand given the binary nature of the output (top ranking/1, not top ranking/0) and the assumption of independence between observations, since "tweeted" opinions about movies are personal; all of which are feature characteristics of these models (Kumar and Babcock 2017).

2 Literature Review

Sentiment analysis is the field of study that analyzes people's sentiments towards other entities by assigning positive or negative dimensions (Mohammad 2016, 6) through computational methods that evaluate subjectivity in a text (Gilbert 2014, 217). The literature of interest in this review focuses on developing and using pre-trained language models and their real-world applications.

Early studies demonstrated the importance of designing lexicons to train language models. Hatzivassiloglou and McKeown (1997) used news articles from a 1987 Wall Street Journal to build a list of positive and negative adjectives to predict whether they were synonyms or antonyms using a log-linear regression model: delineating binary orientation (positive-negative). To account for the degree of sentiment, Turney and Littman (2003) inferred semantic orientation from semantic association by calculating the strength of a word's association with a set of positive words such as [good, nice, excellent, positive, fortunate, correct, superior] minus the strength of its association with its corresponding opposing pairs [bad, nasty, poor, negative, fortunate, wrong, inferior] (6).

Once text mining tools automatically collected and analyzed large corpora, improved lexicons applied extended language rules to social media. Zhang et al. (2011) applied a Pearson's

chi-square test to a corpus of tweets to identify and add opinion indicators to a lexicon-based model by calculating how dependent a term such as cute was on the positive or negative tweets surrounding it. A Support Vector Machine (SVM) classifier was applied to assign sentiment polarity to the newly identified opinionated tweets (7).

Real-world applications of pre-trained models emerged following a period of extensive model fine tunings. Philander and Zhong (2016) used a pre-trained sentiment lexicon to assign sentiment scores to tweets mentioning specific resorts in Las Vegas. By assigning an average score to each resort, the researchers could visualize public opinion over time and compare sentiment scores between resorts (19). Guda, Srivastava, and Karkanis (2022) predicted yelp scores based on restaurant reviews by comparing six machine learning models to different model setups, including using meta-features (good/bad or 1-5 range star classification), text reviews represented as TF-IDF vector inputs or a combination of TD-IDF vector inputs and meta-features (3). Results showed that using a BETA text encoder and passing the features through a feed-forward neural network outperformed all other machine learning models for all three setups (7).

3 Data

Netflix shares a monthly list of dates on which new movies stream on their platform. The online repository of Netflix movies, "What's on Netflix", is the source for the list of 106 streaming Netflix movie names manually extracted for querying on Twitter (Moore 2022). Netflix also releases a weekly list of top ten ranked movies and television series going back to June 2021 on their website "top10.netflix.com", downloadable as a CSV file. This information helps create a manual data frame of movies and their corresponding classification of top ten or not used as binary labels (1/0) to train the machine learning classifiers.

The data mining process in this study uses Tweepy, a Python-based application programming interface (API), to access Twitter's backend server to collect its users' public tweets (Ahmad 2020; Parack 2022). Querying for mentions of hashtags containing one of the chosen streaming movies creates a corpus of 46,270 tweets to perform sentiment analysis. Depending on the level of access granted, Twitter's API contains restrictions such as a limit on the number of monthly tweets and a limit to tweets posted in the past seven days. This study bypasses these constraints by scheduling weekly queries.

4 Methods

4.1 Data Pre-Processing

Text cleaning and data pre-processing steps performed prior to applying sentiment analysis methods require special attention due to the presence of exclamation points, degree modifiers (e.g., extremely, marginally), contrastive conjunctions (e.g., but), negations (e.g., isn't, not), emoticons, and emojis; all of which express emotions. VADER, the lexicon-based model used to calculate sentiment tweet scores, uses exclamation points and emojis to finetune words' positive or negative affect. The data cleaning performed includes converting text to lower-case, removing URLs and HTTP links, removing usernames beginning with an @, removing hashtags beginning with a #, and removing special characters except for exclamation points.

4.2 Sentiment Analysis

Sentiment analysis, the process of extracting positive and negative sentiments from a text, can be performed either through lexicon-based methods or supervised machine learning methods (Ahmad 2020). This study uses VADER, a lexicon-based sentiment analysis tool developed by Hutto and Gilbert (2014) at Georgia Tech, specifically tailored to microblogs yet effective in

multi-disciplinary corpora. Drawing upon existing world-banks such as LIWC, ANEW, and GI, it incorporates lexical features such as emoticons, sentiment-related acronyms, initialisms, slang, punctuation, capitalization, degree adverbs, the conjunction "but" and negated sentences to finetune intensity ratings on lexical features (220). Sentiment scores reflect sentiment polarity, positive to negative, and sentiment intensity on a scale from -4 to +4, where good has a positive valance of 1.9 while great is 3.4 (220). Once the model produces the sentiment score for all tweets, calculating a count of tweets and an average sentiment score for each movie provides the features to train the classification models.

4.3 Classification Models

Machine learning classification models predict and classify data into discrete classes. Figure 1 displays the modeling approach that predicts whether specific Netflix movies are listed as top ten or not based on the sentiment analysis performed on tweets discussing those specific movies. Supervised learning algorithms Bernoulli Naïve Bayes and Logistic Regression train the model to predict the binary class top ten, listed as 1, or not, listed as 0. The training features are the mean sentiment score per movie and the count of tweets per movie, with 63 movies in the training set and 43 movies in the test set.

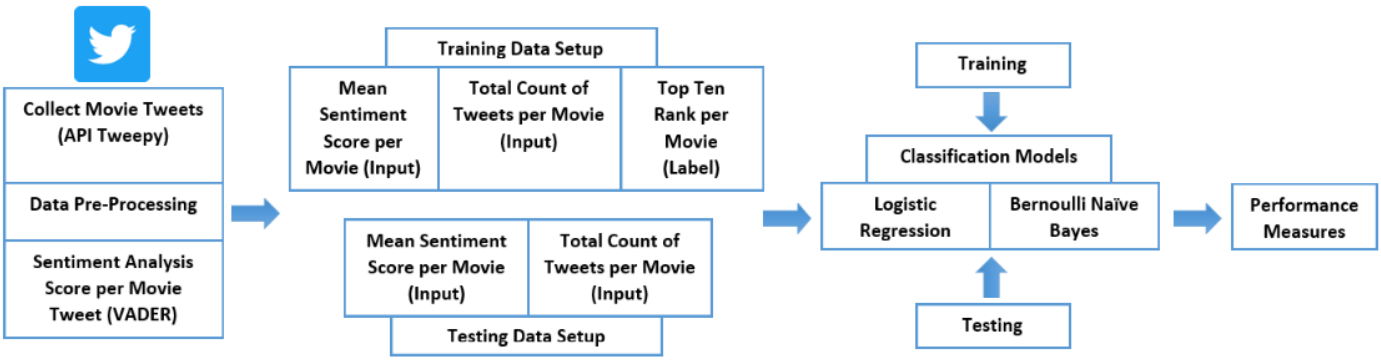

Figure 1. Sentiment analysis modeling approach using supervised learning classification models Logistic Regression and Bernoulli Naïve Bayes.

4.3.1 Bernoulli Naïve Bayes

The classification algorithm Bernoulli Naïve Bayes, based on Bayes' Theorem, measures the independent probability of events A (label) and B (features) as

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

(Medhat, Hassan, and Korashy 2014, 1099). Specifically, Bernoulli Naïve Bayes determines the probability of binary outcomes, where trials are independent, and each outcome has a constant probability per trial (Esposito and Esposito 2020, chap.14). The goal of Bernoulli Naïve Bayes in this study is to calculate the conditional probability of the features, i.e., mean sentiment score or number of tweets given a class or label, i.e., top ten, or not for a particular movie (Esposito and Esposito 2020, chap.14).

4.3.2 Logistic Regression

The second classifier, Logistic Regression, estimates the probability that an input belongs to a particular class, where a probability greater than 50% predicts that the input belongs to the desired class, labeled 1; otherwise, it does not belong, labeled 0 (Geron 2019, chap. 4). This binary classifier weighs the sum of the input features plus a bias term. It then measures the logistic of the result by using a sigmoid function shown in Figure 2, which outputs a number between 0 and 1 (Geron 2019, chap. 4).

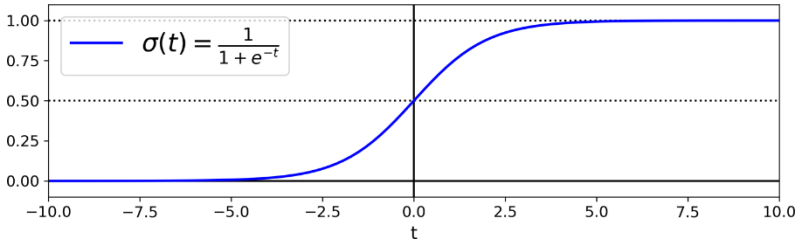

Figure 2. Logistic function.

Source: (Geron 2019) https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch04.html#idm45022189757752

5. Results

Results show that the Bernoulli Naïve Bayes classifier slightly outperforms the Logistic Regression model with F1-scores of 0.54 and 0.49, respectively. This study uses the F1-score instead of the accuracy rate given the presence of an imbalanced dataset where 63 out of 106 movies do not belong to the desired top ten rank class, which affects the accuracy rate since it is dependent on the distribution of classes (Karabulut, Özel, and Ibrikci 2012, 41). The confusion matrices shown in Figure 3 illustrate the main issue, the models' bias toward movies not in the top ten class.

Bernoulli NB	Predicted: 0 Not Top 10	Predicted: 1 Top 10
Actual: 0 Not Top 10	25	1
Actual: 1 Top 10	15	2

Logistic Regression	Predicted: 0 Not Top 10	Predicted: 1 Top 10
Actual: 0 Not Top 10	25	1
Actual: 1 Top 10	16	1

Figure 3. Confusion matrices of Bernoulli Naïve Bayes and Logistic Regression model predictions.

The receiving operating characteristic (ROC) curves in blue shown in Figure 4 indicate the models' tradeoff between the True Positive Rate and the False Positive Rate. In contrast, the black dotted lines represent the results of random guessing (Molin 2019). Both models perform slightly better than a random guess, with the Logistic Regression model at times performing worse, as shown by the blue line dipping under the dotted line.

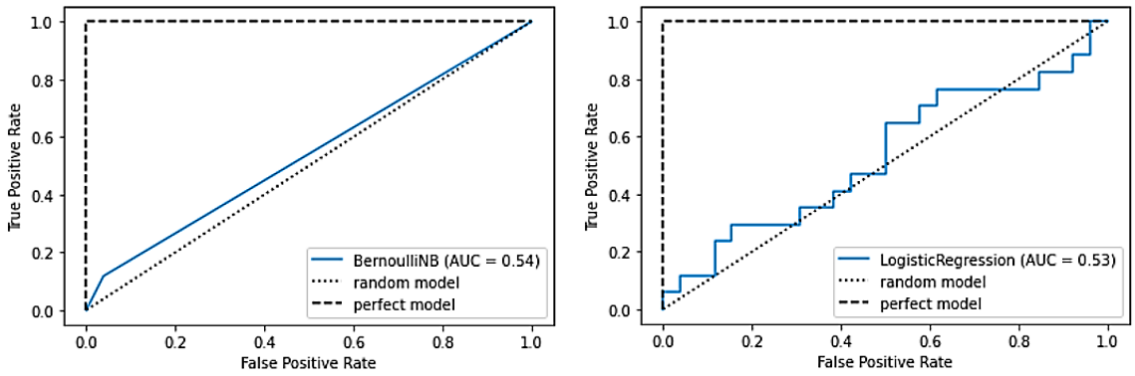

Figure 4. ROC curves of Bernoulli Naïve Bayes and Logistic Regression model predictions.

6. Analysis and Interpretation

This study's methodology differs from the conventional literature on analyzing tweets for sentiment analysis and opinion prediction. It trains two classification models on the average sentiment score per movie and the total number of tweets generated per movie instead of training on the actual tweets and their corresponding sentiment scores. However, the accuracy of the models implemented is low, possibly due to constraints such as corpus size, training methodology, and twitter demographics.

The Twitter API standard access level limits tweet querying to the past seven days and caps the number of tweets per session and per month. Given the constant addition of new streaming movies on Netflix, tweets about older movies become scarce as new movies become a popular topic. Therefore, the two weeks following the release of a movie are pivotal to building a comprehensive corpus of tweets for that movie. Missing this time frame or capping the number of tweets affects the training model due to an incomplete dataset. For example, in this study, only 28% of movies queried have 100 tweets or more from which to assign an accurate mean sentiment score. This limited corpus size partly explains the low performance in predicting a movie's top ten rank status.

The methodology is another potential explanation for the model's low performance. Given the low average number of tweets queried per movie, training the models on the actual tweets and their associated sentiment scores would have increased the number of training and test instances, possibly allowing for better performance. Furthermore, the mean sentiment scores in this study include neutral sentiments, tweets that do not include opinion word markers and therefore have a score of 0. As shown in Figure 5, many tweets are neutral, diminishing the

average sentiment score; consequently, movies with more negative sentiment scores appear less negative, and movies with more positive sentiment scores appear less positive when averaged.

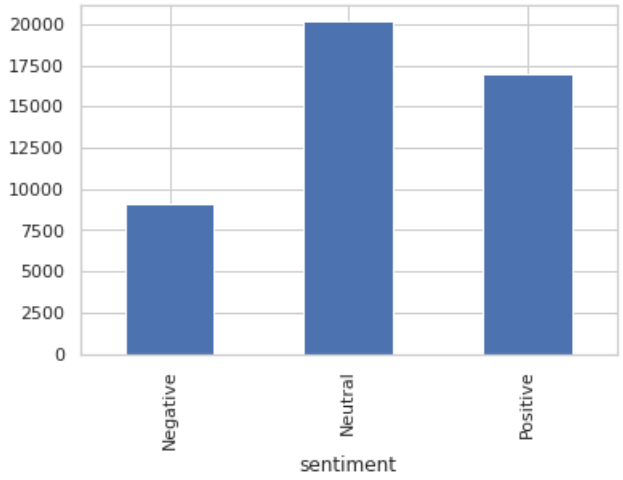

Figure 5. Total number of Negative, Neutral and Positive tweets from which the average scores per movie are computed.

Finally, the restricted demographics of Twitter may be one of the biggest reasons for the lack of performance. A Pew Research Center study (Shah, Remy, and Smith, 2020) found that 92% of tweets come from 10% of their users, meaning that most users engage by re-tweeting, liking, or commenting on original posts. Another report found that 59% of Twitter users are between the ages of 25 and 49, male users outnumber females, and the platform only comprises 8% of social media users (Dean 2022). Figure 6 illustrates how a small corpus and an underrepresented demographic can result in more positive tweets for non-top-ranked movies listed as 0 than for top-ranked movies listed as 1.

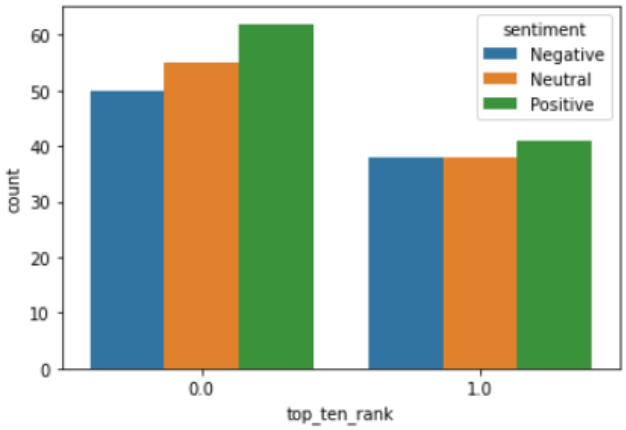

Figure 6. Bar chart of sentiment scores based on class (1 - top ten rank, 0 - not top ten rank).

7. Conclusions

This study aims to predict the popularity of Netflix movies based on their appearance on Netflix's weekly top ten movies list by training two classification models on the average sentiment score of tweets regarding these movies and the number of tweets generated for each movie. The classification models Bernoulli Naïve Bayes and Logistic Regression are trained using a labeled set of movies identified as being in the top ten class, labeled as 1 or not, labeled as 0. Results show that Bernoulli Naïve Bayes slightly outperforms the Logistic Regression model; however, the ROC curve shows that both models perform only slightly better than a random guess. Some reasons may be a small corpus due to Twitter API restrictions, using the direct mean sentiment scores rather than the actual tweets to train the model, and the disproportionate demographics of Twitter users.

8. Directions for future work

This study's topic can evolve into a valuable tool for film production companies to anticipate the genres, characters, and plots that audiences enjoy. However, as discussed, corpus size, methodology, and a broad demographic are essential factors to develop for improving

model performance. Hence, promising areas for future work include increasing corpus size, evaluating different machine learning models and data features, and expanding social media platforms to increase user market reach.

Ensuring a higher Twitter API access level is crucial to bypass data limit restrictions such as the last seven days' query limit and the monthly and per session tweets cap. Furthermore, since most Twitter users do not post original comments, future iterations of this study could include re-tweets and likes as features representing popular opinions about movies. Re-tweets, in particular, would increase the corpus size significantly.

Future updates to this study could remove tweets with neutral sentiments to improve the mean sentiment score and explore a more extensive set of classification algorithms, including Support Vector Machine (SVM), Decision Trees, Random Forrest Classifier, and Gradient Boosting Classifier. A different method could build on the Guda, Srivastava, and Karkhanis (2022) study, which trains a Bidirectional Encoder Representations from Transformers (BERT) model with both text and meta-features for sentiment classification (13). As applied to this study, meta-features could include sentiment score, number of tweets and re-tweets, number of likes, and movies genre.

As a final point, to access a more diver user-base, this study could incorporate data from Google Trends by using the PyTrends python package to identify a movie's favorability by searching for combinations of movie titles with different positive and negative opinion terms. Another way would be by leveraging the Instagram API to access and analyze text on Instagram posts mentioning select Netflix movies.

References

Ahmad, Imran. 2020. *40 Algorithms Every Programmer Should Know*. *O'Reilly*. Birmingham, UK: Packt Publishing Limited. https://learning.oreilly.com/library/view/40-algorithms-every/9781789801217/.

Baysinger, Tim. 2021. "Netflix Dominates Emmys with 44 Wins, Led by 'the Crown' and 'the Queen's Gambit'." TheWrap. TheWrap, September 20, 2021. https://www.thewrap.com/emmy-winners-by-the-numbers-netflix-laps-the-field-with-44-wins-led-by-the-crown-and-the-queens-gambit/.

Bengfort, Benjamin, Rebecca Bilbro, and Tony Ojeda. 2018. *Applied Text Analysis with Python*. *O'Reilly*. Sebastopol, CA: O'Reilly Media, Inc. https://learning.oreilly.com/library/view/applied-text-analysis/9781491963036/.

Dean, Brian. "How Many People Use Twitter in 2022? [New Twitter Stats]." 2022. Backlinko. https://backlinko.com/twitter-users.

Esposito, Dino, and Francesco Esposito. 2020. *Introducing Machine Learning*. *O'Reilly*. Microsoft Press. https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch04.html#idm45022189758376.

Geron, Aurelien. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems*. *O'Reilly*. https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch04.html#idm45022189758376.

Guda, Bhanu Prakash Reddy, Mashrin Srivastava, and Deep Karkhanis. 2022. "Sentiment Analysis: Predicting Yelp Scores." *CoRR* abs/2201.07999 (February 1, 2022). https://doi.org/ https://doi.org/10.48550/arXiv.2201.07999.

Hatzivassiloglou, Vasileios, and Kathleen R. McKeown. 1997. "Predicting the Semantic Orientation of Adjectives." *Proceedings of the 35th annual meeting on Association for Computational Linguistics* -, July 1997, 174–81. https://doi.org/10.3115/976909.979640.

Hutto, C., and Eric Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1):216-25. https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Karabulut, Esra Mahsereci, Selma Ayşe Özel, and Turgay Ibrikci. 2012. "A comparative study on the effect of feature selection on classification accuracy." Procedia Technology 1: 323-327. https://doi.org/10.1016/j.protcy.2012.02.068.

Kedia, Aman, and Mayank Rasu. 2020. *Hands-on Python Natural Language Processing: Explore Tools and Techniques to Analyze and Process Text with a View to Building Real-*

World NLPApplications. *O'Reilly*. Birmingham, UK: Packt Publishing Limited. https://learning.oreilly.com/library/view/hands-on-python-natural/9781838989590/.

Kumar, Ashish, and Joseph Babcock. 2017. *Python: Advanced Predictive Analytics*. *O'Reilly*. Birmingham, UK: Packt Publishing Ltd. https://learning.oreilly.com/library/view/python-advanced-predictive/9781788992367/.

Md, Mohiuddin, Abdul Qudar, and Vijay Mago. 2020. "TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis." *arXiv preprint*. arXiv:2010.11091 (October 17, 2020). https://doi.org/arXiv:2010.1109.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. "Sentiment Analysis Algorithms and Applications: A Survey." *Ain Shams Engineering Journal* 5, no. 4 (December 2014): 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

Mohammad, Saif M. 2016. "Sentiment analysis: Detecting valence, emotions, and other affectual states from text." In *Emotion measurement*, pp. 201-237. Woodhead Publishing. https://doi.org/10.1016/b978-0-12-821124-3.00011-9.

Molin, Stefanie. 2019. *Hands-on Data Analysis with Pandas: Efficiently Perform Data Collection, Wrangling, Analysis, and Visualization Using Python*. *O'Reilly*. Birmingham, UK: Packt. https://learning.oreilly.com/library/view/hands-on-data-analysis/9781789615326/3e34bb4a-05e5-491e-9f38-d25b65930ce6.xhtml.

Moore, Kasey. "What's Coming to Netflix in February 2022." What's on Netflix, February 20, 2022. https://www.whats-on-netflix.com/coming-soon/whats-coming-to-netflix-in-february-2022-02-20/.

Parack, Suhem. 2022. "A Comprehensive Guide for Using the Twitter API V2 with Tweepy in Python." DEV Community. DEV Community. https://dev.to/twitterdev/a-comprehensive-guide-for-using-the-twitter-api-v2-using-tweepy-in-python-15d9.

Philander, Kahlil, and YunYing Zhong. 2016. "Twitter Sentiment Analysis: Capturing Sentiment from Integrated Resort Tweets." International Journal of Hospitality Management 55: 16–24. doi: http://dx.doi.org/10.1016/j.ijhm.2016.02.001.

Shah, Sono, Emma Remy, and Aaron Smith. 2020. "How Democrats and Republicans Use Twitter." Pew Research Center - U.S. Politics & Policy. Pew Research Center. https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/.

Spangler, Todd. 2021. "Netflix Wins Seven Oscars, Biggest Haul among All Studios This Year." Variety. Variety, May 13, 2021. https://variety.com/2021/awards/news/netflix-oscars-most-wins-1234959949/.

Tapp, Tom. 2021. “The Most-Watched Netflix TV Shows & Movies Ever – Update.” Deadline. Deadline, October 13, 2021. https://deadline.com/feature/most-watched-netflix-shows-movies-1234846488/.

“Top 10 by Country.” Netflix Top 10 - By Country: United States, 2022. https://top10.netflix.com/united-states.

Turney, Peter D., and Michael L. Littman. 2003. “Measuring Praise and Criticism: Inference of Semantic Orientation from Association.” *ACM Transactions on Information Systems* 21, no. 4 (2003): 315–46. https://doi.org/10.1145/944012.944013.

Zhang, Lei, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. "Combining lexicon-based and learning-based methods for Twitter sentiment analysis." *HP Laboratories, Technical Report HPL-2011* 89 (2011): 1-8.