Submitted to *Marketing Science*
manuscript (Please, provide the manuscript number!)

(Authors' names blinded for peer review)

Existing works have predicted user behavior on social media using either image or text data, however, rarely work with both content types simultaneously. We propose and validate a methodology for combining image and text data for predicting user behavior on social media. So far, studies have shown the tendency for images to improve user engagement [Somerfield et al., 2018, Wagner et al., 2015], especially for images with positive emotional content [Chen and Dredze, 2018]. Using 30k Facebook firm-generated content (FGC) posts, we find that the combined model can predict for a greater comment sentiment, comment count, and share count 93%, 65%, and 63% of the time. The model achieves a 3.5x improvement in MSE when predicting share count and a 14% improvement for comment sentiment. Finally, the study demonstrates the ability to pick more performant advertisement. The results validate the need for using both image and text data for predicting user behavior on social media.

Key words: social media, user behavior modeling, advertisements, Facebook

1. Introduction

Social media serves as a platform where brands create and maintain an online presence [Greenwood et al., 2016]; the goal is that tangible user engagements result in faster user conversions Authors (2013). Social media can create tangible value that improves business performance [Barreto, 2013]. Tangible benefits include a decreased time needed for users to make a buying decision [Barreto, 2013]. Intangible benefits include how advertisements influence buyer decisions [Barreto, 2013]. Both the decreased time to buy and improved influence positively affect business performance. In addition, with better forecasting, advertisers can improve planning their sales cycles and projected revenue

[Imsa and Irwansyah, 2020]. Advertisers remain concerned with social media metrics, especially those that promote engagement [Tiago and Verissimo, 2014], including click-through rate (CTR), brand awareness, and word-of-mouth buzz. Advertisers associate these with advertisement return on investment (ROI), which is known as the Holy Grail of social media [Fisher, 2009]. Nevertheless, advertisers calculate ROI, which often includes an increase in user interaction [Romero, 2011], (Anonymous for Blind Review).

In order to impact future sales from untapped markets [Guo et al., 2020] advertisers aim to brand stickiness, improve user relationship quality, create unique visitors, increase average time per visit to their website, get repeated visitors, and increase visit frequency [Bhat et al., 2002]. There are many ways to improve advertisement campaign performance [Imsa and Irwansyah, 2020]. Examples that improve advertisement performance might include the use of pictures in advertisements [Wagner et al., 2015] and providing content with positive emotional content [Chen and Dredze, 2018]. However, neither of these provides a direct forecast of the advertisement's performance. Given the cost of creating and showing ads, quicker feedback mechanisms that can predict advertisement performance are useful in curating content and publishing on social platforms with a greater degree of confidence concerning the advertisement's performance [Hu et al., 2016].

However, a gap exists in current research when modeling user behavior with social media data because researchers fail to incorporate multiple data types, despite the availability of both image and text data. This gap is driven in part by the maturity of image processing methodologies and corresponding technological requirements. Image processing is an emergent application in marketing, which still requires considerable computing power and technological fluency. Many prior works of note only focus on a single data type, we assume because working with a single data type is more simple and often performs quite well

for its purpose, including predicting gender classification [Hassner and Tal, 2015], detecting sarcasm [Poria et al., 2016], profiling [Segalin et al., 2017], and predicting social media popularity [Gelli et al., 2015]. Some attempts exist which recognize the linking of the two data types, i.e., performing sentiment analysis of posts with images [Wang et al., 2015] but not incorporating the post's text into the model or using text to predict a post's CTR but ignoring its associated image data [Hudson et al., 2020]. Very little research exists which treats text and image data concurrently. The research gap is not utilizing multiple available data types when modeling user behavior on social media.

Our first research question asks whether a combination of text and image data better predict user engagement on social media using machine learning. The question explores the existing gap in designing a performant machine learning architecture that digests both image and text data to predict user engagement. Such an architecture might include text-based NN, CNNs, and popular models like decision trees. The predicted user engagement consists of the count of likes, comments, shares, and comment sentiment. Models that predict numbers use regression and mean-squared error (MSE) as their loss function. We explore whether model architectures that combine text and image better produce a model with a lower loss using regression and mean-squared error (MSE) as the loss function than their text and image counterparts.

The second research questions asks what content type, text or images, are better for predicting user comment sentiment using machine learning? Existing studies use images in CNN models to predict visual sentiment [Segalin et al., 2017, Xu et al., 2014]. If a relationship exists between visual sentiment and user response, we expect image-based CNNs to perform well at predicting user sentiment. Nevertheless, text-data provides a great deal of the post's content. It is worthwhile to juxtapose how well text-based and image-based

models compare. This research question will examine the ability for image-based models to predict the sentiment of users, via their comments compared to text models.

The final research question explores to what extent can machine learning models predict which of any two social media advertisements will perform best? An eventual goal of advertisers and marketing research is in having a low(er) cost mechanism to suggest which of two advertisements would have higher ROI. Machine learning models that predict user engagement might be able to select, from a group of advertisements, which one will perform the best. The ability to select the best performing advertisement beforehand allows brands to better select advertisement content. One implication is the ability to choose the best post. A further implication might be automated A/B testing with different variations of the same advertisement. Brands can use such a machine learning model to better tweak their advertisement so it best performs on social media. As far as we are aware, successful selection of the best performing advertisement, from a group of ads, has not been done with machine learning models with social media advertisements. This paper explores whether the curated machine learning models can predict which advertisement, will perform best on social media.

1.1. Solution

In response to the research questions, we provide a method for combining text and image data for modeling user behavior on social media. We demonstrate the successful implementation of a model combining image and text data and demonstrate its improved performance over single-data type models. The chosen method makes use of an ensemble model whose input is a text-based Neural Network (NN) and image-based Convolutional Neural Network (CNN). The combined model outperforms the text and image models when predicting user share, comment, and comment sentiment. The results of twelve tested machine

learning models are provided along with insights concerning model performance and its application to social media advertising.

The remaining paper consists of five sections. The related works covers existing studies that model user behavior on social media, delineating studies relying on text data or image data. We also include and compare methodologies for processing text and image data within the related works section. The methodology section describes and justifies twelve machine learning models: three text-only, three image-only, and six combined models with different architectures. The result section outlines the result of the combined model, juxtaposed with text-only and image-only models. The discussion delineates why the combined model produces an improved performance. The paper concludes with a use case that demonstrates its ability to reliably choose better performing advertisements. The model can predict for a greater comment sentiment, comment count, and share count 93%, 65%, and 63% of the time. In applica- tion, advertisers can utilize these or similar models for tangible benefits concerning their campaign's performance. This research applies its findings to an actual scenario concerning choosing best performing advertisements. The conclusion and future work outline ways future research can adopt these methods to better model user behavior on social media.

2. Related Work

2.1. Modeling User Behavior on Social Media

Modeling user behaviors can provide beneficial research insights about social analytics. Existing research understands user behaviors through behavior models, which often include machine learning models [Hudson et al., 2020, Straton et al., 2017, Xing et al., 2021, Aggarwal and Gupta, 2017, Liu, 2012, Li et al., 2015]. Examples include logistic regression models [Li et al., 2015], neural networks [Straton et al., 2017], text tf-idf models [Ohsawa and Matsuo, 2013], opinion mining [Aggarwal and Gupta, 2017, Liu, 2012], and

sentiment analysis of images [Wang et al., 2015]. Many models trained on text data exist to understand user behavior. Facebook likes have been predicted for hospital data using only post text data [Nash et al., 2017]. Other researchers have used NLP to predict the likelihood a user will follow a Facebook page [Xing et al., 2021, Ohsawa and Matsuo, 2013]. Each example is an exemplar demonstrating a common methodology of modeling user behavior via machine learning models in order to better understand social analytics and user behavior. Each example demonstrate the common methodology of modeling user behavior using machine learning models in order to better understand social analytics and user behavior. Modeling user behaviors can provide beneficial research insights about social analytics.

Convolutional Neural Networks (CNNs) perform well at working with images. Already by 2015 CNNs were considered to be the standard in the realm of social media for image analysis [Hassner and Tal, 2015]. They have been used for gender classification [Hassner and Tal, 2015], for visual sentiment [Segalin et al., 2017, Xu et al., 2014], to detect sarcasm on Twitter [Poria et al., 2016], for detecting stress in social media images [Lin et al., 2014], to perform social media profiling [Segalin et al., 2017], to predict social media popularity [Gelli et al., 2015], and to predict which posts will receive the post clicks [Khosla et al., 2014].

2.2. Industry need for Modeling User Behavior

Companies calculate social media revenue return on investment (ROI) via their advertisement performance on the platform [Fisher, 2009]. Therefore, ROI is the Holy Grail of social media [Fisher, 2009]. When asked which social media metrics marketing managers care about most, they replied with brand awareness, word-of-mouth buzz, customer satisfaction, user-generated content, and web analytics [Tiago and Verissimo, 2014]. However, ROI is difficult to track (Anonymous for Blind Review). Most companies are unable

to get revenue or cost savings from social media because of specific platform affordances [Romero, 2011]. Instead, ROI is measured via user consumption metrics, i.e., (Anonymous for Blind Review) performed a cross-platform analysis of ROI on Facebook, Twitter, and Foursquare showing that advertising-focused tweets could predict rising Foursquare check-ins.

Users visit social media sites to gain information [Schröter et al., 2021, Fisher, 2009, Clark and Melancon, 2013]—for example, 34% of participants post products about opinions from online information. Moreover, traffic to blogs kept increasing to 50% alone that year, compared to 17% at CNN, MSNBC, and the New York Times. 70% of consumers visit social media sites for information and 49% of that 70% buy based on social media content. 36% of participants better rate companies with an online presence and 60% of users pass along social media data to other users.

2.2.1. Text-based Social Media Models Facebook and twitter are commonly used platforms in text-based analyses of advertising effectiveness. Considering Facebook advertisements, [Lotze et al., 2021] categorizes all posts into engagement categories, e.g., low, medium, and high then predicted user metrics include page likes, shares, and comment counts from this data. The employed Neural Network trained on text and time data. Their model can accurately predict for lower user engagement but fails to predict for higher levels of engagement. The study's sample size was 100k posts and did not incorporate images or comment text in its predictions. Nevertheless, the study found that text data can predict limited levels of user engagement.

A CTR study focuses on the likelihood of user clicks based on user interests [Li et al., 2015]. The study is essential because it models user behavior toward advertiser data. The study performs its prediction by modeling the Twitter feed and the click rates for

each type of user interest. As a result, the study successfully predicted user click-through rates based on how well user interests coincide with the advertisement's content. User CTR can be reliably predicted based on user interests.

Research exists that to measure user sentiment using either text or image data. Text data is useful for opinion mining [Aggarwal and Gupta, 2017], where opinion mining uses keywords as sentiment indicators. Fortunately, existing sentiment lexicons are available for predicting sentence sentiment [Georgiou et al., 2015]. In contrast, there is research that detects image sentiment by clustering images [Wang et al., 2015]. Methods exist that use either text or image data to predict user sentiment. Newer studies exist that calculate user sentiment by correlating image and text sentiments [Zhao et al., 2019]. Existing research and the newest studies show that new research should explore ways to combine image and text data to predict user sentiment.

2.2.2. Image-based Social Media Models Many studies use Convolutional Neural Networks (CNN) for image analysis. The use cases are varied, and include: age and gender classification [Hassner and Tal, 2015]; image polarity [Poria et al., 2016]; sarcasm detection [Poria et al., 2016]; and image popularity classification [Khosla et al., 2014]. Existing research has produced visual sentiment classifiers with CNNs [Segalin et al., 2017, Xu et al., 2014] to identify stress within social media images, [Lin et al., 2014], use supervised CNNs to performed social profiling to identify personality traits [Segalin et al., 2017], perform sentiment analyses and estimated social media popularity with CNNs [Gelli et al., 2015], use images to predict which types of images are popular on social media [Gelli et al., 2015], and predict which posts will receive the most clicks [Khosla et al., 2014]. CNNs are frequently used in combination with images on social media for understanding user behavior.

3. Methods

The method section describes the data context, collection, and steps for processing, obtaining, and evaluating results by this study. First we describe the data collection from Facebook and Hootsuite. The second section will describe the data processing, including the cleaning of text and image data necessary for widespread usage of such a process. The next section handles the creation of machine learning input vectors. For example, text data is processed via nlp, converted into sentence vectors, and transformed into sentiment. Other integer engagement data is converted into vectors for machine learning models. The latter section of the methodology will concerns the machine learning models, their architectures, and the method for their evaluation. This section will provide details on the data collection, processing steps, and both the decision and their justifications for this research.

3.1. Data Collection

The research team focuses on studies Facebook posts on company brand pages and the resulting user response. These brand pages consist of Firm-Generated Content that users can interact with. Brand pages are public company pages where brand-content is posted in order to influence user behavior. Often this content and page have the purpose of building the company's brand. Given the public nature of these pages, they serve as easily accessible FGC content. As a plus, Facebook allows easy access to public Facebook pages. The study looked at public brand pages and examined the posts and the resulting user response.

We obtain a set of brand URL links to Facebook from AdEspresso. Hootsuite owns the website and is a social media management platform created in 2008 that allows advertisers to manage advertising from multiple social media sites from the app (https://www.hootsuite.com). The website features over 100,000 demo Facebook advertisements that have used Hootsuite to advertise from which we scraped 281,090 links to

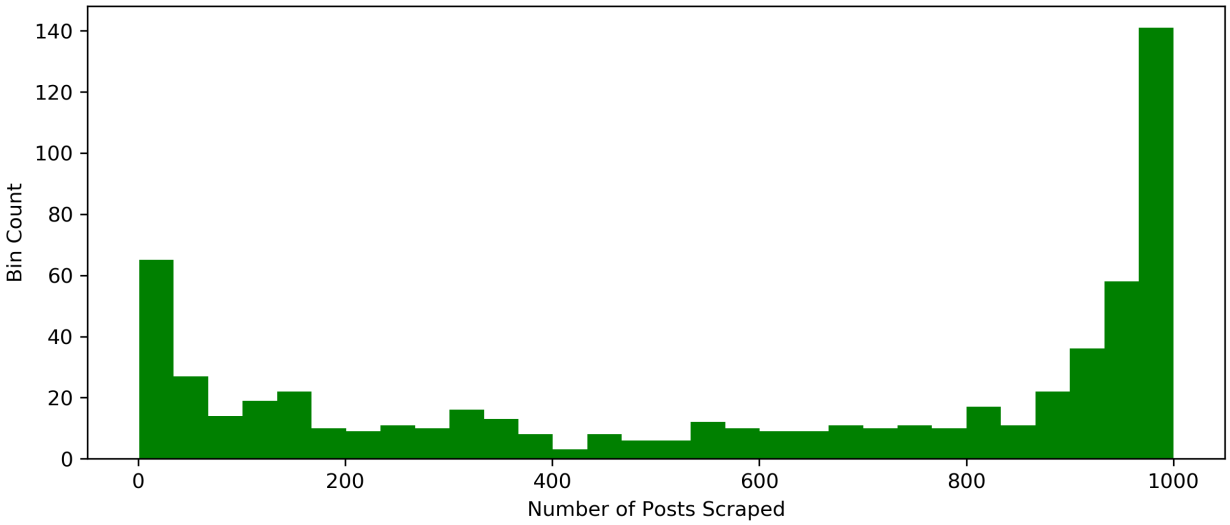

Figure 1 Histogram of the Number of Posts Scraped Per Facebook Page

advertiser pages, as linked on this website. After obtaining advertiser pages on Facebook, we scraped user engagement data with a Python web scraper to crawl the Facebook brand pages. Facebook provides the Graph API, a publicly-available API, for accessing Facebook data. The crawler scrapes the most recent 1,000 posts from each brand page. The histogram for the number of posts scraped per brand page are shown in Figure 1. Thousands of posts is a sufficient number of posts to capture a sample of the Facebook page's posts. For context, [Luarn et al., 2015] only collects 1,030 posts for its study. Another Facebook study only collects 28,595 posts [Settanni and Marengo, 2015]. In total, the study collected 366,415 Facebook posts and 1,305,375 million comments. An abundant amount of posts and comment data are collected by this study that provide a good large range and sample of advertisements on Facebook.

After obtaining advertiser pages on Facebook, we scraped user engagement data from each Facebook page. We use a Python web scraper to crawl the Facebook brand pages. Facebook provides the Graph API, a publicly-available API, for accessing Facebook data. The crawler scrapes the most recent one-thousand posts from each brand page. While

some pages contain more than one-thousand posts, we felt that one-thousand posts per advertiser is sufficient data per brand page. The histogram for the number of posts scraped per brand page are shown in Figure 1. In total, the study collected 366,415 Facebook posts and 1,305,375 million comments. This study collects a great deal of brand-page data via more than 350k Facebook posts and 1.3 million comments.

3.2. Data Quality

The study ensures data quality by relying on existing, third-party resources specializing who specialize in both sourcing and providing data and includes both manual and automatic checks to ensure its analysis is representative of Facebook brand pages. For example, Hootsuite is a company that specializes with advertising on social media across multiple platforms. By collecting brand page URLs from Hootsuite, the study ensures its data comes from advertising pages. Another example of sourcing data is the use of Facebook's API to collect post data. Given a public Facebook page, the Facebook Graph API allows apps to pull Facebook data. The extracted data is performed by the platform itself, ensuring accuracy, and is provided in an easy-to-digest manner, aiding with data processing. The study utilizes existing businesses and resources for compiling this study's data.

The research team also performed manual checks to verify the correctness of both Facebook brand pages and their post statistics. The study reviewed many Facebook brand pages to ensure the pages are brand pages. On Facebook, many brand pages have alternative URLs, where instead of using their brand page id, Facebook allows them to use the name of their brand for the URL. For string URLs, the research team checked to ensure that a reasonable number of those strings represent a business on Google via making requests with Python. For example, Nike's Facebook page is at https://www.facebook.com/nike, and the study can verify that nike.com is indeed a valid website.

The study uses a third-party tools specialized in social media text data for generating text sentiment, the Python library VADER. Text sentiment rates the valence of the text towards either negative or positive and machine learning tools are commonly used for generating text sentiments Haddi et al. (2013). VADER is a parsimonious, rule-based model for Sentiment Analysis for social media text data Gilbert. VADER's output is a score between -1 and 1 to denote the text's positive or negativity, where 1 denotes a positive sentiment, 0 is neutral, and -1 denotes a negative sentiment. The machine learning models in this study predict the average user sentiment for the Facebook post, as scored by VADER. The post's user sentiment is the average sentiment of all user comments. This output serves as the target value for the machine learning models.

For modeling, the study uses standard deep-learning architectures to train its NN and CNNs. These architectures are freely available to import and train via Keras, a Python machine learning library. The CNN model uses VGG16 (https://keras.io/api/applications/) and the NN is a generic Keras, deep NN with seven hidden layers. The The use of existing performant NN and CNN architectures gives us confidence the deep learning models will satisfactorily learn the data. Moreover, future research can also train with these models for extending our findings.

3.3. Data Processing

This study follows standard text processing steps [Camacho-Collados and Pilehvar, 2019], including text cleaning, creating tokens, using a port stemmer, part-of-speech (POS) tagging, lemmatization, and transforming text sentences into a td-IDF vectorizer. First, the program splits data into word tokens using whitespace as a delimiter. Next, the program grouped these tokens into sentences, lowercases words, and removes common English stopwords, as well as words with three or fewer characters. The remaining words are fed into

a port stemmer, which creates word stems. These stems serve as input into a POS tagger, which provides details like noun and verb declensions. Such word details are useful in determining the sentences structure and contents. The program then extracted stems with a word lemmatizer, which takes the stem and the POS tag as input. Finally, the program fed the lemmatized text sentences to a td-IDF vectorizer. The vectorizer creates word vectors, which serve as input for the neural networks. Each of the many steps to process text data to ensure the data is informative and easily digestable for the neural network.

In a similar fashion, the study takes multiple, iterative steps to prepare images. Given that images are inherently highly dimensional, the first step is to reduce each images' dimensionality. First, we used principal component analysis and reduce the number of image dimensions to 20. We experimented with dimensions from 10 to 40 but had good initial model performance with 20 dimensions. This preserves the image directions that contain the most data. Afterwards, the image is further processed via denoising in order to reduce both noise and the size of the data. In order to emphasize edges, the study then applies Gaussian blurring, with a standard deviation of five to each image, which is the default in Python's computer vision library, which both applies image blurring and emphasizes edges. Dilation and erosion are also applied to the image to remove noise in the edge space. The final result are highly filtered images that emphasize shape and edges. An example of such an image is displayed in figure 2

The sentiment dataset consists of 201,215 posts. These posts contain comment count, share count, and comment sentiment data. Of those posts, the study collected and trained on 1,305,375 million user comments. The study rates the sentiment of user response toward a Facebook post as the average sentiment of its comments. The post text denotes the text created by the brand that is present in the post. The comment text includes all comments

Figure 2 Example Image after Preprocessing

made on the FGC post. Comment sentiment is averaged for each post and represents the user sentiment toward the post. Each text is scored from -1 to 1 for its positivity where -1 is negative, 0 is neural, and 1 is a very positive sentiment. Figure-5 shows a histogram of comment sentiments for the Facebook posts.

We also consider brand page variables and how it might affect user engagement. Specifically, the study explores whether general brand page characteristics are correlated with

user engagement. For example, do more popular brand pages also receive more engagement? The study documents these correlations and notes how they might affect the results. We find that two brand page variables had correlations with user engagement: comment sentiment (0.44 adjusted R-squared) and the number of users talking about a brand page (0.38 adjusted R-squared). Unsurprisingly, these two variables are highly correlated with one another with a 0.38 adjusted R-squared. Concerning user engagement, the variable most correlated with user engagement is 'talking about count.' When users talk about a particular brand, they are more likely to share its posts (0.22 adjusted R-squared). The highest correlation is between comment count and comment sentiment, which indicates that posts which elicit positive user comments also receive a greater number of comments.

3.4. Modeling

There are three types of models, text, image, and combined models. The text model uses a neural network architecture with seven hidden layers and the CNN uses the VGG16 architecture. There are two types of combined models, a decision-tree based model and an ensemble of the NN and CNN models. There are three training datasets for share count, comment, count, and comment sentiment. Training each of the four architectures on the three data types makes for a total of twelve models. This study trains four model architectures, based on different data types.

The text-based NN are quite simple. They consist of NN with two hidden layers. Each layer halves the number of layers and uses dropout, ranging from 0.2 to 0.3 to aid with regularization. The input dimension to the NN is 512 and the output is a single node. The model operates a large input, and therefore uses a batch size, both to quickly process the data and ensure it moves toward more performance weights more quickly. The model is set to have 100 epochs but is also set to stop training as soon as the test model performance

becomes worse, i.e., begins overtraining. We initially experiment with word vector sizes from 1k-400k. Good performance and fast training for the NN text-based models occurred with a word vector size of 10k. The NN models used 10k as the word vector size. The model uses the adam optimizer and outputs a numeric value. The numeric value corresponds to the model's prediction of the user engagement metric. For example, the comment count model's output represents how many comment counts a post will receive. The sentiment model predicts the average comment sentiment of users predicting on a post. The share count NN output predicts the number of times the post will be shared. The text-based NN model is simple and can be easily produced for replication and future research.

The convolutional neural networks explored both existing architectures and simple CNNs with many convolutional and max pooling layers. The network used existing architectures for ease of replicability, and because those models have proved themselves as performant for many types of image types. However, these existing architectures are very large and take a long time to train. In contrast, the study also uses five repeated combinations of 2D convolutions and max pooling. This model is faster to train and optimize with hyperparameters. The study uses the VG16 model, which is easily imported and used in Keras `https://keras.io/api/applications/vgg/`. We use a 2x2 pool size and relu activations on each layer. The model uses the adam optimizer with a momentum of 0.0001. Both models produce a single numerical output that represents the model's prediction for user engagement. Both CNN architectures are easily replicated and perform well for processing images.

The model also trained a decision tree on image and text vector data. The decision tree regressor trains with cross validation. The model learns using the squared error and chooses the best split at each node. We train the decision tree with text data, image data, and

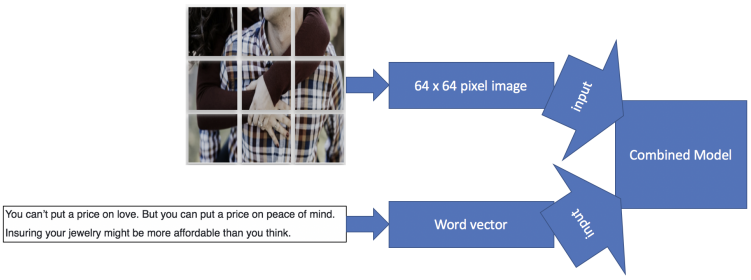

Figure 3 Combined Model Inputs

an ensemble model that takes the output of the text-based decision tree and image-based decision tree.

The second combined model is an ensemble combined model that concatenates both text-based NN and image-based CNN. This model did not take very long to train since it reused the individual CNN and NN models that were already trained on their respective data type. The model operates on the output of each individual model, e.g., the text-based NN produces a single regressor output that is fed into the ensemble model. The ensemble model takes one input from the text-based NN and one from the image-based CNN. Each model is trained with k-fold cross-validation, its use, and methods for training and testing are publicly available on Github at `https://github.com/cpluspluscrowe/Marketing_Science`. The study makes use of common or default hyperparameters for training the readily available and imported architectures.

Regression is used to optimize the models since the data consists of numerical data. For example, comment sentiment ranges from -1 to 1, and both share and comment count vary from zero to millions. The study uses Mean Squared Error (MSE) as its loss function for the machine learning models, which heavily penalizes values straying from the observation since the difference between the target and prediction is squared. The resulting models have a smaller prediction range but with more accurate predictions of which posts which outperform others.

Metrics / Model	Text-Based NN	Image-Based CNN	Combined Decision Tree	Combined NN
Share Count	3.44	1.01	2.58	1.00
Comment Count	1.02	1.01	1.29	1.00
Comment Sentiment	1.42	1.14	4.20	1.00

Table 1 Model Mean Squared Error Reported as a Ratio to the Best Model's Performance

4. Results

The results display the MSE for the twelve ml-models. The twelve models use a combination of different data types as input, including only-text, only-video, and a combination of text and video data. The relative MSE for the twelve models is shown in Table 1. The combined models include a decision tree and an ensemble CNN and NN. The study provides relative MSE performance across both different architectures and combinations of input data types. Rather than provided raw MSE across all models, we provide ratios. We only provide ratios since the resulting MSE is large due to viral posts that heavily skew the MSE, so the resulting number is difficult to interpret. Instead, a ratio between each model provides feedback on the relative model performance. The histograms of MSE for each model is shown in the Figures 4, 5, and 6. We also provide a histogram of the MSE to provide an understanding of how well the model performs on the test datasets.

The ensemble combined model best predicted all user behavior metrics. Each machine learning model had its lowest MSE predicting for share count. The CNNs achieved a lower MSE than the NN on all metrics. Moreover, the CNN performance is best on data exhibiting a higher variance. Figure 5 and 6 show the predicted vs actual distribution for the combined model.

The decision tree performed far worse than the ensemble combined model for all user engagement types. While the exact reason is unknown, there are a few differences between

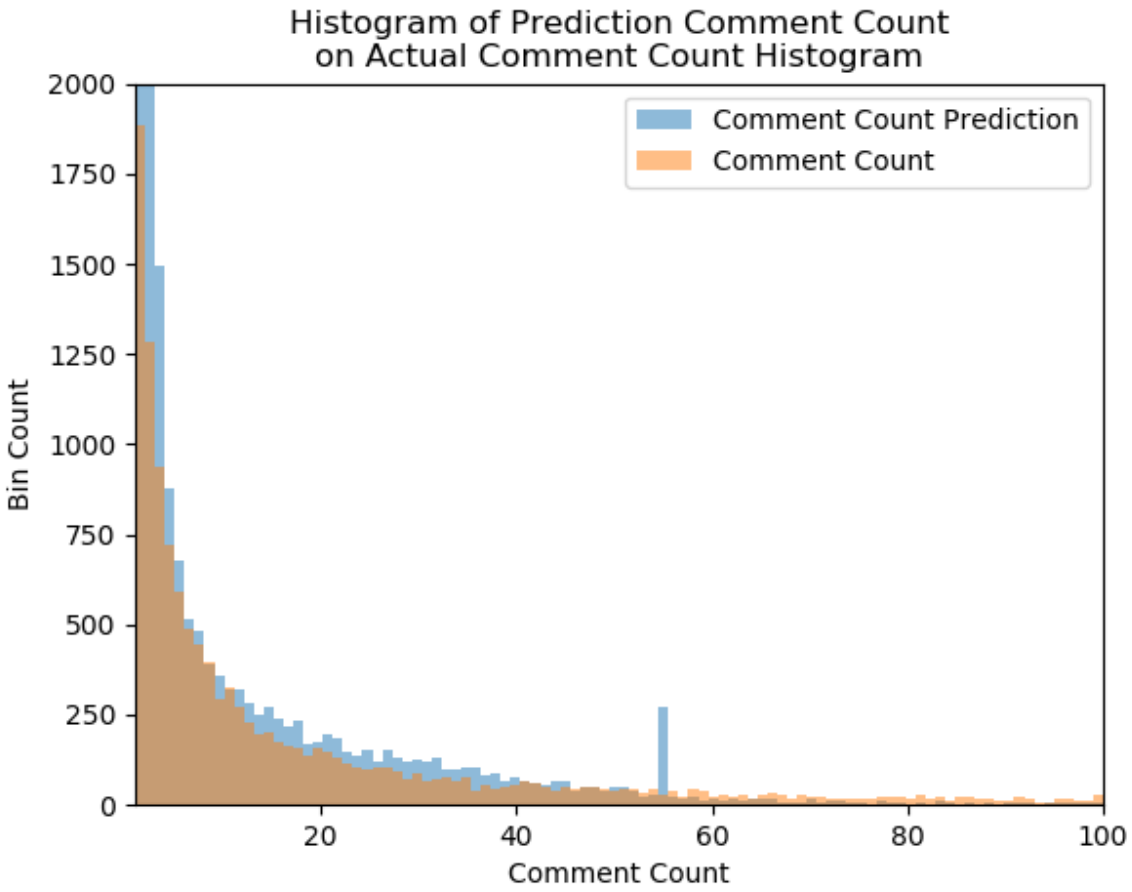

Figure 4 **Actual vs Predicted Comment Count Histogram**

the two models. First, the combined model closely integrates with the text-based NN and image-based CNN. Second, when the combined model trains, it also trains its two-parent models. As a result, the decision tree did not learn alongside its inputs. Likely, parent models compensate for mistakes by training together.

Popular pages do not generate more comments per post. This might capture a phenomenon where brands receive many followers but produce less engaging content. Therefore generating many followers does not necessitate an engaged user base. On social media, a large page also needs to produce quality content in order to generate user engagement.

Correlations exist in the occurrence of some user engagement metrics. Principally, there is a general correlation between user comment sentiment and other metrics, specifically

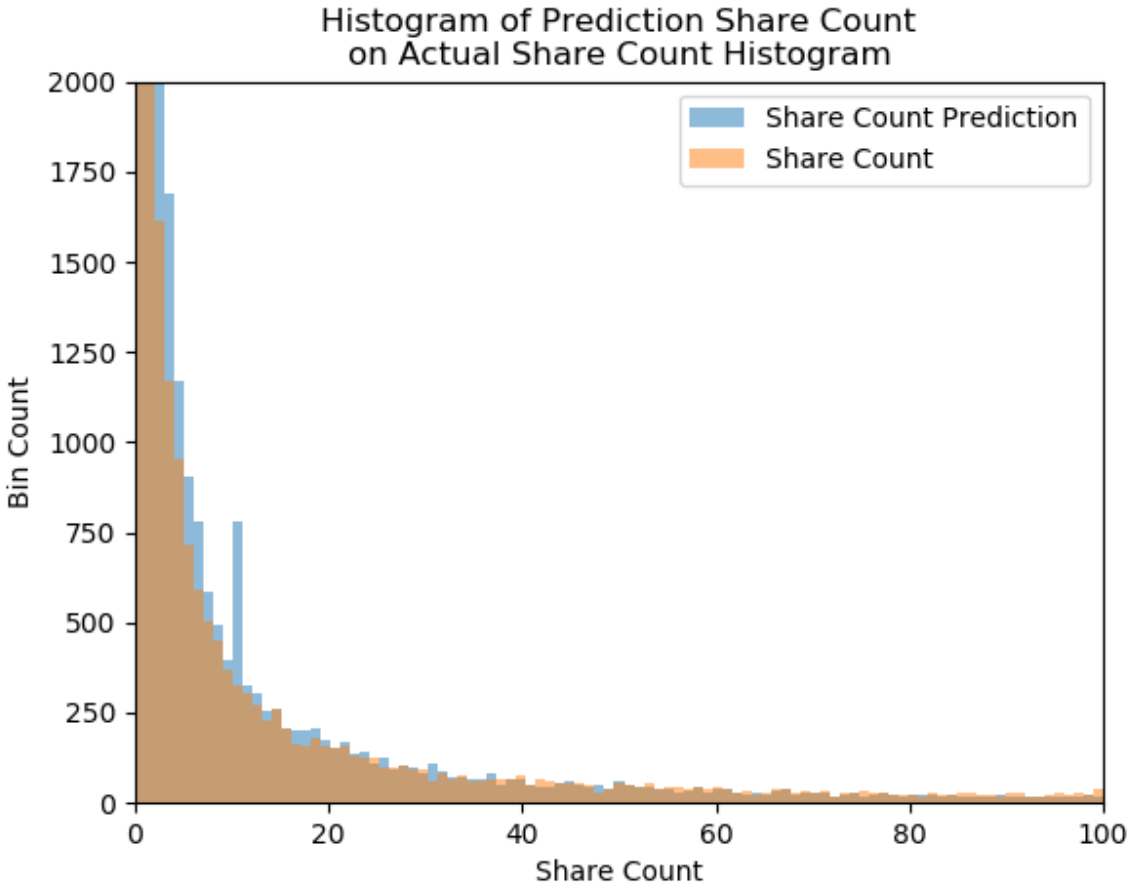

Figure 5 **Actual vs Predicted Share Count Histogram**

comment count and share count. However, the correlation between comment and share count is much weaker. The implication is that posts that elicit positive comments will also receive a greater number of comments and shares.

The models perform the best on distributions that skew toward fewer counts, such as comment and share counts. These are distributions that resemble the geometric distribution; those that heavily decrease for higher counts. The actual vs predicted histograms look very similar and how lower model losses, as shown in Figure 5 and Figure 4. As can be seen from the figures, the models do especially well at predicting share count and comment count on Facebook brand posts.

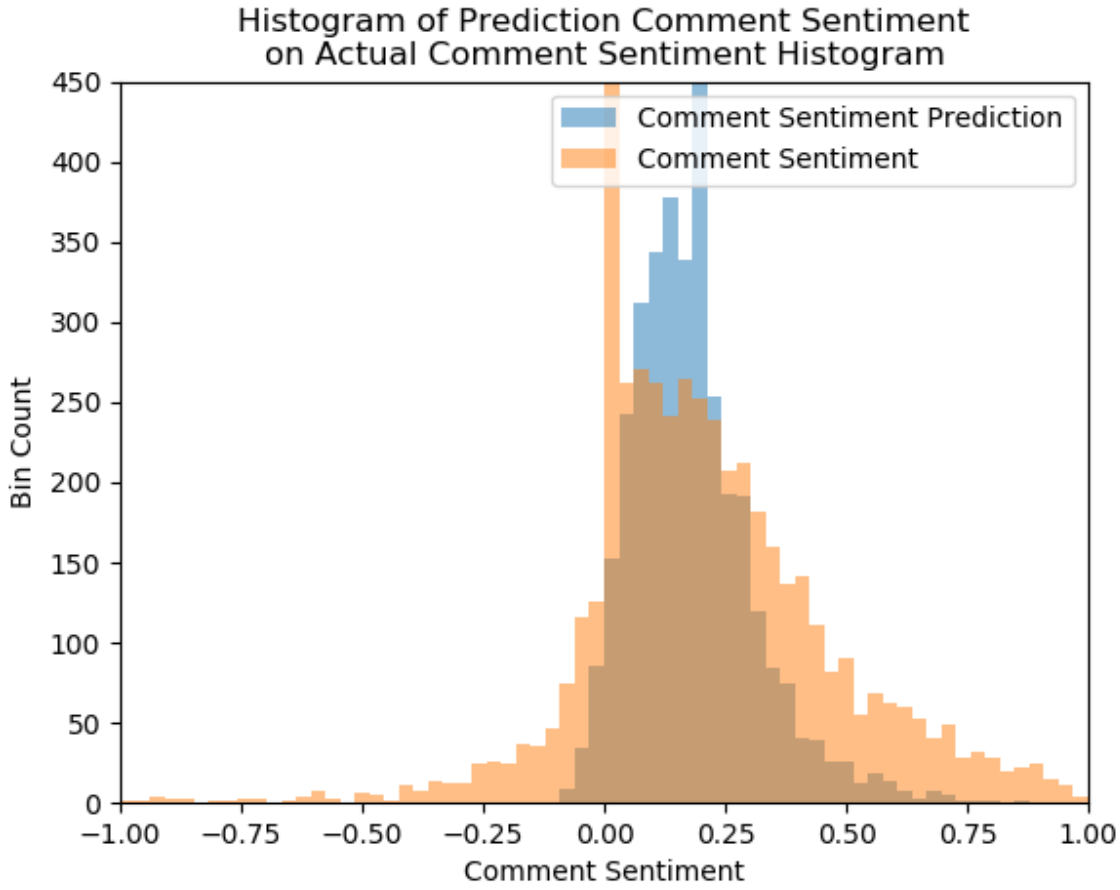

Figure 6 Actual vs Predicted Comment Sentiment Histogram

The model achieves a quite low model loss for predicting user sentiment for posts. This is largely possible due to the close relationship between post sentiment and comment sentiment. The resulting actual vs predicted comment sentiment is shown in Figure 6. The study successfully predicts comment sentiment for Facebook brand posts.

4.1. Performance of Images vs Text Models

Image-based models outperform text-based models the most when predicting share count. The share count metric resembles the long tail of the geometric distribution where most posts have very few shares, similar to comment count. However, share count has a higher variance than comment count. The inference is that image-based models did better at handling the variance within the metric share count. It is worth noting that even a sim-

ple decision tree with image data outperforms a deep neural network with text data for predicting share count. Image data seems especially useful for producing shares, which is reflected in the better performance of image-based models. Image-based models show a greater ability at handing distributions with more variance.

Both text-based NN and image-based CNN perform almost equally well for predicting comment counts, i.e., they achieve a similar MSE. Both models receive the same input in terms of word vectors and both models are able to process this data. The similar performance might mean that both models similarly process the data or tend to find the same patterns in the existing data. Both text-based and image-based models perform similarly for predicting comment counts.

The combined decision tree performed poorly compared to all other models on almost all metrics. The combined decision tree's poor performance indicates the complexity of the input data, both word and image vectors. A decision tree is likely a poor candidate for handling more complex inputs, however, one might expect it would be able to find patterns in comment sentiment, which has more simple inputs based on parts of speech. Nevertheless, the decision tree even performed poorly when predicting comment sentiment. Based on this we do not recommend a combined decision tree as it is not performance in learning the patterns within the provided word and image vectors for predicting user engagements.

Image-based models performed better than text-based models on all metrics, likely because it emphasizes images' importance for predicting user interaction on social media. Text-based models were poor predictors for many metrics. Text-based models did especially badly in prediction share counts. One may guess that users are sharing content they consider worthy of sharing.

4.2. Use Case

The research team applied the models in a real-world application to demonstrate the model's ability to choose which two advertisements will receive the greatest user engagement. The research found that models are unlikely able to differentiate between advertisements with similar performances. However, we are able to consistently differentiate between ads with larger difference. We show that the model can differentiate a difference in ad performance for ads with at least a standard deviation difference in user engagement. The combined model performs all predictions since it performed best across all metrics. Moreover, the application is the most useful if it can detect poor-performing and best-performing advertisements.

The study predicts which of two advertisements will receive greater user engagement. Given two advertisements, the model predicts user engagement for both posts. We then check whether the model is correct at choosing which advertisement will have more engagement. A correct prediction is an instance where the model correctly predicts which advertisement received a greater amount of user engagement for the particular metric. As a baseline, a random model correctly predicts the higher performing advertisement 50% of the time. For the results, the combined model scored 93% for comment sentiment, 65% for comment count, and 63% for share count. The prediction model is able to consistently predict which of two advertisements will receive the most positive user comments.

In light of the newly created data, there are no other baseline measures for what constitutes good model performance. One goal was to produce a model that performed better than random guessing. Random guessing alone is not representative of the data. A better guess is based on the input's distribution. A good random guess would consider the data's value at each point along with its distribution. Such a distribution would weigh each value

by the frequency of data. This calculation is the expected value. For a normal distribution, the expected value is equal to the mean. If the model always predicts the mean, the MSE is equal to the data's variance. The variance is a squared order of the data's distance from the mean. The MSE is also a squared order of the model's prediction from the actual value. Producing a model whose loss is below the variance is a general measure of demonstrating the model is doing a significant amount of learning. Fortunately, the combined model's performance was always much lower than each output data's variance.

Models with a higher variance achieved the worst overall loss/variance ratio. It seems that the larger the data variance, the higher the resulting mode's MSE. Share count had the highest variance of all the measured metrics, 1000x more than the comment count. The high MSE likely reflects that share counts are less related to the image and text data. Share counts might be a factor of other features, like page popularity.

5. Discussion

One consideration might be that both NN and CNN show a similar performance on the data but major improvements can be made in terms of share count and comment sentiment in the use of a model that combined both image and text data. The 14% improvement for comment sentiment and 3.5x improvement on share count can be made across either the NN or CNN. The combined model demonstrates dramatic improvement on at least one metric for each of the single-data type models.

The use case showed the models' ability to predict for greater user engagement between any two advertisements based on their text and image data. Platforms could employ the models to aid advertisers in choosing the best performing advertisement before paying to advertise it on social media. The model can tell advertisers which ads would perform best on the platform, which allows advertisers have their ads vetted. The vetting could

prevent advertisers from spending large amounts of money showing worse ads. Moreover, the vetting would allow advertisers to only show ads that will perform best. In the context of billions, 93%, 65%, and 63% accuracy is a substantial monetary difference.

6. Conclusion

A gap exists in current user behavior modeling research with social media data as researchers have not yet capitalized upon multiple data types, despite the availability of both image and text data. This research provides a methodology for predicting user behavior in response to advertising. The paper concludes with a use case that demonstrates its models ability to reliably choose better performing advertisements. The research found that machine learning with both image and text data results in large improvements for predicting comment sentiment. The combined model showed a 14% increase in predicting user combined sentiment vs. the text-based NN and image- based CNN. The implication is that a combination of image and text data is best when predicting how users will respond to an advertisement. The study found significant gains in the performance of a model combining image and text data when predicting user sentiment. The study found that text-based models can be improved by 42% for comment sentiment and a factor of 3.5x for share count. Both the image and combined models far outperformed text-only models for comment sentiment and share count. Text models that ignore image post data are likely missing large opportunities for improved model performance from incorporating image data via a CNN.

The study demonstrates an accuracy of 93%, 65%, and 63% for comment sentiment, comment count, and share count when predicting which of two advertisements will receive a greater amount of user engagement. This provides an application of the research on the combined model for advertisements. Advertisers can expect tangible gains in predicting

user engagement when utilizing ml models that predict user engagements based on post text and image data.

Many studies only use images or text; this study provides both an insight and a methodology for how one might go about improving their model's performance by combining the two. The study the use of an ensemble model of NN and CNN achieved an improved performance across all metrics. The study also provides its architectures, hyperparameters, and code examples for the models. This study demonstrates the degree of improvement that a combined model can achieve for predicting user engagement on advertisements.

6.1. Limitations

The study is not able to show the improved benefit of the combined model on pre-existing study data or models, however, the study makes use of well-known architectures to ensure model performance and study replicability. The study is not able to acquire existing ml models and study data concerning single-data type models to thoroughly benchmark against. The data and models are not for public use. The result is that this study made use of existing, high-performing model architectures for both CNN and NN. This allows the study to ensure its results are replicable. Using pre-existing model architectures also ensures that existing research's best practices for ml models and architectures are included within this study.

6.2. Future Work

The models created from this research could generate data to train a generative model. The generative model could transform images and text into advertisements that should generate more user interaction. The transformed and original advertisements could both be shown on social media and their user interactions compared. Thus, the study might demonstrate that generative models generate and improve existing advertising

References

[Aggarwal and Gupta, 2017] Aggarwal, R. and Gupta, L. (2017). A hybrid approach for sentiment analysis using classification algorithm.

[Barreto, 2013] Barreto, A. M. (2013). Do users look at banner ads on Facebook?: An International Journal. *Journal of Research in Interactive Marketing*, 7(2).

[Bhat et al., 2002] Bhat, S., Bevans, M., and Sengupta, S. (2002). Measuring users' web activity to evaluate and enhance advertising effectiveness. *Journal of Advertising*, 31(3):97–106.

[Camacho-Collados and Pilehvar, 2019] Camacho-Collados, J. and Pilehvar, M. T. (2019). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. pages 40–46.

[Chen and Dredze, 2018] Chen, T. and Dredze, M. (2018). Vaccine Images on Twitter: Analysis of What Images are Shared. *J Med Internet Res*, 20(4):e130.

[Clark and Melancon, 2013] Clark, M. and Melancon, J. (2013). The influence of social media investment on relational outcomes: A relationship marketing perspective. *International Journal of Marketing Studies*, 5:132.

[Fisher, 2009] Fisher, T. (2009). Roi in social media: A look at the arguments. *Journal of Database Marketing and Customer Strategy Management*, 16(3):189–195.

[Gelli et al., 2015] Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A., and Chang, S.-F. (2015). Image Popularity Prediction in Social Media Using Sentiment and Context Features. *Proceedings of the 23rd ACM international conference on Multimedia - MM '15*, pages 907–910.

[Georgiou et al., 2015] Georgiou, D., MacFarlane, A., and Russell-Rose, T. (2015). Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. *Proceedings of the 2015 Science and Information Conference, SAI 2015*, pages 352–361.

[Greenwood et al., 2016] Greenwood, S., Perrin, A., and Duggan, M. (2016). Social Media Update 2016.

[Guo et al., 2020] Guo, L., Lu, R., Zhang, H., Jin, J., Zheng, Z., Wu, F., Li, J., Xu, H., Li, H., Lu, W., Xu, J., and Gai, K. (2020). A Deep Prediction Network for Understanding Advertiser Intent and Satisfaction. In *International Conference on Information and Knowledge Management, Proceedings*, pages 2501–2508.

[Hassner and Tal, 2015] Hassner, G. L. and Tal (2015). Age and Gender Classification using Convolutional Neural Networks. *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*, 24(3):2622–2629.

[Hu et al., 2016] Hu, Y., Shin, J., and Tang, Z. (2016). Incentive problems in performance-based online advertising pricing: Cost per click vs. cost per action. *Management Science*, 62(7):2022–2038.

[Hudson et al., 2020] Hudson, N., Khamfroush, H., Harrison, B., and Craig, A. (2020). Smart Advertisement for Maximal Clicks in Online Social Networks without User Data. In *Proceedings - 2020 IEEE International Conference on Smart Computing, SMARTCOMP 2020*.

[Imsa and Irwansyah, 2020] Imsa, M. A. and Irwansyah (2020). Online Advertising Effectiveness for Advertiser and User. 459(Jcc):216–221.

[Khosla et al., 2014] Khosla, A., Das Sarma, A., and Hamid, R. (2014). What Makes an Image Popular ? *Proceedings of the 23rd International Conference on World Wide Web*, pages 867—-876.

[Li et al., 2015] Li, C., Lu, Y., Mei, Q., Wang, D., and Pandey, S. (2015). Click-through Prediction for Advertising in Twitter Timeline. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 1959–1968.

[Lin et al., 2014] Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Cai, L., and Feng, L. (2014). User-level psychological stress detection from social media using deep neural network. *Proceedings of the ACM International Conference on Multimedia - MM '14*, pages 507–516.

[Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. (May):1–108.

[Lotze et al., 2021] Lotze, T., Klut, S., Aliannejadi, M., and Kanoulas, E. (2021). Ranking clarifying questions based on predicted user engagement. *ArXiv*, abs/2103.06192.

[Luarn et al., 2015] Luarn, P., Lin, Y. F., and Chiu, Y. P. (2015). Influence of Facebook brand-page posts on online engagement. *Online Information Review*, 39(4):505–519.

[Nash et al., 2017] Nash, E. L., Gilroy, D., Srikusalanukul, W., Abhayaratna, W. P., Stanton, T., Mitchell, G., Stowasser, M., and Sharman, J. E. (2017). Facebook advertising for participant recruitment into a blood pressure clinical trial. *Journal of Hypertension*, 35(12).

[Ohsawa and Matsuo, 2013] Ohsawa, S. and Matsuo, Y. (2013). Like Prediction: Modeling Like Counts by Bridging Facebook Pages with Linked Data. *Proceedings of the 22Nd International Conference on World Wide Web Companion*, pages 541–548.

[Poria et al., 2016] Poria, S., Cambria, E., Hazarika, D., and Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks.

[Romero, 2011] Romero, N. L. (2011). ROI. Measuring the social media return on investment in a library. *Bottom Line*, 24(2).

[Schröter et al., 2021] Schröter, J. M., Dutzi, A., and Withanage, E. (2021). Can firm performance and corporate reputation be improved by communicating csr in social media? *International Journal of Applied Management Sciences and Engineering*.

[Segalin et al., 2017] Segalin, C., Cheng, D. S., and Cristani, M. (2017). Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding*, 156:34–50.

[Settanni and Marengo, 2015] Settanni, M. and Marengo, D. (2015). Sharing feelings online: studying emotional well-being via automated text analysis of facebook posts. *Frontiers in Psychology*, 6.

[Somerfield et al., 2018] Somerfield, K., Mortimer, K., and Evans, G. (2018). The relevance of images in user-generated content: A mixed method study of when, and why, major brands retweet. *International Journal of Internet Marketing and Advertising*, 12(4):340–357.

[Straton et al., 2017] Straton, N., Mukkamala, R. R., and Vatrapu, R. (2017). Big social data analytics for public health: Predicting facebook post performance using artificial neural networks and deep learning. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 89–96.

[Tiago and Verissimo, 2014] Tiago, M. T. P. M. B. and Verissimo, J. M. C. (2014). Digital marketing and social media: Why bother? *Business Horizons*, 57(6):703–708.

[Wagner et al., 2015] Wagner, T. F., Baccarella, C., and Voigt, K. (2015). Antecedents of Brand Post Popularity in Facebook: The Influence of Images, Videos, and Text. *Cognition the Arts eJournal*.

[Wang et al., 2015] Wang, Y., Wang, S., Tang, J., Liu, H., and Li, B. (2015). Unsupervised sentiment analysis for social media images. *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2378–2379.

[Xing et al., 2021] Xing, B., Si, H., Chen, J., Ye, M., and Shi, L. (2021). Computational model for predicting user aesthetic preference for gui using dcnns. *CCF Trans. Pervasive Comput. Interact.*, 3:147–169.

[Xu et al., 2014] Xu, C., Cetintas, S., Lee, K.-C., and Li, L.-J. (2014). Visual sentiment prediction with deep convolutional neural networks.

[Zhao et al., 2019] Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., and Liu, M. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing and Management*, 56(6).