SMU Data Science Review

Volume 7 | Number 1 Article 2

Professor Text: University Fundraising Optimization

Braden Anderson
Southern Methodist University, bradenandersonee@gmail.com

Connor Dobbs
Southern Methodist University, dobbs@smu.edu

Hien Lam
Southern Methodist University, hien.lam09@gmail.com

John Santerre
Southern Methodist University, john.santerre@gmail.com

Professor Text: University Fundraising Optimization

Braden Anderson[1], Connor Dobbs[1], Hien Lam[1], John Santerre[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Marketing and Communications Department, 6425 Boaz Lane,
Dallas, TX 75205 USA
{bradena, dobbs, hienl}@smu.edu
jsanterre@smu.edu

Abstract. University fundraising campaigns are a unique type of cause-related marketing with its own challenges and opportunities. Campaigns like this typically last an extended period, such as five or more years, and goals exist beyond the dollar amount raised. These supplemental goals, such as awareness among potential future donators or brand reputation within the local community, are important to consider and strategize. There can also be unique limitations, such as requiring advertising specifically on recent large gifts or endowment programs. This research explores how machine learning techniques such as natural language processing can be used to optimize a fundraising campaign strategy, execution, and overall performance.

1 Introduction

Educational institutions worldwide have traditionally relied on primarily government funding and tuition payments to support faculty, facilities, and research. While these primarily public funded universities are still common, particularly outside the United States, private funded nonprofit universities are increasingly popular. 27% of total undergraduate and graduate students are enrolled in a private university, although those private universities make up approximately 59% of universities (Bureau, 2022). Fundraising is becoming more important even for public universities as state funding has decreased over time. In 2017, the total funding was $9 Billion below pre-recession 2008 levels, according to the Center on Budget and Policy Priorities (Mitchell et al., 2017). For private universities funding is vital in allowing them to compete with more publicly funded institutions. Private schools often must charge higher tuition, to begin with, but private funding helps reduce this gap and provide more opportunities for the school to provide scholarship opportunities.

According to personal communications with Neil Robinson, an Executive Director for Southern Methodist Universities Planning and Campaign Marketing team, private university fundraising typically takes the form of one of three gifts. The first is money donated towards a specific capital project. These include instructional buildings, living spaces, research facilities, and/or sports facilities. The money given goes towards those projects chosen by the donor. (Depending on the

amount donated, can have the donators name on a plaque in the facility, or even an entire building named after them). The second type of gift is towards an endowment, a considerable sum of money that the university invests. The returns from the investments then go back towards the university, allowing a gift to exist in perpetuity. Lastly, donors can direct the funds toward purchasing something more immediate, such as lab equipment.

Many of the common fundraising approaches exist as options for universities. These include direct solicitations through e-mail, mail, phone, or door-knocking. Another method is hosting fundraising events or partnerships with companies that may have a mutual interest. In traditional fundraising, these organizations are often for a specific cause, such as funding a boy scout troop, campaigning for a politician, or raising money to clean up oil spills in the gulf. Universities are broader in scope, presenting a layer of complexity and strategic options in focusing efforts. A more precise solicitation could be donations directed to the university, endowment programs, or direct-use gifts. These gifts could behoove components of the university e.g., athletics programs, research facilities, general upkeep of the campus, and more. This large scope is where the differences between typical nonprofit fundraising and private university fundraising begin.

Audience targeting also comes with its complexities for universities. While a common cause nonprofit, such as one raising money for cancer research, has a broad geographical audience to target, this is not typically the case for a university. Instead, an alumni network presents itself as a strong initial fundraising base, as well as the local community that is invested in the success of an institution that has a strong impact on the shape of the community. It is often challenging to extend the fundraising base past those two groups, limiting the number of people that may act. This increases the importance of brand awareness and sentiment among alumni and the local community.

Messaging content can also be restricted in some circumstances during university fundraising. For example, when certain types of gifts are given, they bring a requirement to advertise and communicate the news about the gifted funds. The cost of this advertising is not always a part of the gift funds itself and instead uses a budget for the fundraising campaign. This can limit the amount of creative and strategic control around what type of content is in paid advertisements.

While the most obvious goal of any fundraising campaign is to raise money, other objectives are just as important. One such example is creating stronger relationships with people that may not have much money to donate currently but may be able to donate larger amounts in the future. This may include recent graduates and individuals who may come across a large amount of immediate wealth (i.e., inheritance). Another common objective is to increase reputation in the local community by partnering with other local nonprofits and organizations that foster relationships with a wide range of people. For example, a partnership with a local elementary school provides short-term benefits with improved awareness and reputation. Still, those children who benefit may be more likely to attend the university in the future.

Southern Methodist University (SMU), a private nonprofit university, is the subject of this research. The university's Marketing and Communications department (MARCOM) helps manage the current university fundraising campaign, *Ignited.*

The current stated goal of the campaign, which started in September 2021, is to raise $1.5 Billion by 2028. Neil Robinson, an executive director within the MARCOM department, stated three objectives these funds would be used to accomplish:

1. Empower and attract outstanding students by providing additional scholarships and financial aid
2. Enrich teaching and research at SMU by providing funds for increased research activities and endowment for faculty
3. Enhance the campus and community as it relates to athletics, clubs, campus communities, events, and partnerships with companies and school districts

This research aims to provide a recommendation and evaluation system utilizing an easily manageable data pipeline and leading machine learning techniques. This system should help optimize fundraising strategies execution for the *Ignited* campaign, a ten-year fundraising initiative, and provide applicable insights into other areas of opportunity for the marketing department. While CRM-type fundraising software help provides strong capabilities for existing donor pools, they lack integrating advertisement data to communicate to people not already subscribed. The campaign planners should be able to make better decisions using previously unavailable data for these types of decisions on where to allocate funds and what type of media to publish. Areas of interest include messaging topics and advertising platforms' performance relative to other platforms and with specific audiences. Exploring the interaction of natural language processing along with multivariate time series analysis should lead to takeaways applicable to more than just university-specific fundraising but language and time analysis in general.

2 Literature Review

2.1 Fundraising Landscape

Fundraising design for private, nonprofit institutions relies heavily on cultivating relationships and partnerships. A substantial proportion (80%-95%) of organizations' donated dollars comes from few donors (5%-20%), while the number of people making small- and medium-sized gifts is declining (Shaker and Nelson, 2022). This concerning trend indicates that maintaining relationships is as important as ever, especially knowing that cultivating relationships takes a long time to secure those large contributions. Since the 1990s, fundraising teams have begun to adopt the "relationship fundraising" philosophy, which emphasizes personalized strategies to engage donors (Breeze, 2017). Shaker and Nelson (2022) indicated that relationship fundraising strategies are conducive to the donor experience. Donors feel involved and that their gifts are impactful, further building loyalty between the two parties.

Haruvy et al. (2020) performed an alternate approach to relationship fundraising. They outlined three different strategies for modeling fundraising design: (1) utility-based, (2) appeals-based, and (3) societal-based. In the utility-based method, donors/volunteers value the utility of a donation. In the appeals-based method, donators/volunteers react to cues/appeals as their primary decision to donate. In the societal-based method, donors/volunteers prioritize the welfare of the community. This novel framework by Haruvy et al. (2020) acknowledged fundraising as a business and recognized how the major actors are interconnected.

Due to the nature of nonprofits and the overall size of typical private universities, a challenge presents itself in the form of limited budgets and data maturity. Data-driven decisions such as where to advertise, who to direct those advertisements to, and what content the ads should contain may be tough to implement, given the limiting data infrastructure. Private universities may lack the required software to obtain and store the data and overall level of detail needed to optimize these decisions. In an ideal environment, large samples of properly labeled data and A/B testing methods can help support the fundraising strategy and focus on how and where to direct funds to raise money, maintain alums and community relationships, and manage brand reputation.

The most popular fundraising software (DonorBox, GiveSmart, Qgiv, amongst others) often acts as a customer relationship management tool. According to information provided on their respective websites. These platforms help maintain and analyze data on existing donors, assist with member management, plan fundraising events, and host donation portals. Many donators give numerous times across a single campaign, so effectively engaging those people and maintaining their relationships is extremely important. According to research by Neon One CRM and University of Texas at Dallas professor Elizabeth Searing, reoccurring gifts made up over 15% of annual revenue for organizations reviewed, although larger nonprofits were less likely to have them (Sarrantonio, 2021). While these services are excellent once someone has an established relationship, they offer little on the side of advertising management and help in attracting new donors. This is due to the diverse nature of potential platforms and ways to market to potential new donors.

2.2 Machine Learning Tools and Marketing Strategies

This section focuses on two principal areas: how Artificial Intelligence (AI) empowered the evolution of the marketing landscape and its application in a university-specific use case. According to Kohli and Jaworski (1990), a "market-oriented organization is one whose actions are consistent with the marketing concept" (p. 1). This early study distinguished a corporate marketing strategy or concept and its actual implementation. With the onset and growth of advanced analytics, including machine learning and AI, there is opportunity more than ever for any given organization to orient towards a specific niche market, and uniquely implement their strategy towards that market. Previous research by Davenport et al. (2020) has stressed how more marketing strategies across the industry are being designed around AI-enabled capabilities.

It is important to explore the benefits and challenges discovered in recent attempts to improve advertising or marketing strategies. Pamuksuz & Humphreys

(2021) implemented a type of social media monitoring that provides close to real-time visibility into how a brand is being perceived, improves a company's ability to judge the success of marketing campaigns, how the branding held through external events, and how feelings toward the brand compared to competitors. Yan & Pengfei (2021) found many obstacles when attempting to implement a customer response scoring model that is widespread across industries when it comes to data analysis. The biggest hindrance encountered was the increase in raw data that often can be unreliable, causing time of processing to be a more significant concern. Another frequently encountered limitation with substantial amounts of potentially usable data is feature engineering – spending significant amounts of time trying to produce usable features out of that large amount of data. Sarkar & De Bruyn (2021) showed that "recent neural network architectures, traditionally used in natural language processing and machine translation, could effectively do away with the complicated and time-consuming step of feature engineering." (p. 93). Specifically, long-short-term memory neural networks were used in predicting customer behavior responding to direct marketing attempts.

Machine learning use cases for universities have also been researched. Langston & Loreto (2017) successfully combined Customer Relationship Management (CRM) software with an extra layer of predictive regression analytics to improve undergraduate admissions recruitment. However, provided the disclaimer that the biggest key to success was still implementing a team of individuals to act on the model's findings. Another university use case also focused on admissions, where decision trees were used to create an interpretable rule-based list for actions. Not only was the model successful at prediction with a ROC of 99.7%, but it provided insight into the fact that prospect demographics and actions were more significant variables than the outreach activity toward the participants (Merritt et al., 2020). Choi & Kim (2020) explored twenty-three various machine learning techniques to see how they could be used in two broad categories of targeted advertising – user-centric and content-centric approaches. In both cases, numerous techniques were deemed applicable and useful to optimizing the target advertising success. A key takeaway is that it is important to incorporate subject matter experts into choosing the correct strategy to apply machine learning to. Rather than just in picking the best model, unless there is a large amount of funding for A/B testing atypical in a nonprofit or university environment.

A similar example more specific to the nonprofit/university fundraising environment was conducted by Chang (2012). Here, two styles of cause-related marketing, product-oriented, and cause-focused, were explored. Advertising can often experience negative consumer reactions, including cause-related marketing and fundraising attempts. Focusing on the product, such as the endowment a person is donating to or the cause of bettering a local institution, can have different results. Depending on the style, different messaging and visual content may perform better and should be considered when developing any marketing strategy.

The wide variety of marketing content and areas often grouped under machine learning can be intimidating to approach. Using keyword classification, researchers (Mariani et al., 2021) could narrow groups of research in this general field to eight key technical topics. The most relevant to this research are neural networks, linguistic analysis, social media and text mining, and social media content analytics. These topics can serve as a strong starting point for further technical research.

2.3 Marketing using Time Series Analysis

This section seeks to review trends in advanced time series methodologies related to forecasting. Traditional time series models utilize a single horizon (timeframe) to forecast the target of interest. An explicit limitation of this algorithm is the unreliable predictions as the modeler looks further into the future and the ignorance of the indicated path, e.g., adjusting for inflation or economic pressures in the duration of the fundraiser. Multi-horizon forecasting is a time series algorithm that predicts the target of interest using many data sources and inputs without prior information on how they interact with the target (Lim et al., 2021). The advent of deep learning architectures in conjunction with multi-horizon forecasting has become a robust method over conventional time series models. While previous studies focused on recurrent neural networks (RNN), attention-based layers have enhanced the selection of relevant time steps in the past beyond the inductive bias of RNN (Lim et al., 2021). Wen et al. (2022) stated that special challenges in time series tasks such as forecasting, anomaly detection, and classification have been successfully addressed through novel transformers. A distinct limitation here is the lack of interpretability in these black-box algorithms. Lim et al. (2021) introduced the Temporal Fusion Transformer (TFT) that combined multi-horizon forecasting with interpretable insights into temporal dynamics. TFT resulted in higher predictive performance while preserving explainability in the underlying models and the forecasts they produced.

The *Ignited* advertising campaign contains a variety of data sources and complex inputs, making multi-horizon forecasting a suitable algorithm. It can take inputs without prior knowledge of how they all interact (Lim et al., 2021). Relevant examples are known information about the future (e.g., holidays, game days), exogenous time-related variables (e.g., performance from previous campaigns), and static metadata e.g., demographics of current donors). A more robust application for *Ignited* is TFT which can leverage multi-horizon forecasting while empowering interpretability.

2.4 Marketing using Natural Language Processing

Modern NLP techniques can be leveraged to enhance marketing strategy in many ways, such as segmenting existing customers, gauging consumer sentiment, and prioritizing leads for expansion. Sukru Ozan (2021) utilized a customer relationship management (CRM) database containing text data on customer interactions to train a set of high-dimensional word embeddings that captured semantic and contextual meaning behind the terms commonly found across their documents. These vectors were used to construct an automated lead labeling system which improved time management and productivity by identifying high value customer leads. They also showed that the information in well-trained word embeddings could be successfully applied to various other text classification tasks. A similar application in the context of university fundraising could focus on improving engagement by identifying the most promising target audiences across each advertising domain. This can be time-consuming and difficult as a manual process due to the high audience variability across advertising platforms.

Methods originally created for language modeling have also been successfully applied to gauge advertising effectiveness in domains that are not traditionally viewed as text oriented. Perdices et al. (2021) investigated the problem of extracting website visits from internet traffic measurements and showed that profiling user browsing data could be cast as a text classification problem by viewing individual domains as words and sets of domains as documents. This type of analysis could provide insight into the effectiveness of advertising campaigns by quantifying ad engagement across geographies or other customer segments.

When applying any machine learning technique to text data, the first step is to transform the raw words into a numeric form that a learning algorithm can interpret. Which language components the transformation is applied to depends on both the algorithm and the analysis's specific goals. Numeric embeddings may be created to represent single characters, parts of words (sub-words), full words, or entire documents. The methods used to transform text to numeric will also change based on the language components used to generate the embeddings.

For creating clusters to uncover categories and relationships within a corpus of small documents (such as the text fields for a set of advertisements), one approach may be to transform each advertisement into a single document vector representing all the words in the ad. The most simplistic type of document vector is a count vector, originally used in the context of information retrieval by G. Sultan (1971). A count vector has a dedicated dimension for each word in the corpus vocabulary. The value in each vector element corresponds to the number of times the associated word occurred in a particular document. This means count vectors are typically remarkably high dimensional (e.g., thousands of dimensions) and are sparse representations since, in most cases, a single document will only utilize a small portion of the full corpus vocabulary. A slightly more advanced method for creating document vectors involves weighting the values in a count vector by each word's inverse document frequency (IDF), Sparck Jones (1972), which results in a Term Frequency Inverse Document Frequency (TFIDF) vector. The impact of using this IDF weighting is that a word's contribution to a document vector will be dampened when the word appears across many different documents. This is an intuitive result based on the assumption that words that show up often across different documents do not provide as much distinctive information about any document.

More recently, several successful methods have been proposed that utilize neural networks to learn embedding vectors for each word in a text corpus. One of the earliest successes is Word2Vec created by Mikolov et al. (2013), which proposed the Continuous Bag of Words (CBOW) and Skip-Gram neural network architectures for training word embeddings. They showed that the dense vectors learned by these architectures improved upon sparse representations that treat words as atomic units (e.g., TFIDF) by capturing both syntactic and semantic aspects of the language while utilizing significantly fewer dimensions. The FastText algorithm is a direct extension of Word2Vec that significantly improves the computational efficiency of the word vector training process, Joulin et al. (2016). FastText also improves the performance of learned embeddings by utilizing sub-word information to account for the internal structure of words, which results in more flexible and reliable representations, especially for rare words and words with multiple inflected forms Bojanowski et al. (2017). Neural network-based embeddings can also be used for transfer learning, which

means they are still effective for NLP applications where only a small amount of domain-specific text data is available. Mikolov et al. (2017) described an efficient process for pretraining FastText vectors on large bodies of text (e.g., Wikipedia). They showed that the semantic and statistical word information contained in those embeddings could be transferred to achieve state-of-the-art performance on various tasks.

The high dimensionality of text embeddings makes exploring the relationships between documents challenging since visualization tools commonly used for exploratory analysis cannot be directly applied to data in more than three dimensions. A solution is to find a dimensionality reduction technique that can map the data to 2 or 3 dimensions such that the overall structure of the high-dimensional data is preserved in a low dimensional plot. T-distributed Stochastic Neighbor Embedding (t-SNE), proposed by van der Maaten and Hinton (2008), has been shown to generate particularly useful representations. It can accomplish this by mapping data to a lower dimensionality while retaining a significant amount of local information as well as some of the global structure of the original high dimensional data. The t-SNE mapping is iteratively optimized using gradient descent, which can be made more computationally efficient and less noisy if another dimensionality reduction technique is used. An example of a reduction technique is Principal Components Analysis (PCA)Hotelling (1933) and is first used to reduce the data to a more modest number of dimensions such as 50.

An alternative to t-SNE, proposed by McInnes et al. (2020), is Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). McInnes argues that UMAP is competitive with t-SNE in terms of visualization quality, is less computationally expensive, and captures more of the global topological structure in the data. They also state that UMAP can scale to significantly larger data sets than are feasible with t-SNE and that the UMAP algorithm can effectively reduce dimensionality for even exceedingly high dimensional data. The latter point has the important consequence that the UMAP algorithm does not require a preprocessing step to reduce the number of dimensions with a secondary technique such as PCA, which is the standard recommendation when using t-SNE.

Machine learning is already being applied heavily in the marketing industry, including nonprofits, to a more limited extent. For companies without robust data talent, "out of the box" marketing software may have built-in basic machine learning classification and analysis techniques but falls short of keeping up with recent developments in neural networks and other more advanced methods. These methods have demonstrated the ability to add value to a campaign in optimizing dollars and awareness achieved when implemented appropriately.

3 Methods

The primary data for this research is SMU *Ignited* social media advertisement data from three platforms – Facebook, Twitter, and LinkedIn. Access to this data has been provided by the SMU MARCOM department and is pulled directly from each platform's ad management platform. Naming conventions, available fields, and overall

data structure differ between each platform, so work must be done to transform and aggregate the data across platforms. All the data contains categorization labels for each level of aggregation, outlined next. In addition, the data is collected as daily observations, providing information on advertisement performance by day. The time range is from January 1, 2021, to October 29, 2022.

The ad campaign is the highest level of ad categorization for all three platforms. These can be months long and are typically where the objectives are set for ads within any given campaign. Example objectives include awareness, reach, post engagement, or clicks. Furthermore, a campaign may not be specifically under the umbrella of *Ignited* but instead a more general advertising campaign by the university, most often to do with admissions and to reach potential undergraduates or their parents. The next level of aggregation is called an Ad Group. This is primarily where specific target audiences and spending categories are set. The target and objective for the ad group is not stored in a separate field but rather called out through naming conventions for the Ad Group name itself. An awareness campaign may have an ad group on the Facebook platform targeting Instagram users interested in sports and another ad group targeting parents of Alumni on Facebook itself. Each ad group is then made up of a set of Ads. Each ad is associated with a "creative," which may be a piece of media, video, straight text, or a combination. This is what is displayed to the audience at the ad group level. Metrics in focus for the scope of this research were available across all platforms. Follow-up studies could be done on platform-specific metrics, such as more detailed demographic information that Facebook provides than other platforms.

Table 1. Social Media Advertisements collected data fields of interest

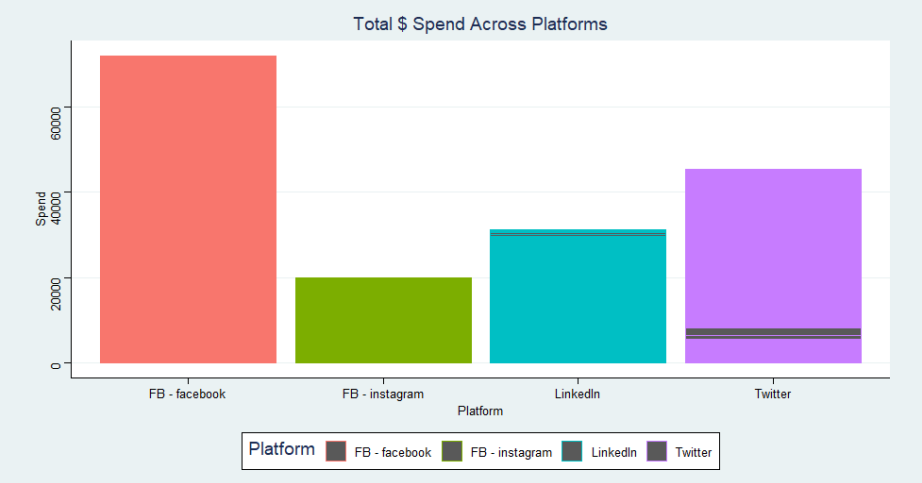
Field Name	**Description**
Text	Text contained in the creative
Spend	The dollar amount spent on that ad on that day
Impressions	Number of views of an advertisement
Clicks	Number of times the ad link to a website was clicked
Reach	Unique number of advertisement views
Video Starts	Times an ad video was played
Video Completions	Times an ad video was completed
Engagements	Count of times a post was commented, liked, or shared. Also includes platform specific engagement options such as quote tweets
Results	Provides insight into the goal of an ad and performance on that goal
Objective	Objective of the associated ad group, can differ from result

In the case of Twitter, reach is not available at the detailed ad level. Instead, it is only available for ad groups. As a result, the data were imputed from ad group detail to the ads, based on the assumption that the percentage of impressions for the ad is

directly correlated with its share of reach. Another needed transformation was for Facebook, where the ad performance data did not have a direct platform tie-in with the creative information. In these cases, an extra join was performed from data containing the creative information for each ad.

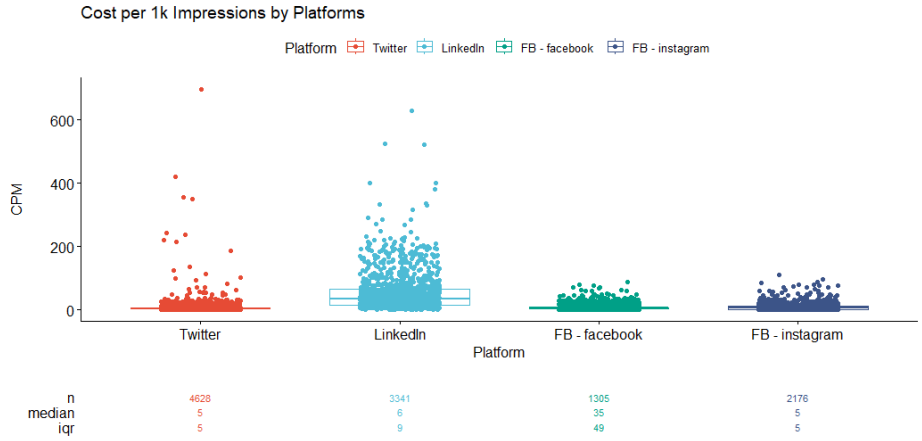
A quick view of the total spends across the primary campaign social media platforms, shown below, indicates that a substantial portion of funding currently goes to Facebook. The total spends there is more than triple what was spent on Instagram. The other two large platforms, LinkedIn, and Twitter are closer in total spending, but there is still a significant difference.

Chart 1. Totals spend across platforms



A common metric to track ad performance is CPM, which is the cost per 1,000 impressions. This provides a quick indicator of the efficiency of the ad. Median CPM is consistent across platforms, except for Facebook, which is significantly higher than others.

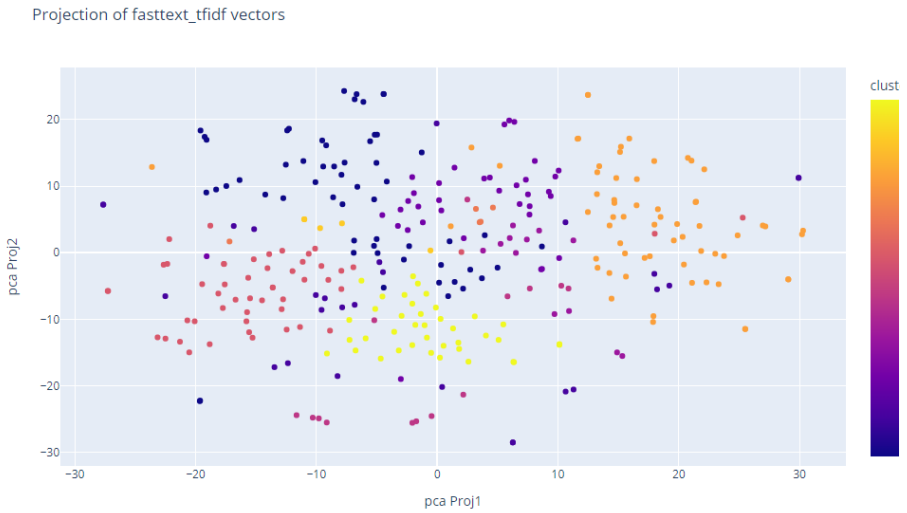
Chart 2. Cost per 1k Impressions across platforms by daily ad



With the data extracted and standardized, the next step in this data pipeline is to create a new field that categorizes the ad into a cluster based on the content of the ad creative. Multiple algorithms will be explored for the clustering and visualizations of the clusters, such as TFIDF, t-SNE, and UMAP. Whichever is determined to provide the most interpretable clusters while still being potentially useful will be used with the final model. The aim of this step is to narrow the considerable number of potential text fields to help categorize the messaging of an ad. For example, if a cluster contains athletic related advertisements, that information could be useful later in performance prediction and understanding how certain messaging does with target demographics.

Dimension reduction techniques can also be useful to supplement the clustering embeddings and will be explored for benefit in the final model. PCA and truncated SVD methods will be attempted to reduce the text dimensions prior to cluster creation to see if any value is gained in useability of the final clusters. After clustering, a manual analysis and manual labeling must be performed on the interpretability of assigned clusters to translate the clusters into something usable for the planning team. The graph below shows an early attempt at visualizing the clusters in 2D with PCA. The yellow cluster in the middle contains all the advertisements that contain some sort of quote.

Chart 3. Initial attempt at combining PCA and TFIDF clustering



After clustering and labeling is complete, the next step in the pipeline will be to create models for predicting ad performance at various levels of aggregation and segmentation. A naïve model will first be created as a baseline of performance to compare against, using the average metric performance as the predictor. Next the Temporal Fusion Transformer (TFT) algorithm will be used to create a prediction model that implements the time series information inherited to the data. While TFT is noted to remain interpretable, another model based on multiple logistic regression will be used without incorporating the time series information. This will serve as another comparison point for the TFT model and can potentially provide interesting insights to the business on the importance of various variables in ad performance. Models will be evaluated based on the root mean squared error (RMSE) of the predictions to actual values, using a rolling holdout window of time to compare forecasted verse actual.

The final output of a model should be able to provide prediction on key metrics based on input variables such as platform, cluster topic, number of days the ad is running, and spend. Key metrics for prediction include cost per impression, click, and engagement. Also included will be a visualization of the clustering to help explain the topics and messaging within each cluster. Ideally, some statistical ANOVA analysis will also provide insight into what variables are or are not statistically significant to performance.

We created Professor Text, an application that encompassed text embedding, dimensionality reduction, clustering, and visualization methodologies referenced in the section above. On the next tab, regression assumptions and metrics are addressed from the corresponding clusters in the previous tab. KPI predictions can be discerned on the same tab based on user's input. The last tab displayed Twitter post generation. The user has ample settings to tune for their own edification of how text clustering algorithms work as well as analysis of the texts themselves.

4 Results

This research aimed to implement a machine learning analysis to optimize and predict advertisement performance for SMU's current fundraising campaign. The goal is for SMU's Marketing and Communications department to execute their desired campaign goals as well as to have better insight into how and why things are or are not working with their current and previous advertising efforts.

We conclude that there is sufficient evidence to suggest that posts from the *Ignited* campaign are not dissimilar. Hugging Face's RoBERTa large model produced the best separated clusters, and we explored this further by lowering t-SNE perplexity hyperparameter, so it mapped the points to various locations from 50 dimensions (**Chart 4**) down to two (**Chart 5**). According to Hugging Face's documentation, RoBERTa large is conducive for tasks that utilize full sentences, and this aligned with the input texts observed to be full sentences as well.

Chart 4. t-SNE of RoBERTa, agglomerative clustering, cosine similarity, 50 dimensions

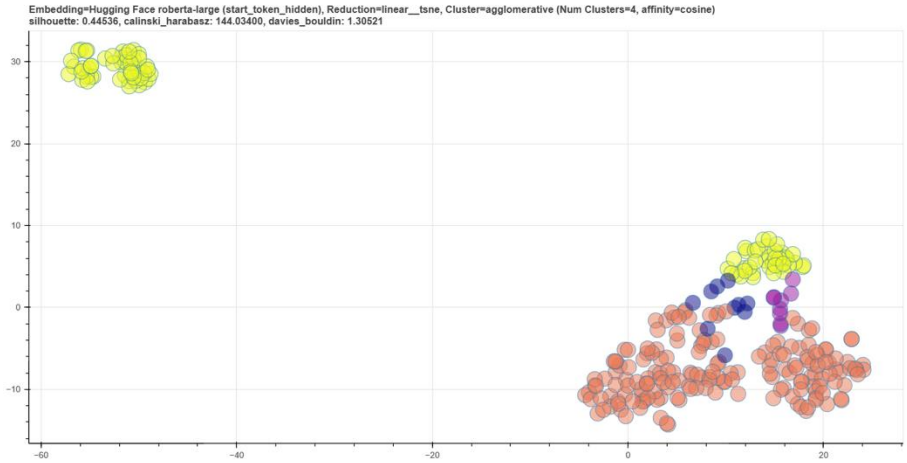


Chart 5. t-SNE of RoBERTa, agglomerative clustering, cosine similarity, two dimensions

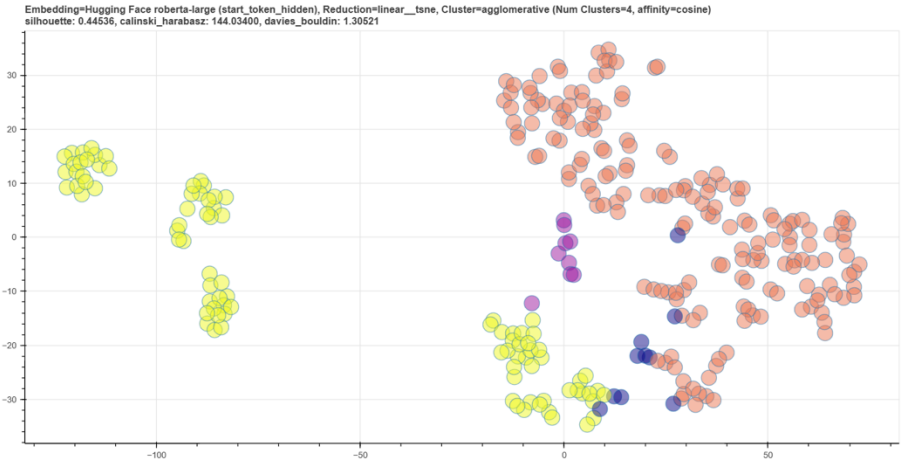
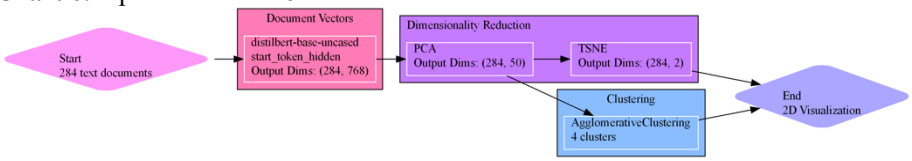


Chart 6. Pipeline of Chart 5



Next, we conducted TF-IDF embedding with the same settings as above and compared their clustering metrics. As depicted in Chart 7, TF-IDF did not produce clusters as nicely separated compared to RoBERTa large and produced a silhouette score 15 magnitudes smaller.

Chart 7. TF-IDF

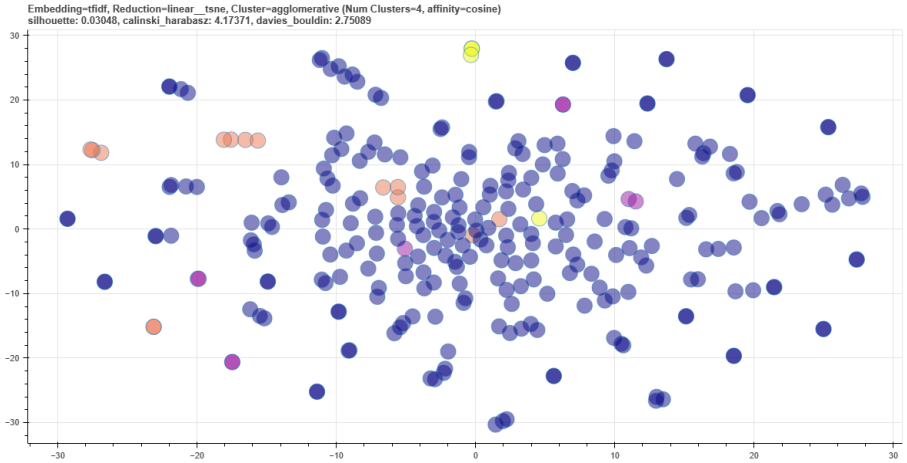
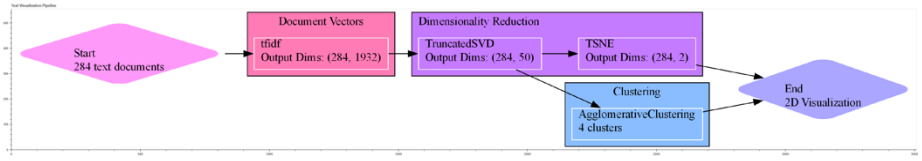


Chart 8. Pipeline of Chart 7



The scope of the research is limited to social media posts from the *Ignited* campaign, which spanned from September 2021 to the current date (February 2023).

5 Discussion

Based on these results, there is evidence to suggest that the content of an advertisement for SMU's social media has an impact on performance. If an advertisement's text is assigned to cluster 1, for example, it would be expected to outperform an advertisement belonging to cluster 5 across all platforms. Furthermore, there is evidence to suggest that certain platforms outperform others based on the raw metric of impression.

With a smaller number of clusters, insights can emerge. A small number of clusters are more likely to be interpretable, so if cluster 1 contains all advertisements relating to research or laboratory investments and is best performing, the ad creative team can use this information when deciding what to center an ad around. While there are times a certain topic may be required to be the focus of the ad content, these results indicate that these topic requirements may come at a cost to overall success in achieving goals of impressions or reach.

Advertising in any context is complex in how to judge success and decide on proper actions, and university fundraising is no exception. Maximizing a single metric like count of impressions ignores many other key factors, such as variety of demographics contribution to those impressions. However, the results from this study confirm that ad content and platform placements are important to keep in mind when trying to reach target audiences. While subject matter expertise may lead a decision maker to value a LinkedIn impression over a Twitter impression, if per dollar twitter can provide more impressions overall, the decision to spend more on LinkedIn is less obvious. These results can help inform SMU's Marketing and Communications department just how large or small the incremental benefits of topic or platform changes are and should be considered as additional context in future spending decisions.

The current research in the text clustering space primarily utilizes transformers, which rely on an abundance of data to extract meaningful data representation of clusters. However, the data used in this study was 90% less than the typical minimum total of observations and consists of a smaller number of unique posts and individual texts, rendering it not conducive for transformers. Furthermore, there is a lack of studies that focus on optimal text clustering techniques for different data types and sizes.

SMU Marketing and Communication is in its beginning state to become a data driven organization; therefore, their data is quite undeveloped in size and quality to conduct machine learning effectively. This limited the suitable text cluster

methodologies and questioned the reliability of the results the clusters produced. For example, TF-IDF may be more useful for smaller data yet does not consider word order, context, or different word meanings. Transformers such as BERT were pre-trained on Web corpora so they can generate contextual representations but cause bias (nonfactual content) and would not be appropriate for the limited data present in the research. A potential solution to combat data limitation is leveraging TF-IDF which narrows the scope of the corpora to be domain-specific.

Regarding the research's ethical impact, the data was provided by SMU Marketing and Communication and did not contain Personally Identifiable Information (PII). It encompassed the aggregation of performance metrics from social media platforms and Google Analytics, which protects personal information to be linked to an individual donor or viewer. This aligned with the legal framework set by the General Data Protection Regulation (GDPR).

There are many opportunities that can be recommended for future efforts. One such opportunity is to incorporate a limited amount of target audience demographic data to help account for differences in how certain ads are placed even within a single media platform. Another key area for improvement would be to add additional analysis based on the present and type of media included in an advertisement, which can be a picture or video. Last, but not least, would be to expand the scope of analysis beyond social media, and look at other streams of campaign communications. This would include data sources such as from Google Analytics, Out of Home ads from mediums like billboards, email communications to previous donors or alumni, Television commercial ads, and more. The type and presence of these ads running at the same time as a social media ad could potentially impact the performance of the social media ad or how likely a person is to click through and donate.

6 Conclusion

The overall research culminated as Professor Text, an application that conducted a wide array of embedding and clustering algorithms while depicting its corresponding regression assumptions and statistics. Text generation was also explored as well as KPI predictions from any given texts. We hope this research is the start of a budding partnership between SMU's Marketing and Communications department and Master of Science in Data Science program. This behooves MARCOM in their journey towards being a data-driven department while simultaneously providing an opportunity for aspiring data scientists to work with real data and encounter challenges as one would in a practical setting. We also hope future MSDS students can build on Professor Text and get it to a working application that provides actionable insights for MARCOM.

Acknowledgments.

A. Jacquelyn Cheun, PhD. – Capstone professor

References

1. Ait Hammou, Ait Lahcen, A., & Mouline, S. (2020). Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. Information Processing & Management, 57(1), 102122–. doi.org/10.1016/j.ipm.2019.102122
2. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with sub word information. Transactions of the association for computational linguistics, 5, 135-146.
3. Chang, C.-T. (2012). Missing ingredients in cause-related advertising. International Journal of Advertising, 31(2), 231–256. doi.org/10.2501/ija-31-2-231-256
4. Choi, & Lim, K. (2020). Identifying machine learning techniques for classification of target advertising. ICT Express, 6(3), 175–180. doi.org/10.1016/j.icte.2020.04.012
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
6. Gkikas, Theodoridis, P. K., & Beligiannis, G. N. (2022). Enhanced Marketing Decision Making for Consumer Behaviour Classification Using Binary Decision Trees and a Genetic Algorithm Wrapper. Informatics (Basel), 9(2), 45. doi.org/10.3390/informatics9020045
7. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
8. Kotras. (2020). Mass personalization: Predictive marketing algorithms and the reshaping of consumer knowledge. Big Data & Society, 7(2), 205395172095158–. doi.org/10.1177/2053951720951581
9. Langston, R., & Loreto, D. (2017). Seamless integration of predictive analytics and CRM within an undergraduate admissions recruitment and marketing plan. Strategic Enrollment Management Quarterly, 4(4), 161–172.
10. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764. doi.org/10.1016/j.ijforecast.2021.03.012
11. Mariani, Perez-Vega, R., & Wirtz, J. (2022). AI in marketing, consumer research and psychology: A systematic literature review and research agenda. Psychology & Marketing, 39(4), 755–776. doi.org/10.1002/mar.21619
12. Merritt, Stephen; Francomano, Anne; and Garcia, Martin (2020). "Optimizing the Enrollment Funnel with Decision Trees and Rule Based List," SMU Data Science Review: Vol. 3: No. 1, Article 3.
13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
14. Ozan. (2020). Case studies on using natural language processing techniques in customer relationship management software. Journal of Intelligent Information Systems, 56(2), 233–253. doi.org/10.1007/s10844-020-00619-4
15. Pamuksuz, Yun, J. T., & Humphreys, A. (2021). A Brand-New Look at You: Predicting Brand Personality in Social Media Networks with Machine Learning. Journal of Interactive Marketing, 56(1), 55–69. doi.org/10.1016/j.intmar.2021.05.001

16. Perlich, Dalessandro, B., Raeder, T., Stitelman, O., & Provost, F. (2013). Machine learning for targeted display advertising: transfer learning in action. Machine Learning, 95(1), 103–127. doi.org/10.1007/s10994-013-5375-2
17. Perdices, Ramos, J., García-Dorado, J. L., González, I., & López de Vergara, J. E. (2021). Natural language processing for web browsing analytics: Challenges, lessons learned, and opportunities. Computer Networks (Amsterdam, Netherlands: 1999), 198, 108357–. doi.org/10.1016/j.comnet.2021.108357
18. Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting, 36(3), 1181-1191.
19. Sarkar, & De Bruyn, A. (2021). LSTM Response Models for Direct Marketing Analytics: Replacing Feature Engineering with Deep Learning. Journal of Interactive Marketing, 53(1), 80–95. doi.org/10.1016/j.intmar.2020.07.002.
20. Shah, Engineer, S., Bhagat, N., Chauhan, H., & Shah, M. (2020). Research Trends on the Usage of Machine Learning and Artificial Intelligence in Advertising. Augmented Human Research, 5(1). doi.org/10.1007/s41133-020-00038-8
21. Urban, Timoshenko, A., Dhillon, P., & Hauser, J. R. (2020). Is Deep Learning a Game Changer for Marketing Analytics? MIT Sloan Management Review, 61(2), 71–76.
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
23. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2022). Transformers in time series: A survey. arXiv preprint arXiv:2202.07125.
24. Yan, & Pengfei, L. (2021). Marketing customer response scoring model based on machine learning data analysis. Journal of Intelligent & Fuzzy Systems, 40(4), 6445–6455. doi.org/10.3233/JIFS-189484.
25. Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. Advances in neural information processing systems, 31.
26. Bureau, U. S. C. (2022, September 27). *American Community Survey Data*. Census.gov. Retrieved November 5, 2022, from https://www.census.gov/programs-surveys/acs/data.html
27. Mitchell, M., Leachman, M., & Masterson, K. (2017, October 23). *A Lost Decade in Higher Education Funding State cuts have driven up tuition and reduced quality*. VTechWorks Home. Retrieved November 5, 2022, from https://vtechworks.lib.vt.edu/handle/10919/83618?show=full
28. Sarrantonio, T. (2021, January 7). *Research validates importance of recurring giving*. NonProfit PRO. Retrieved November 6, 2022, from https://www.nonprofitpro.com/post/new-industry-research-validates-importance-of-recurring-giving/
29. Shaker, G. G., & Nelson, D. (2022). A Grounded Theory Study of Major Gift Fundraising Relationships in U.S. Higher Education. Nonprofit and Voluntary Sector Quarterly, 51(5), 1054–1073. https://doi.org/10.1177/08997640211057437
30. Haruvy, E., Popkowski Leszczyc, P., Allenby, G., Belk, R., Eckel, C., Fisher, R., Li, S. X., List, J. A., Ma, Y., & Wang, Y. (2020). Fundraising design: key issues,

unifying framework, and open puzzles. Marketing Letters, 31(4), 371–380. https://doi.org/10.1007/s11002-020-09534-8

31. Kohli, A. K., & Jaworski, B. J. (1990). Market Orientation: The Construct, Research Propositions, and Managerial Implications. Journal of Marketing, 54(2), 1–. https://doi.org/10.2307/1251866

Appendix

Use if needed for additional information