SI 699

Are Celebrity Endorsements Worth The Money?

Nitanshi Mahajan and Emily Schemanske

Introduction

Are celebrity endorsements worth the money?

Many companies use celebrities to promote their products; a wise move since, according to Forbes, a company could see as much as a 4% increase in sales almost immediately after an endorsement deal is signed (Olenski). At the same time, these lucrative deals are not without risk. Recent public scandals and the rise of “cancel culture” have raised questions about the potential financial impact on companies when things go wrong.

This paper aims to understand the degree to which a company’s financial performance may be impacted by a celebrity endorser’s scandal, using empirical data and case study to provide a preliminary analysis of the issue. The overarching question will be, how can we measure such a thing? By shedding light on this topic, this study offers valuable insights for companies looking to weigh the risks and benefits of celebrity endorsements in today’s highly competitive market.

Methodology

Our first task is to quantify a “scandal”. We will do this by applying Natural Language Processing techniques to extract sentiment from published articles - in particular, Us Weekly magazine. We will combine our insights from that analysis with stock data for the company to understand the degree to which it impacts a stock’s performance. For this part of the analysis we will use linear regression to build a model and interpret the results.

Finally, we will visualize the data to get a holistic view of the relationship between celebrity and company. For this case study, we will look at the nine year long partnership of Kanye West and Adidas.

Data

Text Data

In order to collect data for the report, it was originally intended to obtain information from Twitter. However, this task proved to be difficult due to the paywall restrictions. As a result, an alternative data source was sought from news and magazine outlets. The initial plan was to collect data from Us Weekly, Peoples, and the New York Times.

However, due to the paywall of the New York Times and the frequent changes to the user interface of People's website, it was challenging to collect data from these sources. Attempts were made to utilize the Wayback Machine to obtain snippets of the People's website from the past, but this proved to be a complicated process. Ultimately, the decision was made to use Us Weekly as the primary source for scraping data. The Beautiful Soup tool was utilized to collect the date, header, and content of the articles from Us Weekly.

Text Processing

Data cleaning was performed to ensure that only relevant articles were used for analysis. The first step in this process was to filter out articles that were not related to the problem. To achieve this, a name-based filtering strategy was applied, where the articles were checked for the presence of certain names of people such as Kanye, Kim, West, Kardashian, etc. The articles that did not contain any of these keywords in either the header or content were removed.This was achieved using Named Entity Recognition (NER) on the articles to extract the names of all individuals mentioned in them, with a focus on the 'PERSON' entity type. This was done using NLTK's pre-trained NER model.The Python library NLTK's Named Entity Recognition (NER) is based on machine learning algorithms, specifically on the Maximum Entropy Markov Model (MEMM) and Conditional Random Fields (CRF) models. These models use statistical methods to identify and classify named entities in text, such as persons, organizations, locations, and dates. NER is a supervised learning task, where the algorithm is trained on a labeled dataset of text with annotated named entities.

Out of the 1,975 articles that were initially collected, around 1,600 articles mentioned Kanye either in the header or the article content, and approximately 500 articles mentioned him in both. Only the articles that mentioned Kanye in both the header and content were used for further analysis.

To further refine the data, the number of 'PERSON' entities found in each article was counted. If the number of 'PERSON' entities was four or more, it was assumed that the article did not revolve particularly around the person of interest, Kanye. Conversely, if the number of 'PERSON' entities was less than four, it was assumed that the article was more likely to be about Kanye.

However, limitation to this process is that even with these filtering steps, there was still no foolproof way to filter the data that solely revolved around Kanye. Despite this caveat, the data cleaning process helped to remove irrelevant articles and provided a more focused dataset for analysis.

Financial Data

To study the impacts of public sentiment on Adidas' financial health, relevant stock data was gathered from investing.com, including opening and closing prices, trading volume, and percent change at weekly[1] intervals.

The next step was to gather data for a market index to use as a benchmark to normalize the company's stock performance. Adidas is a German company, and in the United States its stock is traded on the OTCQX International Premier market. The OTCQX Composite Index tracks the overall performance of that market and is similar to the Dow Jones Industrial Average and other more well-known indices (Israel). However, the OTCQX Composite was created on January 2, 2015, while Kanye West's relationship with Adidas began in 2013. It was preferable to find an index that covered the entire partnership. Thus, to determine if it would be appropriate to substitute DJIA or the S&P 500, data was collected for all three (also from investing.com) and the correlation of the week-over-week percent change for each index was computed.

Correlation of week-over-week percent change			
	OTCQX Composite	**Dow Jones Industrial Average**	**S&P 500**
OTCQX Composite	1.00	0.83	0.82
Dow Jones Industrial Average	0.83	1.00	0.96
S&P 500	0.82	0.96	1.00

Unsurprisingly the Dow and the S&P were nearly perfectly correlated, but importantly, they were also highly correlated with the OTCQX Composite. Either index would be a sufficient substitute; ultimately, the Dow was chosen due to its slightly higher correlation with the OTCQX as compared to the S&P.

[1] Weekly intervals were chosen to align with the frequency of our text data.

Models

Sentiment Analysis

Sentiment analysis is a popular application of natural language processing (NLP) that involves identifying the sentiment expressed in a given piece of text. Typically, sentiment analysis is performed using supervised learning algorithms that require labeled training data. However, in this case, as labeled data is not available, we can either perform unsupervised sentiment analysis using methods like clustering or use pretrained models that are trained on domain-specific data like social media data which can perform social NLP tasks.
One advantage of using pretrained models is that they can be used to perform sentiment analysis on a wide range of texts, without requiring any additional training data. Another advantage is that these models are often very accurate, due to their large training datasets and advanced training algorithms. However, one disadvantage of using pretrained models is that they may not perform well on domain-specific texts or texts with unique nuances. As the models we used in our analysis revolve around social media (magazine in our case), using models that are pre-trained on social media would provide a great base.

In contrast to pretrained models, unsupervised clustering methods can be used to perform sentiment analysis without requiring any labeled training data. Clustering methods work by grouping similar texts together based on some similarity metric, such as cosine similarity or Jaccard similarity or any other similarity metric. Once the texts are clustered, the sentiment of each cluster can be inferred based on the sentiment of its constituent texts. Lexicon-based methods rely on sentiment dictionaries or lexicons that contain lists of words and their associated sentiment scores. The sentiment scores can be positive, negative, or neutral. Rule-based methods: Rule-based methods use a set of predefined rules to identify sentiment in text. These rules can be based on grammatical patterns, semantic rules, or a combination of both. Rule-based methods can be useful in cases where lexicon-based methods are not effective, such as when analyzing text in a specific domain or when dealing with sarcasm or irony.
However, there are many limitations to the above mentioned methods especially in the context of social media data. Social media data can be highly informal and often contains misspellings, slang, and abbreviations. This can make it difficult for unsupervised sentiment analysis methods to accurately identify sentiment, as they may not recognize the nuances in language. Social media data is highly dynamic and constantly evolving, with new words, expressions, and trends emerging all the time.

Unsupervised sentiment analysis methods may not be able to keep up with these changes, leading to outdated sentiment scores and inaccurate results. Since Unsupervised sentiment analysis methods rely on lexicons or rules to identify sentiment in text, which can be limited in their accuracy. These methods may not capture nuances in language or context-specific meanings.
In conclusion, unsupervised methods are often less accurate and more complex than supervised methods. Using pretrained models which are customized for social NLP tasks are context-sensitive. Therefore we decided to move forward with using pretrained models for the sentiment analysis and scoring task.

In this analysis, we used 2 pretrained models - NLTKs' pretrained Sentiment Intensity Analyzer, another was a pre-trained model from Hugging Face which was trained on twitter data and designed specifically for social NLP tasks.

NLTK's pre-trained Sentiment Intensity Analyzer(SIA) is a tool that can be used for sentiment analysis of text data. It is a pre-trained model that uses a lexicon-based approach to determine the sentiment of a piece of text. The SIA assigns a sentiment score to each piece of text, with positive scores indicating positive sentiment and negative scores indicating negative sentiment. The score ranges from -1 (most negative) to 1 (most positive), with 0 indicating neutral sentiment. It uses a variant of the Valence Aware Dictionary and Sentiment Reasoner (VADER) lexicon, which was specifically designed for sentiment analysis of social media text.
The VADER lexicon is pre-loaded with sentiment scores for thousands of words and emoticons, which are used to calculate the overall sentiment score for a piece of text.

'FiniteAutomata/bertweet-base-sentiment-analysis' is a pre-trained transformer-based model available on the Hugging Face model hub. This model is based on BERTweet, a RoBERTa-based model that has been specifically trained on a large corpus of English tweets for social NLP tasks.

Linear Model

The goal of this model is to explain the percent change of Adidas stock week-over-week, using the Dow Jones Industrial Average to represent market effects, and leaving the unexplained (or excess) returns to be accounted for by the other variables in the model (namely, social media sentiment). The regression equation takes the following form:

$$a = \beta_0 + \beta_1 m + \beta_2 s_1 + ... + \beta_n s_i + ... + \epsilon ,$$

where

- a is the vector of Adidas stock weekly percent changes (26-week moving average)
- m is the vector of DJIA weekly percent changes (26-week moving average)
- s_i are one or more vectors of sentiment features,
- ϵ is the unexplained component of the model.

The β values are the regression coefficients given by the linear model. The model was based on the work of Fekrazad et al (2022).

Two different sets of sentiment scores were tested as predictors. The first set came from Hugging Face's BERTweet model (Nguyen et al. 2020), and the second from NLTK's VADER sentiment tool (Hutto and Gilbert, 2014). Base models give a starting point to understand the effects - if any - of gossip magazine article sentiment on the market. They are univariate models whose only predictor is the Dow Jones Industrial Average, and there is one for each of the NLTK and Hugging Face sentiment scores.

Hugging Face Model	NTLK Model
Residuals: Min 1Q Median 3Q Max -2.45728 -0.51758 0.01816 0.54109 1.73822 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -0.04695 0.05169 -0.908 0.364 djia_PercentChangeMA 1.61906 0.13898 11.650 <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.8016 on 307 degrees of freedom Multiple R-squared: 0.3065, Adjusted R-squared: 0.3043 F-statistic: 135.7 on 1 and 307 DF, p-value: < 2.2e-16	Residuals: Min 1Q Median 3Q Max -2.45728 -0.51758 0.01816 0.54109 1.73822 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -0.04695 0.05169 -0.908 0.364 djia_PercentChangeMA 1.61906 0.13898 11.650 <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.8016 on 307 degrees of freedom Multiple R-squared: 0.3065, Adjusted R-squared: 0.3043 F-statistic: 135.7 on 1 and 307 DF, p-value: < 2.2e-16

These results are the same since the same predictor from each data frame was used - they simply provide confirmation that the next steps will be comparable. $R^2 = 0.3065$, which means that the DJIA alone explains roughly 31% of the variance in Adidas stock performance over the last ten years.

Next, for each set of sentiment scores, features were selected using backward elimination; a model with all predictors is built, and the predictor with the highest $p-$value that is greater than the desired significance level is deleted. A new model is made from the remaining features and the process is repeated until all predictors are significant. The results:

<table>
<tr><th>Hugging Face Model</th><th>NTLK Model</th></tr>
<tr><td>
<pre>
Residuals:
     Min   1Q   Median   3Q   Max
-2.16677 -0.49365 -0.01292 0.50323 1.76221

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.4052   0.1090 -3.718 0.000239 ***
djia_PercentChangeMA   1.6531   0.1365 12.112 < 2e-16 ***
pn_score_MA   0.6818   0.1837   3.712 0.000244 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7855 on 306 degrees of freedom
Multiple R-squared: 0.3364,	Adjusted R-squared: 0.3321
F-statistic: 77.57 on 2 and 306 DF, p-value: < 2.2e-16
</pre>
</td><td>
<pre>
Residuals:
     Min   1Q   Median   3Q   Max
-2.16372 -0.51496 0.01823 0.41608 1.71222

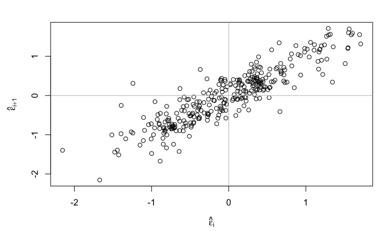
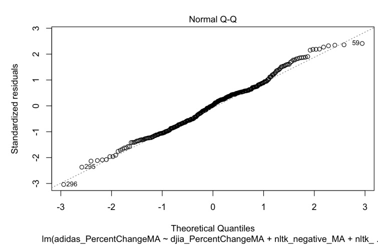
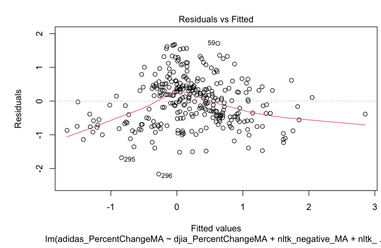
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.5859   4.7350   6.671 1.20e-10 ***
djia_PercentChangeMA   1.7929   0.1339 13.390 < 2e-16 ***
nltk_negative_MA   -57.7529   10.9951 -5.253 2.82e-07 ***
nltk_neutral_MA   -30.5085   5.0806 -6.005 5.45e-09 ***
nltk_compound_MA   -4.4275   0.6386 -6.933 2.48e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7153 on 304 degrees of freedom
Multiple R-squared: 0.4533,	Adjusted R-squared: 0.4462
F-statistic: 63.03 on 4 and 304 DF, p-value: < 2.2e-16
</pre>
</td></tr>
</table>

The adjusted R^2 looked more promising for the NLTK model, so it was used going forward.

Next, the assumptions of normality, independence, and constant variance were examined.

<table>
<tr><th>Independence</th><th>Normality</th><th>Constant Variance</th></tr>
<tr><td></td><td>
</td><td>
</td></tr>
<tr><td>The lag plot shows strong evidence of autocorrelation.</td><td>The QQ-plot supports our assumption of normality.</td><td>The structure present in this plot suggests a non-linear relationship.</td></tr>
</table>

The primary concern from these diagnostics is that there is strong evidence for autocorrelation - a common problem for time series data. To correct it, the predictors of the full model were transformed using the Cochrane-Orcutt estimation (Verbeek, 2005), and the model was refitted using the backward elimination process.

NLTK Transformed Base Model	NLTK Transformed Final Model

NLTK Transformed Base Model

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8035 -0.1684  0.0061  0.1603  1.2071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.00657    0.01600   0.411    0.682
x.trans      0.91243    0.10625   8.587  4.6e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2802 on 306 degrees of freedom
Multiple R-squared:  0.1942,    Adjusted R-squared:  0.1916
F-statistic: 73.74 on 1 and 306 DF,  p-value: 4.596e-16
```

NLTK Transformed Final Model

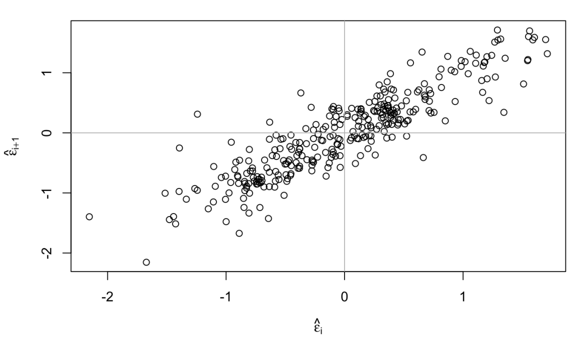
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.81880 -0.15390  0.00414  0.16331  1.18406

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06305    0.03797  -1.661   0.0978 .
x.trans      0.92271    0.10584   8.718   <2e-16 ***
x1.trans    24.15942   11.96050   2.020   0.0443 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2788 on 305 degrees of freedom
Multiple R-squared:  0.2048,    Adjusted R-squared:  0.1996
F-statistic: 39.28 on 2 and 305 DF,  p-value: 6.61e-16
```

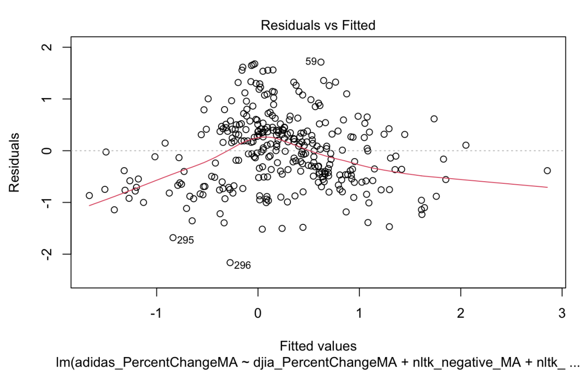
Unfortunately some of the explanatory power was lost after the transformation, with $R^2 \approx 0.2$. This value is roughly the same in both the base model with one predictor, and the model that includes the negative sentiment score (the only other significant feature).

Happily, though, the problems diagnosed in the plots above are improved, indicating that our assumptions of normality, independence, and constant variances are now upheld, which is required before anything can be concluded from the model.

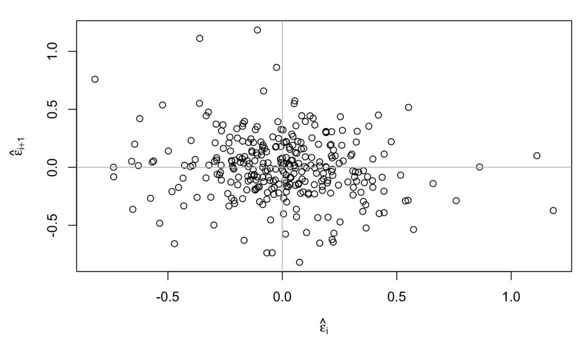
Before Transformation	After Transformation

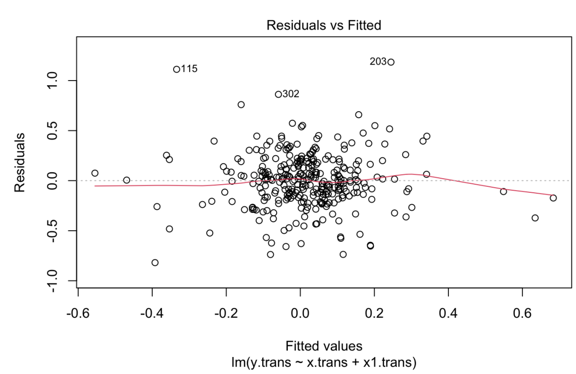
Before Transformation





After Transformation





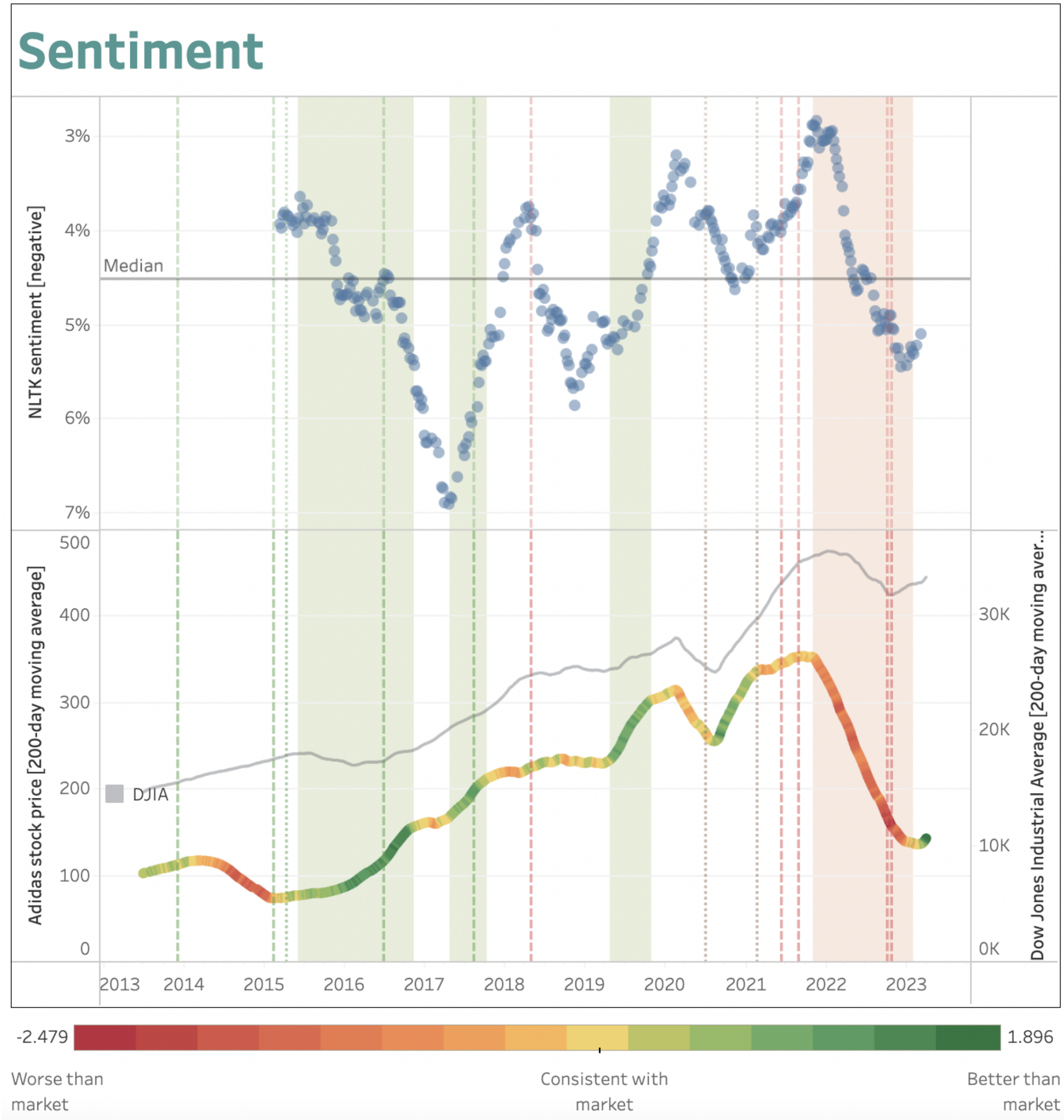
Results

With so many factors that can affect stock prices, such a low R^2 is to be expected with only two features in the model. What we were hoping to see was that the addition of a sentiment feature to the Dow Jones Industrial average would explain much more of the variance. It did not. The adjusted R^2, which is a better indicator of the impact of added features, increased by less than one hundredth. Thus, it seems that the sentiment features we collected did not add any explanatory power.

This could have happened for any number of reasons. Probably one of the biggest shortcomings in our work was the quality of the text data we collected. Twitter would have been the ideal source, but there were serious questions about whether we would have reliable access to it for the duration of our project. Unfortunately, Us Weekly articles could never be published at the frequency and quantity of Tweets, which means our text data probably falls short of capturing true sentiment toward Kanye West in a timely manner.

However, despite the fact that its significance is highly questionable, our model did find that the negative sentiment score was meaningful. This was consistent with the findings of Fekrazad et al. (2022), who observed that “a higher proportion of negative Tweets about a company posted within an hour/day leads to lower returns and a higher short volume for its stock”, while their study of positive and neutral Tweets did not reveal any meaningful relationship. Practically, it makes sense - people are unlikely to think much about a celebrity who is not causing any trouble, but when a public relations disaster occurs they will demand accountability and consequences.

To get a more comprehensive picture of the partnership of Kanye West and Adidas, you can visit **Tableau Public** to walk through the data. While we can not draw any statistically sound conclusions, interesting patterns are revealed when the data is visualized:



Here we see some support for the notion that negative Tweets tend to give more information about a stock's performance - in particular, the orange band around 2022-2023 shows an uncanny similarity in the decrease in stock price, stock value, and the sentiment towards articles published about Kanye West in Us Weekly.

Future Work

Twitter data can provide valuable insights into public sentiment and can potentially yield better performance when studying the impact of celebrity scandals on company stock prices. For this study Us Weekly magazine was used instead, which contains more structured text and provides contextual information. While it may not be an exact representation of public sentiment, we hoped the data was close enough to draw preliminary conclusions. To support this assumption, future work might incorporate more similar sources like People, Star, and other gossip magazines.

Regardless of the source, when working with such a large amount of text it can be difficult or impossible to know if the sentiment that is captured by the pretrained model is accurate. A potential approach to solving this problem is to label some of the text by hand and then use a model to check the accuracy.

The study used a linear regression model with stock sentiment scores as independent variables and stock prices as dependent variables, but other models like neural networks with additional features could potentially perform better.

In the Linear Regression model, to isolate the direct impact of public sentiment on stock prices, the stock data needs to be normalized to remove other factors that can affect stock prices. This includes endorsements from other celebrities, company's marketing efforts, etc.

Further analysis could be performed to study partnerships between other celebrities and companies to see if these results scale. This might be difficult to do, since companies tend to drop celebrities when things are not going well, so there is often not an opportunity for a meaningful pattern to develop.

Impact

In the world of advertising, endorsement deals with celebrities have become a popular way for companies to increase their brand recognition and credibility. However, such deals can be expensive, often costing millions of dollars. To make an informed and evidence-based decision on whether to invest in such deals, companies need to understand how much stock prices are affected by public perception of the celebrity endorsing their brand.

Researchers need to be careful when using data to analyze certain situations, since there is always the possibility that a celebrity's public crisis is the result of a mental health issue. Great care must be taken to not perpetuate biases or exacerbate mental health disparities. Each situation is unique and must be approached with caution to ensure the celebrity's rights and privacy are respected, and that the data is used in a way that is mutually beneficial.

By using data-driven models made from sentiments of public perceptions of celebrities, companies will be able to evaluate the return on investment (ROI) of endorsement deals and make informed decisions about the worth of these associations. With this study, companies can make more effective marketing strategies and optimize their budgets by prioritizing the endorsement deals that deliver the best ROI.

Conclusion

The aim of the study was to investigate the relationship between public sentiment towards celebrities and the companies they endorse, particularly in the wake of scandals or controversies. By examining data and trends especially 2015 to 2018 and 2021 to 2023, we were able to identify instances where negative public sentiment had a strong correlation with a decrease in stock values.

Although the study focused on sentiments expressed in Us Weekly articles, it provides a starting point for this study to develop an understanding of the impact of public sentiment on the stock performance. By tracking negative sentiment across a range of sources and normalizing stocks by eliminating the effect of other factors on financial performance, this study could provide valuable insights for making well-informed marketing investments based on data-driven analysis. By extending this study to other sources of public sentiment, companies can gain a more comprehensive understanding of how public perception affects their financial performance.
There are still many experiments to be performed and many dimensions to be explored, this could act as a baseline for progressing in this research area.

Statement of Work

This project was a collaborative between two contributors:

1. Nitanshi Mahajan, responsible for scraping, processing, and analyzing text data, and selecting the text features used in the model.
2. Emily Schemanske, responsible for collecting financial data, analyzing the model, and creating the Tableau dashboard.

References

Fekrazad, A., Harun, S. M., & Sardar, N. (2022). Social media sentiment and the stock market. *Journal of Economics and Finance*, *46*(2), 397-419. SpringerLink.

Hutto, C.J., & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. https://www.nltk.org/_modules/nltk/sentiment/vader.html

Israel, S. (2019, February 27). *6 Indexes That Can Help You Track The OTC Markets*. Yahoo Finance. Retrieved April 14, 2023, from https://www.yahoo.com/now/6-indexes-help-track-otc-161248134.html

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. *Proceedings of the 2020 Conference on the Empirical Methods in Natural Language Processing: System Demonstrations*, 9-14.

Olenski, S. (2016, July 20). *How Brands Should Use Celebrities For Endorsements*. forbes.com. Retrieved April 13, 2023, from https://www.forbes.com/sites/steveolenski/2016/07/20/how-brands-should-use-celebrities-for-endorsements/?sh=2fd34b275593

Verbeek, M. (2005). *Econometria*. Zanichelli. Retrieved April 15, 2023, from https://www.rdocumentation.org/packages/orcutt/versions/2.3/topics/cochrane.orcutt

*Some text for this report was edited using ChatGPT.