When should I tweet? Estimating the causal effect of tweets' publication hour on their popularity

Omer Madmon and Gal Sasson

Project repository: CausalInferenceProject

1 Introduction

Social networks became a major part of our day-to-day life. Besides their use for socializing, social networks are also used for advertising and promotions of products, companies, events, etc. Product growth and marketing analysts in companies often try to optimize their internet campaigns via various parameters such as content, publishing time, design, audience and many more. Note that not only companies care about this type of optimization, but also many private users take those considerations into account when posting. As such, optimizing social networks engagement became a point of interest for many users, both individuals and business companies.

In this project our scope will be limited to Twitter data. We will focus on identifying and estimating the causal effect of tweets publication time on users' engagement and tweet popularity. Specifically, we will look at the hour-of-day as the treatment and at a measure of engagement/ popularity (e.g., relative number of likes, retweets, etc.) as the outcome.

2 Problem definition

In this project we aim to identify the conditional average treatment effect (CATE) of tweets publication time on tweet popularity, in two levels of granularity. The first level will be by conditioning only on the tweets' domains, and the second level will be by conditioning on high-dimensional covariate vector, including the tweets' domains, text, user features, etc. Understanding the causal effect of publication hours on tweets' popularity in different granularity levels may enable publication time optimization by users or companies seeking to increase engagement.

Formally, let us define the treatment, outcome and potential outcomes. We consider discrete treatment indicating the window of hours in which a tweet is published as follows:

T_{0-6} – Tweet was published between 12 am – 6 am

T_{6-12} – Tweet was published between 6 am – 12 pm

T_{12-18} – Tweet was published between 12 pm – 6 pm

T_{18-0} – Tweet was published between 6 pm – 12 am

The potential outcomes are defined as the tweets' popularity has it been published in a specific window of hours. Formally, the potential outcome is defined as follows:

$$Y_t = \frac{L - \bar{L}}{s}$$

Where L is the number of favorites the tweet received had it been published in the hours interval $t, t \in \{0-6, 6-12, 12-18, 18-0\}$. $\bar{L}$ is the user's average number of favorites on its tweets and s is the standard deviation of the user's number of favorites. Therefore, we refer to this measure as the scaled favorite count. This normalization enables us to measure a tweet's popularity relatively to the user's popularity. For example, consider a user with an average favorite count of 1,000 and a user with an average favorite count of 10. If both published tweets receiving 1,000 favorites, it is clearly a very popular tweet only for the second user while for the first user it's just a normal tweet.

We assume consistent outcome (as we discuss in more details in the identification section), i.e. the outcome of a tweet is $Y = \Sigma_t \mathbb{I}_{\{T=t\}} \cdot Y_t$.

Therefore, we aim to identify the following CATEs:

1. $\mathbb{E}\left[Y_{t_{a-b}} - Y_{t_{c-a}} \middle| D = d\right]$
   $\forall\, t_{a-b}, t_{c-a} \in \{0-6, 6-12, 12-18, 18-0\}, d \in \{health, music, technology, politics\}$
   That is, we only look at adjacent pairs of treatment, as the treatment is cyclic.
2. $\mathbb{E}\left[Y_{t_{a-b}} - Y_{t_{c-a}} \middle| X = x\right]$ $\forall t_{a-b}, t_{c-a} \in \{0-6, 6-12, 12-18, 18-0\}$ and x is a high dimensional covariates vector.

3 The data

We have created a dataset of tweets from 4 domains – health, politics, music and technology. All tweets were published in the USA in order to ensure a unified time zone, and under the assumption that the majority of audience for such tweets are also users from the USA.

3.1 Data collection

We collected the data using the Twitter API and the Tweepy package. The API enables retrieval of tweets from the past 7 days, and has limitations on the amount of requests that we are allowed to make. Under those constraints we have collected as many tweets as possible from each domain, between the dates 14.7.22 to 23.7.22. The tweets sampling of a specific domain was done using a query of a hashtag corresponding to the domain (e.g., #music). All the tweets were in English, from the USA and in a unified time zone. Tweepy enabled us to retrieve the tweets with additional metadata such as timestamp, location, language, text, favorite count (number of likes), number of retweets, and user metadata.

In order to measure the popularity of tweets as defined in the problem definition section, we filtered the dataset so it will only contain tweets that belong to users that have more than one tweet. Otherwise, the empirical standard deviation is not well defined. After the filtering, we were able to add a "scaled favorite count" measurement to each of the tweets.

After taking the above steps, the dataset contains the following number of tweets from each domain:

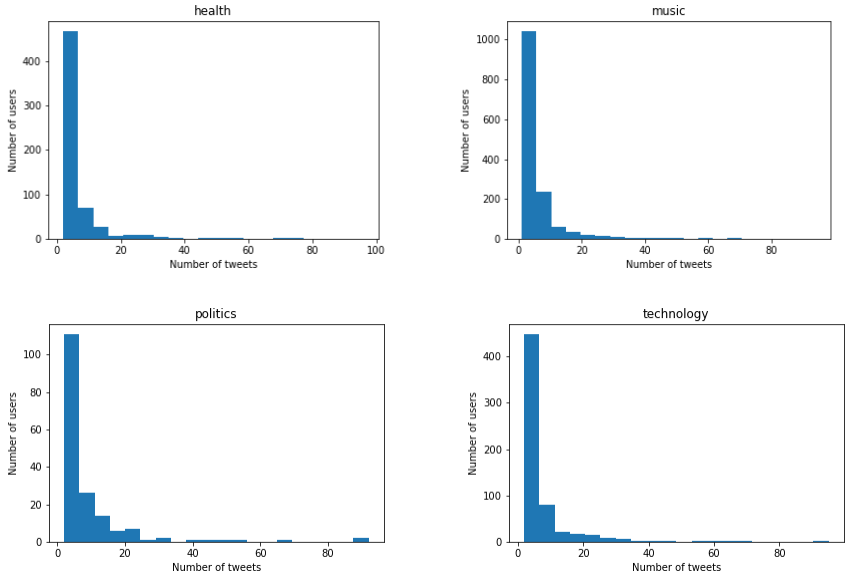
Domain	Number of tweets
Health	4,157
Music	9,035
Politics	1,483
Technology	4,741

Note that within this data collection process, the dataset only contains tweets from a specific range of dates, location, language and domains. This means that extrapolating the results beyond these characteristics may require additional data, as the causal effect might depend on these features which are fixed in our dataset.

3.2 Exploratory data analysis

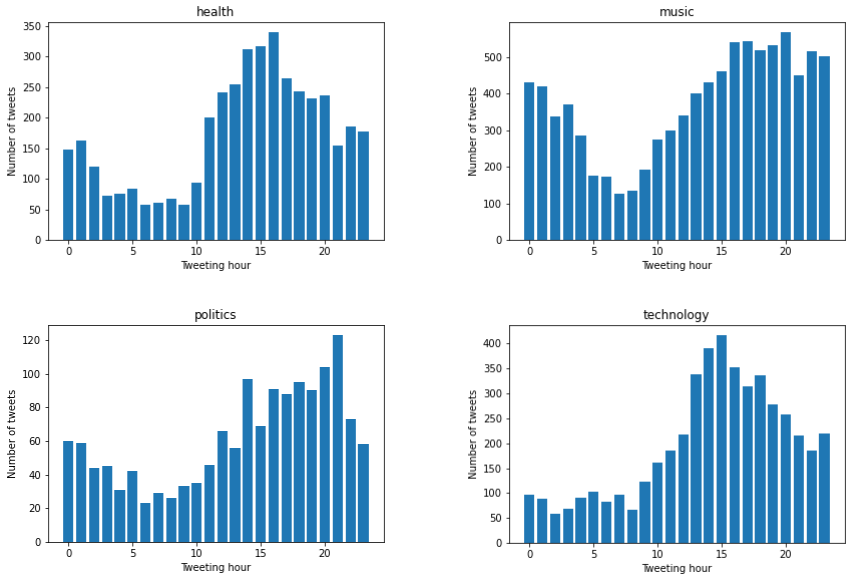
We started by exploring a few distributions for each domain separately, in order to validate our data collection process within and between domains.

1. **Tweets-count per user distribution**: we look at the distribution of the number of tweets published by each user in our dataset.



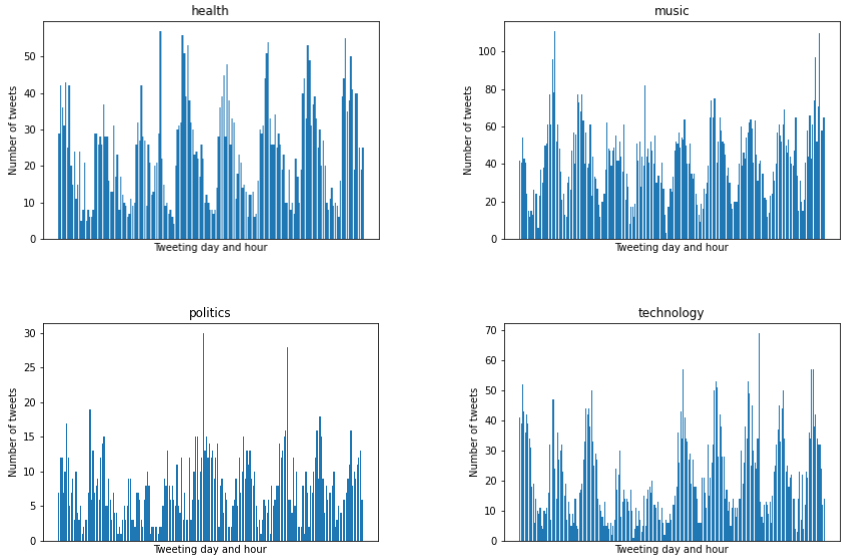
As expected, we can see a heavy-tailed distribution of the tweets count. This makes sense as most of the private users do not post tweets too frequently, and the few users who does might be celebrities, companies, news channels, etc.

2. **Tweeting hour distribution**: we start by looking at the publishing hour of the tweets, regardless of their date.



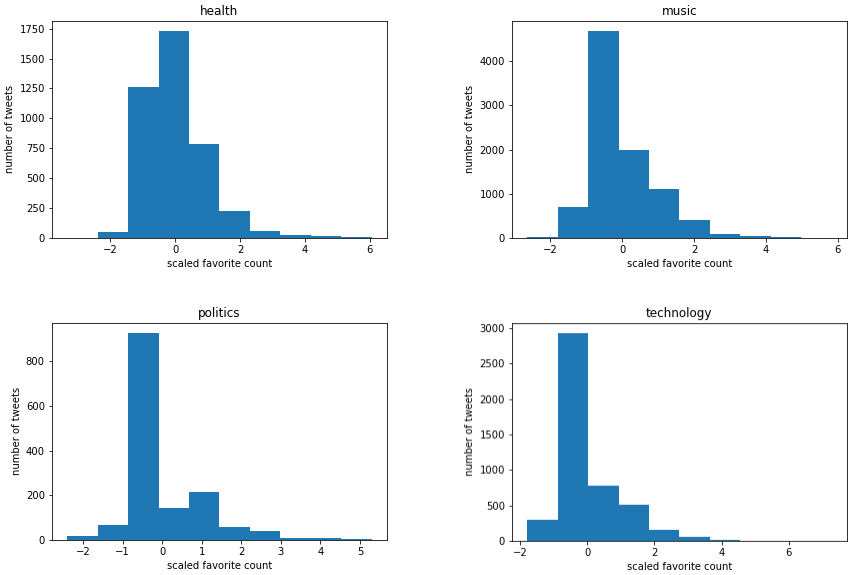
We can see that users tend to tweet more during the day in comparison to night hours. There is a gradual decrease in tweets during the night followed by a gradual increase in tweets as the day proceeds. Note that there is a sufficient amount of tweets from each hour.

3. **Tweeting day-hour distribution**: we now look not only at the hourly distribution, but also on the distribution between different days.



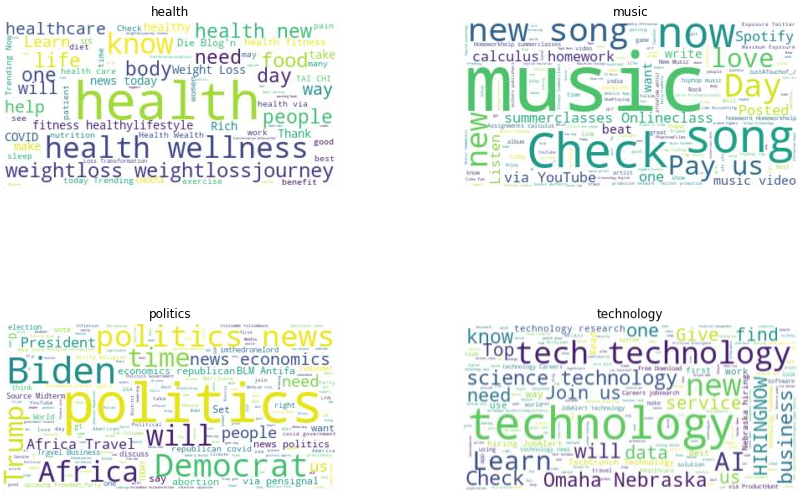
We clearly see a similar pattern across the days, which fits the hourly distribution from the previous point, and demonstrates seasonality in the data. We also see less tweeting during the weekend.

4. **Scaled favorite count distribution**: this shows the distribution of the popularity measure as defined previously.



The graphs fit the empirical property of normal distribution – approximately 95% of the values are no more than 2 STDs from the mean. In our case, this means that most of the tweets are neither very popular nor unpopular (relatively to the specific user).

5. **Word clouds:** we plot the text of the tweets after minimal preprocessing, using word clouds, in order to ensure the subjects of the tweets fit the domain.



4 Identification

In the following section, we will discuss the identification assumptions and possible failures in our setting. We consider those assumptions within in between the domains.

4.1 SUTVA

The SUTVA consists of two assumptions:

1. **The potential outcomes for any unit do not vary with the treatments assigned to other units:** in most cases it is a reasonable assumption, as users are free to engage with a tweet regardless of their engagement with other tweets or their publication hours.
   However, in some cases the will of a user to engage with a tweet may depend on the publication time of other tweets. For example, Fox News posts a tweet exposing a big item that had never been published before on 12 pm. Its potential outcome would have decreased had CNN posted this item 2 hours earlier.
   Notice that if we care about SUTVA within each domain, then the more high-level the domain is, it is less likely the situation above will occur.
2. **For each unit, there are no different forms or versions of each treatment level which lead to different potential outcomes:** this part of the assumption weakly holds, as the treatment was originally continuous and was discretized into multiple hour intervals. This means that receiving the treatment $t = 0 - 6$ has many forms – the tweet can be published at 1 am or at 5 am, which may lead to different potential outcomes. This can be solved by shortening the intervals, which introduces a tradeoff between reliability of the assumption and the number of possible treatments (which leads to more complicated estimation).

4.2 Consistency

We say that consistency holds if for a unit that receives treatment T, we observe the corresponding potential outcome Y_T. We indeed observe the corresponding potential outcome for every tweet (in this case, user engagement). However, notice that the way we have defined the scaled favorites count is actually an approximation of the measure we really care about: where we have used the user's mean and empirical standard deviation of the favorites count we should have used the mean and standard deviation of the underlying distribution, which is unknown.

4.3 Ignorability

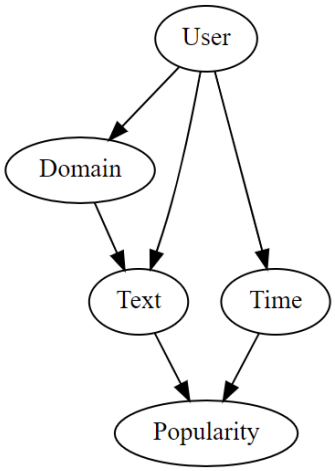
To show that ignorability holds we will use the CGM approach. We will first describe and justify a causal graph encoding our assumptions on the causal relations in our problem. Then, we will show that the backdoor criteria holds, implying the causal effect can be estimated. Lastly, we will consider several possible hidden covariates and show that our backdoor set is still valid.

We start by defining a causal graph encoding our beliefs. We consider groups of covariates rather than treating each covariate separately:

- Time (treatment) – the hour interval as defined previously.
- Popularity (outcome) – the scaled favorites count as defined previously.

- User – includes user features such as account creation date, followers count, textual description of the user, etc.
- Domain – one of the 4 tweets domains. Note that this covariate is only relevant for the ignorability assumption between domains. All following conclusions also apply if we remove this node.
- Text – features extracted from the tweet's text.

According to our hypothesis, the time of publication affects the popularity of a tweet. The user attributes seem to have an effect on the domain (by the user's interests), text (by the user's writing style) and the time of posting (by the user's habits). Note that we have defined the tweet's popularity relatively to other tweets posted by the same user, and not with respect to other users, and therefore there is no direct effect of the user on the tweet's popularity. The domain effects the text of the tweet as it directly dictates the content and the writing style, as the audience differs between domains. Finally, we believe that the tweet's text directly affects its popularity.



Note that in this graph all covariates are observable, and the set of $\{User, Text, Domain\}$ satisfies the backdoor criterion. We believe that there are no hidden confounders, i.e. no hidden covariate H that satisfies both $H \rightarrow Time$ and $H \rightarrow Popularity$. However, there are possible hidden covariates that should be taken into consideration:

- **H such that ($H \rightarrow Time$ and $H \rightarrow Text$) or ($H \rightarrow Time$ and $H \rightarrow Domain$)**. For example, a big news event that motivates users to tweet about it right after it happens.
- **H such that ($H \rightarrow Time$ and $H \rightarrow User$)**. For example, the users' age determines various users' characteristics (such as the creation date of the account) and may also influence the time of publication of most tweets by that user (teenagers tend to stay up late and tweet all night while adults tweet when they're bored at work).
- **H such that ($H \rightarrow Popularity$ and $H \rightarrow Text$) or ($H \rightarrow Popularity$ and $H \rightarrow Domain$)**. For example, when there is a trending subject on social networks (such as "Black Lives Matter"), a user might decide to post about it, and if he gets lucky his tweet might more popular than usual as this subject is now trendy.

Note that even if we consider such hidden covariates, the same set we discussed earlier still satisfies the backdoor criterion. Lastly, notice that there might be some observed/hidden reverse confounders (Z such that $Time \rightarrow Z$ and $Z \rightarrow Popularity$). For example, such covariate might be misspelling in the tweet's text – if a tweet is written at 3 am it is more likely to contain spelling mistakes, which might lead to decrease in the popularity. In this case, we don't want to include Z in our backdoor set (since it mustn't include a node that is a descendent of the treatment), and when not included the previous set still satisfies the backdoor criterion.

All backdoor sets discussed in this section were validated using causal graphical models (the code can be found here).

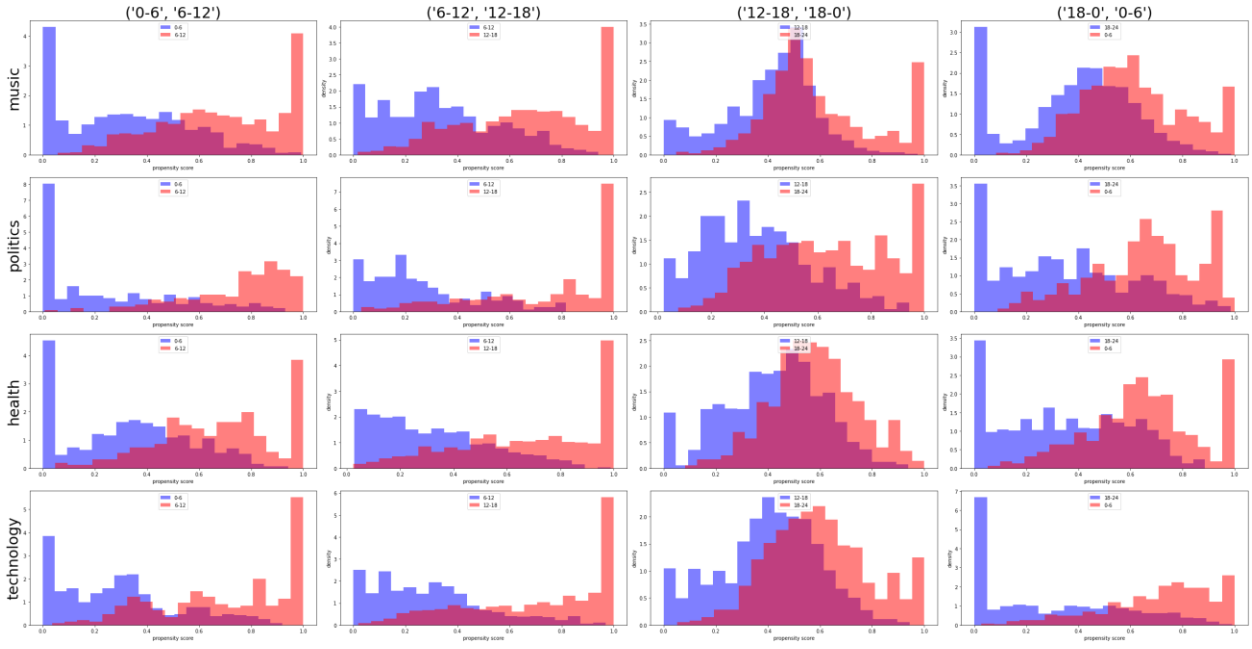
4.4 Overlap

The common support (overlap) assumption states that $P(T = t|X = x) > 0 \; \forall t, x$. In our case, since we're only interested in estimating the CATE of every two adjacent treatments, we consider this assumption only between the relevant pairs. Meaning, the assumption holds if for every tweet there are positive probabilities to be published in the next bin of hours and in the previous bin of hours.

There are a few reasons to believe that overlap indeed holds. First, we know for sure that technically speaking, nothing prevent a tweet from being published at a certain hour. Second, although there are tweets that are more likely to be published at a specific time (for example, "I am now eating the best breakfast ever!") it is still not entirely deterministic (for example, if that breakfast is eaten in Benedict, a restaurant that serves breakfast all day long).

In order to empirically validate our assumptions, we train a logistic regression model for each pair of adjacent treatments, and confirm overlap as for the corresponding binary case. The covariates vector X contains representations of the tweet's text, user textual description and other user covariates. This representation will be further explained in the "Data representation" section.

The following plots visualize the propensity scores of each such binary case in each of the domains:



As can be noticed from the graphs, the overlap assumption seems to hold for all relevant pairs of treatment. Interestingly, we see different patterns between the pairs that repeat in all domains. For example, notice that in all domains the pair of $(12-18, 18-0)$ demonstrates higher overlap than the other pairs. This is

intuitive as people likely behave in a more similar manner during the afternoon $(12-18)$ and evening hours $(18-0)$ rather than, for example, between the evening $(18-0)$ and after midnight $(0-6)$.

In addition, in order to verify that our logistic regression models are reasonable (in terms of prediction quality), we have computed the AUC of each classification task. The average AUC score was 0.74.

5 Estimation

5.1 Data representation

In this section we will discuss how we processed and represented the covariates used to estimate the CATE. Our raw features can be split into three categories:

1. Tweet text
2. User description: a short description written by the user in his Twitter account.
3. User covariates:
    a. User_created_at
    b. User_protected
    c. User_followers_count
    d. User_friends_count
    e. User_listed_count
    f. User_favorites_count
    g. User_statuses_count
    h. User_default_profile

    Explanations on each field can be found here.

We have used BERTweet in order to represent both textual fields – tweet text and user description – as two separate 1024 dimensional vectors (after averaging overall tokens). Each vector was reduced to 512 dimensions using UMAP. We denote these representations as X_{text} and $X_{user\_description}$ respectively. Note that all user covariates are either binary or numerical, so the only transformation applied to them was standard scaling. We denote the scaled user covariates as $X_{user\_covariates}$. Finally, the tweet covariates vector X is $\left(X_{text}, X_{user\_description}, X_{user\_covariates}\right)$, of dimension $512+512+8=1032$.

When estimating the high dimensional CATE in which the domain is also a feature in X, we encoded it as a one-hot vector, adding 4 more features to the tweets' representation.

5.2 Methods

In order to estimate the ATE within each domain separately (which is basically CATE conditioned on the domain) we used the S-Learner and T-Learner frameworks. For the S-Learner we used a single linear regression model, and for the T-Learners we used two Lasso (L1 regularized) linear regression models.

When trying to use Lasso for the S-Learner as well we got $CATE = 0$, as the coefficient corresponding to the treatment was 0 due to the regularization. In both methods, in order to evaluate the models' quality, we split the data to 90% train set and 10% test set. The models were trained only on the train set, evaluated on both train and test set (to negate over-fitting), and CATE was computed on the entire data set. We calculated bootstrap confidence intervals with 100 iterations, for the CATE as well as for the models' MSEs (for each relevant pair of treatments).

To estimate the high-dimensional CATE we used the X-Learner framework with Lasso models for generating the pseudo-CATE labels as well as for predicting the labels. The latter two models can be combined in many different ways into the final CATE model (e.g. propensity scores, relative to sample size, etc.). Hence, we decided to rather be focused on interpreting the two models separately, and deduce insights on which covariates determine the high-dimensional CATE of a tweet. Again, we negated over-fitting for all four models by splitting the data into train and test sets as described previously.

5.3 Results

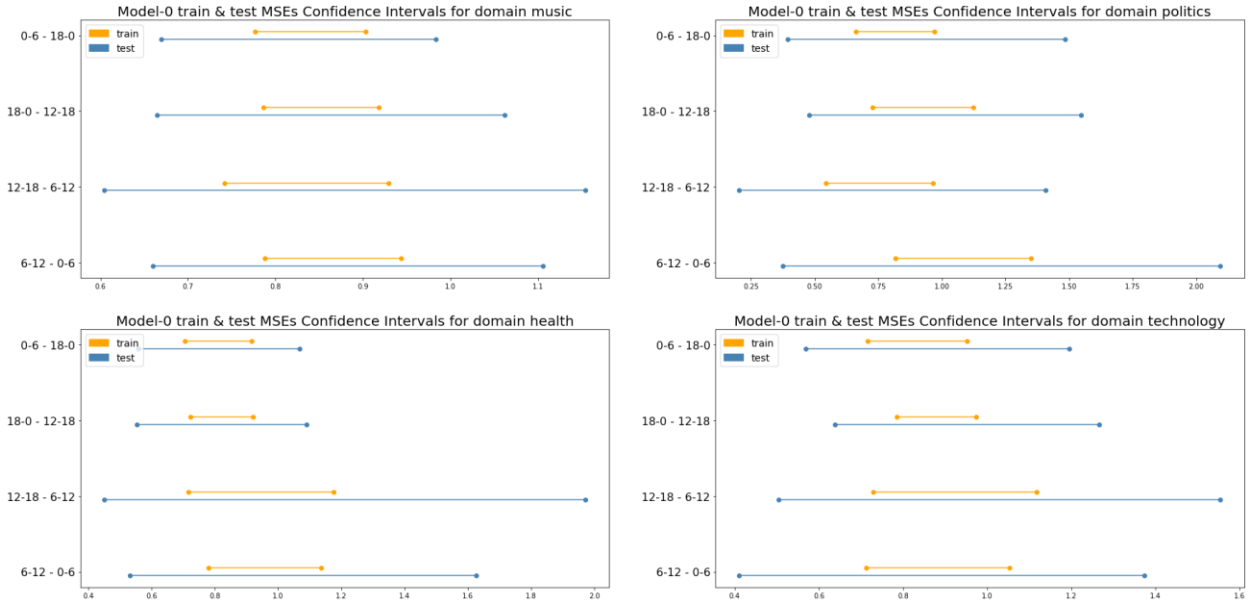
Some results were omitted from the report for brevity. The full results can be found here.

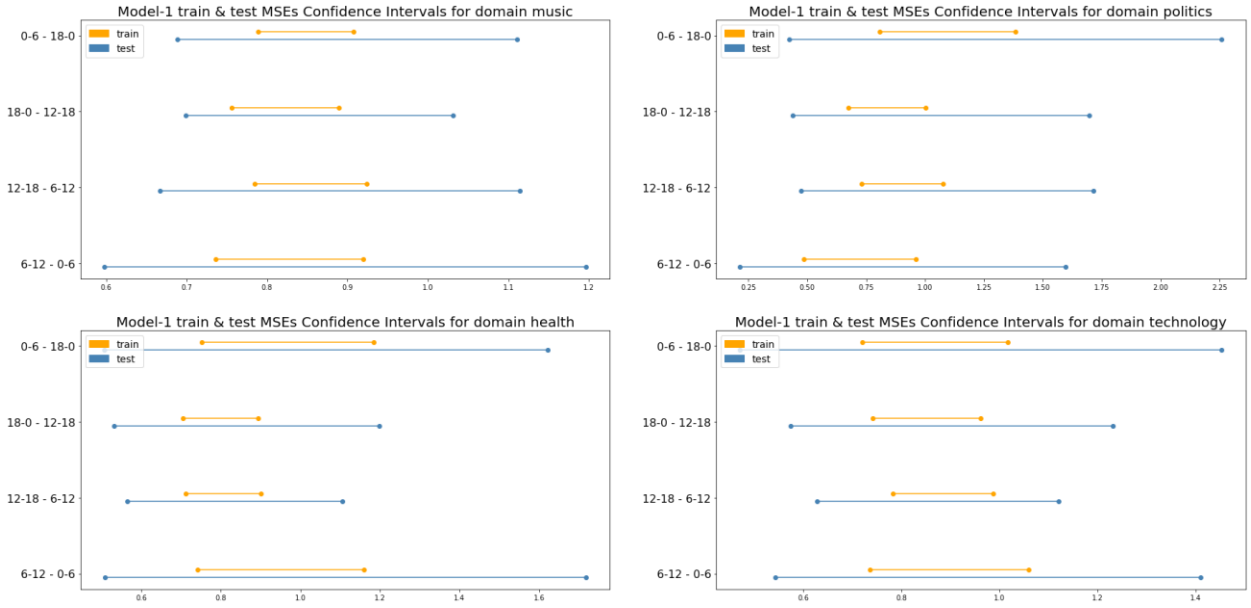
5.3.1 S-Learner

In most cases we observed high variance in the models' MSE on the test set, as well as large gaps between the train and test MSEs. These observations (that were also reproduced when trying different models) imply low reliability of the method. Therefore, we omit these results from the report.

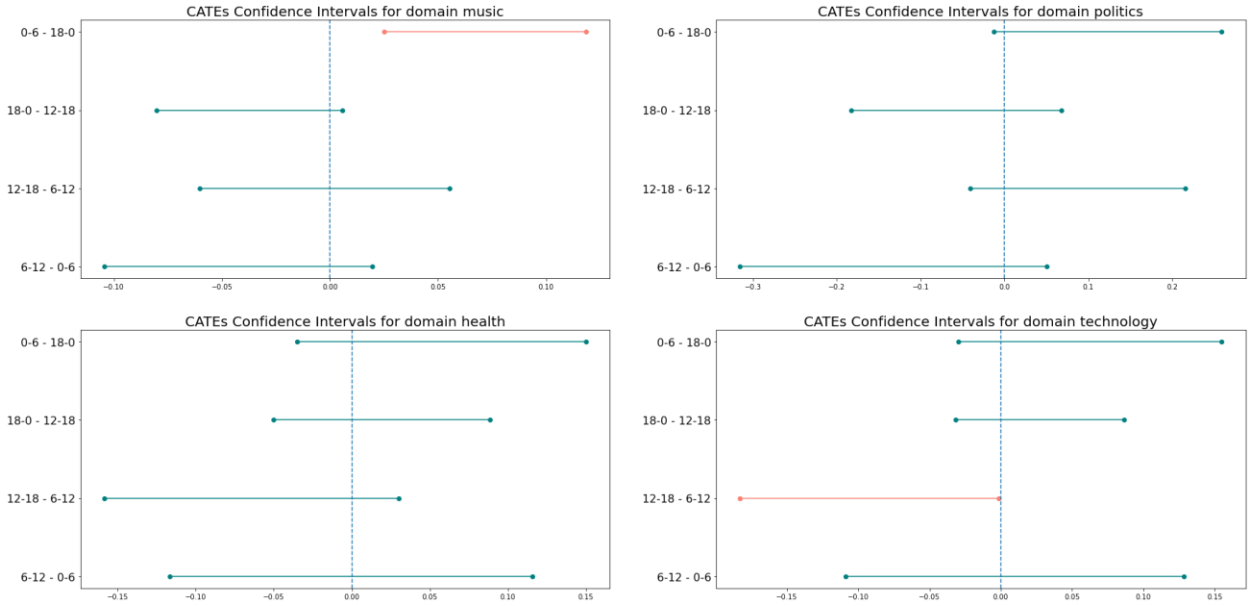
5.3.2 T-Learner

Using this framework we observed lower variance in the models' MSE on both the train and test sets (e.g. shorter confidence intervals). In addition, there were no significant gaps between the train and test MSEs as the confidence intervals overlap, which can be seen in the following graphs for both models:





These results visualized in the above graphs imply high reliability of the T-Learner method, and therefore we used it to estimate the 16 different CATEs (4 domains with 4 treatment pairs each). The following graphs visualize the confidence intervals of these CATEs:



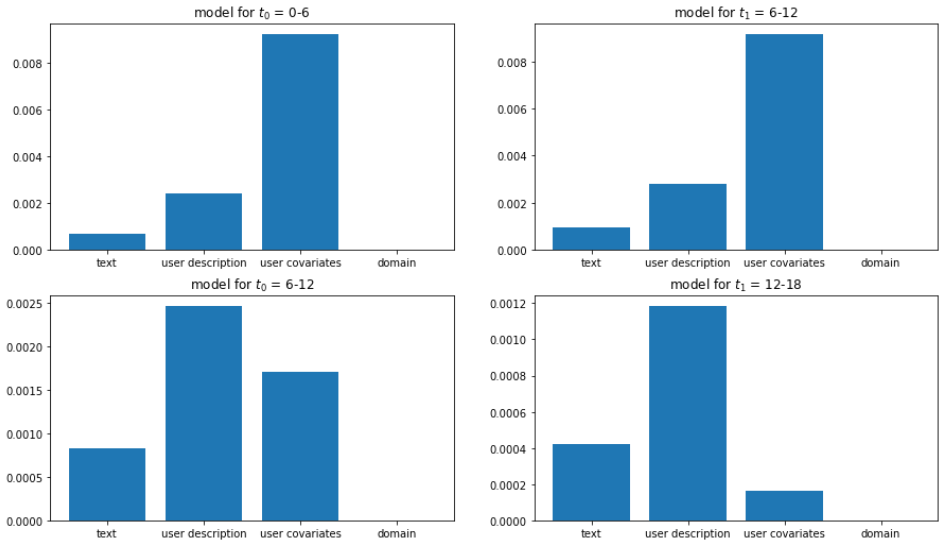
As can be seen above, only two CATEs are significantly above or below zero. In the music domain, we have $\mathbb{E}[Y_{t_{0-6}}|D = music] > \mathbb{E}[Y_{t_{18-0}}|D = music]$, implying larger relative popularity of tweets had they been published after midnight and before dawn, rather than during the evening and early night. In the technology domain, we have $\mathbb{E}[Y_{t_{6-12}}|D = technology] > \mathbb{E}[Y_{t_{12-18}}|D = technology]$, implying larger relative popularity of tweets had they been published in the morning, rather than in the afternoon.

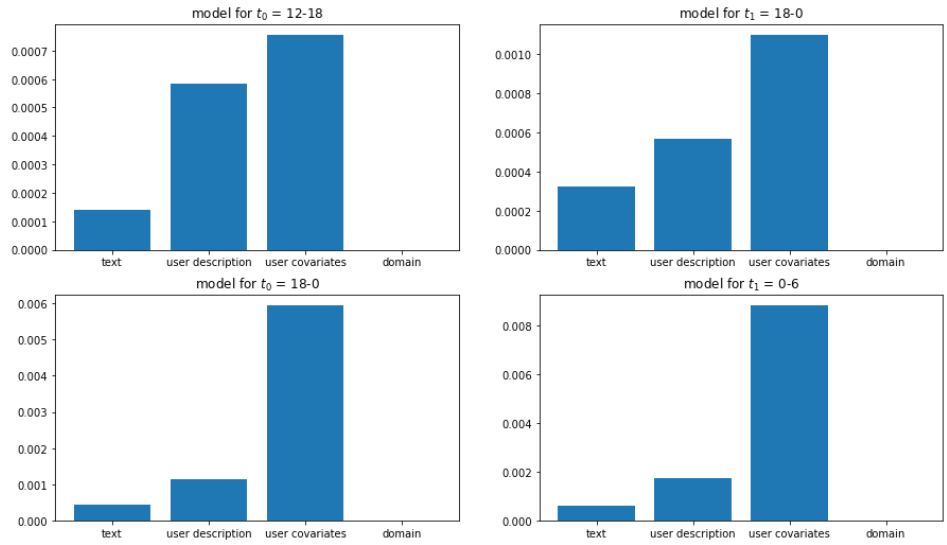
Note that when applying the Bonferroni correction for multiple comparisons (using $\alpha' = \frac{\alpha}{16}$) the above CATEs are also insignificantly different than zero. Thus, in overall we believe that there is no significant causal effect of the publication hour of a tweet on its relative popularity within each domain.

5.3.3 X-Learner

We used the X-Learner framework to estimate the high-dimensional CATE. All four Lasso models resulted in low MSEs both for the train and test sets, implying high reliability of the models. As the output of this framework is a pair of linear models predicting CATE pseudo labels, denoted as $\hat{\tau}_0$ and $\hat{\tau}_1$, we first wanted to identify the most impactful features that affect the prediction of the pseudo labels. This can be used as a proxy to assess which covariates determine the CATE of a tweet.

For each pair of treatments $t_0, t_1 \in \{0-6, 6-12, 12-18, 18-0\}$ we have two linear models, $\hat{\tau}_0$ and $\hat{\tau}_1$, both predicting the CATE and trained on tweets having $T = t_0$ or $T = t_1$ respectively. As explained in the "Data representation" section, we have four groups of covariates: text, user description, user covariates and domain. In order to give an importance score for each group of covariates (with respect to a linear model $\hat{\tau}$) we average the absolute values of the coefficients corresponding to covariates of that group. This way, for each pair of treatments we have defined the groups' importance scores for both $\hat{\tau}_0$ and $\hat{\tau}_1$:





First, we observe that the domain has no importance in determining the CATE of a tweet. This fits our previous results when estimating the CATE conditioning only on the domain. On the other hand, the groups related to the user (both the user covariates and the user description) are the most important. In particular, each of them is more important than the text of the tweet, which might be surprising.

This might have an important implication – when we want to advise a user regarding the best time to publish its tweets, our advice should depend more on the user characteristics rather than on the text of the tweet or its subject.

To demonstrate how these four CATE models can be used in order to find the optimal interval to publish a new tweet, let us consider the following example. Let X be the high-dimensional representation of the tweet, and let $g(X)$ be the function by which we balance between $\hat{\tau}_0$ and $\hat{\tau}_1$ in all pairs of treatments. Suppose that we get:

- $\hat{\mathbb{E}}\left[Y_{t_{6-12}} - Y_{t_{0-6}} \middle| X\right] = g(X) \cdot \hat{\tau}_0^{0-6}(X) + \left(1 - g(X)\right) \cdot \hat{\tau}_1^{6-12}(X) = 1$
- $\hat{\mathbb{E}}\left[Y_{t_{12-18}} - Y_{t_{6-12}} \middle| X\right] = g(X) \cdot \hat{\tau}_0^{6-12}(X) + \left(1 - g(X)\right) \cdot \hat{\tau}_1^{12-18}(X) = -1$
- $\hat{\mathbb{E}}\left[Y_{t_{18-0}} - Y_{t_{12-18}} \middle| X\right] = g(X) \cdot \hat{\tau}_0^{12-18}(X) + \left(1 - g(X)\right) \cdot \hat{\tau}_1^{18-0}(X) = -1$
- $\hat{\mathbb{E}}\left[Y_{t_{0-6}} - Y_{t_{18-0}} \middle| X\right] = g(X) \cdot \hat{\tau}_0^{18-0}(X) + \left(1 - g(X)\right) \cdot \hat{\tau}_1^{0-6}(X) = -1$

This induces the following topological ordering:

$$6-12 \succ_X 12-18 \succ_X 18-0 \succ_X 0-6$$

And therefore we would advise to publish the tweet in the interval $6-12$. However, such an order does not always exist. For example, if all four CATEs have the same sign this results in a cyclic ordering (every interval has a different interval which is preferred).

6 Discussion

In this project we demonstrated and implemented a methodology for estimating the causal effect of the publication hour of a tweet on its relative popularity. This required dealing with several challenges. First, handling continuous treatment which had to be discretized into hour intervals. This introduced a tradeoff between the plausibility of the SUTVA assumption and the number of CATEs to be estimated, both determined by the length of the intervals. In particular, the wider the intervals are, the less plausible the SUTVA assumption is. On the other hand, in this case, we have to estimate less CATEs and for each treatment we have a larger sample size, which induces more accurate results. This is a tradeoff between identification and estimation.

The next big challenge was the variety and complexity of the covariates used for the models. We had to combine textual, numerical and categorical features into a single tweet representation. In this work we have used linear models, which might not capture accurately the relationships between the covariates, the treatments and the potential outcomes. Further research might be focused on applying more complex models within the scope of this methodology.

When estimating the causal effect conditioned only on the domain, we couldn't identify a significant causal effect of the hour on the relative popularity. A possible explanation is that within very general domains the causal effect is different for different specific sub-domains, but on average they cancel each other. For example, within the music domain, it might be the case that tweets about classical music would gain more popularity had they been published in the morning, while tweets about techno music would gain more popularity had they been published late at night. These two non-zero CATEs might cancel each other when considering the entire music domain.

When considering the high-dimensional CATE we used the X-Learner framework, which has a built-in limitation. In the first phase we generate pseudo labels, which might be inaccurate, and are then used training the models for predicting the labels. If these models are inaccurate themselves, we end up with two types of errors which might enhance each other. In addition, we have seen that the user covariates play a major role in determining the high-dimensional CATE of a tweet. This might imply that by collecting more user data, one can more accurately estimate the tweet's CATE.

Finally, a new research direction might focus on developing a new methodology which will be more business-oriented, i.e. that will give a more accurate answer to the question – what is the best time to publish a tweet? One possible high-level idea is hierarchical CATE estimation. Given a tweet X, first estimate the $\mathbb{E}\left[Y_{t_{0-12}} - Y_{t_{12-0}} \middle| X\right]$ to determine whether the tweet should be published in the first or second half of the day. Suppose that the CATE model produced $\widehat{\mathbb{E}}\left[Y_{t_{0-12}} - Y_{t_{12-0}} \middle| X\right] > 0$, which means that the preferred interval is $0-12$, therefore the next step would be to estimate $\mathbb{E}\left[Y_{t_{0-6}} - Y_{t_{6-12}} \middle| X\right]$. Now again we decide which interval is preferred, and continue the process until reaching an interval as small as we wish. Notice that the number of CATE models required for this method grows exponentially in the desired interval length. If we wish for the models to be independent of each other, then we must use an

exponentially growing number of large-enough datasets. In addition, some sort of multiplicity adjustment might be needed somewhere along the way.

We believe that all of the above conclusions show fascinating insights, and although the results in this project are limited to specific dates, location, language and domains, the methodology can easily be used on different datasets with different characteristics. With the fast growing and constantly evolving social networks technologies, deepening in those research directions may lead to a huge impact on the way we use Twitter and other social media in the future.