Does Tiktok show viewers the content relevant to them?

Ekaterina Fedorova

Department of Economics

University of California, Berkeley

Undergraduate Honors Thesis

Advised by Dr. Clair Brown

Abstract

On social media, how users interact and what content they see characterizes their experience. How interaction occurs and content is shown to the viewer varies across platforms and is determined by undisclosed algorithms. This paper explores how TikTok's algorithms curate videos and how users engage with the videos and each other. The analysis is done in two parts. First, how four different professional groups hashtag their content is compared in order to understand whether groups communicate with the algorithm in distinct ways. I find that while some tags are commonly used across all types of content, the remaining tags form clusters, which represent the interests of the four pooled groups. Second, a statistical analysis of a particular hashtag is conducted. Using video posting date, video length, and video description, I construct a model to predict whether a TikTok video is relevant to the hashtag. Increased hashtag relevancy is valuable to both users looking for specific topics through hashtag search and also to social media platforms that want to hold the attention of users for as long as possible. Empirically, video relevancy can be predicted with 94% accuracy. This thesis contributes to our understanding of user interactions on social media by demonstrating that shared topics are a driving factor in TikTok community interactions, and it also provides an accurate prediction model that is able to classify content by hashtag relevancy using only anonymous video attributes.

1 Introduction

Social media platforms and the algorithms that govern the content they recommend to viewers are important drivers in information dissemination and community formation. As such, academic studies of the content shown and to whom it is recommended contribute to our understanding of how information spreads online. Well-known and perhaps even infamous, TikTok is still a new application on the social media scene and even newer to academic social network analysis. As a result, TikTok has very little existing literature, despite its status as a major rival of Instagram [1] and its achievement in 2020 as the most downloaded app of the year[1].

With such massive user counts and a young audience, [13] what creators see and are able to post within their communities has significant influence both for users and the population not on the app. Political movements, for example, are a key aspect of the TikTok platform. In 2020, political activists on the app instigated a large false registration drive for the Trump rally in Oklahoma[2]. Whether anyone sees this information is determined by whether it shows up in their feeds. On all platforms, users may see two different types of content suggested: content from followed topics/people, and content that is algorithmically recommended based on the user themselves. Across social media, all algorithmic recommendations are consistently opaque, [18] [4] but TikTok's opaque algorithmic recommendation is also the biggest driver of its video curation. [15] As a result, the TikTok model is one that distinctly centers its user experience on algorithmically recommended content above all else. This can be most clearly seen in the platform's emphasis on the algorithmically curated "For You Page" over the alternative video feed, which is a compilation of followed topics and creators. The For You Page (also called FYP) is an algorithmically curated feed of videos that is "a personalized video feed based on what you watch, like, and share."[3] It is the default home page and, amongst users, synonymous with the app itself. The heavy implementation of content recommendation algorithms on TikTok means many creators desire to make algorithmically favorable content (i.e., primed to be spread by the recommendation) in order to achieve popularity. At the same time, making content that is likely to be recommended to others is very difficult when the underlying algorithm is hidden.

1.1 Motivation

This thesis aims to conduct an investigation into user-algorithm interactions in order to better understand what methods creators employ to make their content primed for recommendation. When these methods are successful, users are better enabled to spread and consume information and form impactful communities

[1] https://www.socialmediatoday.com/news/tiktok-was-the-most-downloaded-app-in-2020-according-to-new-data-from-app/591910/

[2] https://www.nytimes.com/2020/06/28/style/tiktok-teen-politics-gen-z.html

[3] https://apps.apple.com/us/app/id835599320

because their content is recommended to more viewers. However, before success can be assessed, the way a user communicates to the algorithm must first be understood. The light shed on these methods is the main contribution of this paper.

Hashtags represent an accessible way for users to communicate with the algorithm by letting the platform know important topics in the video's content. They are also a way by which users can participate in trends and make content that might be recommended for its recent popularity. This paper addresses both of these aspects in two separate analyses. First, four large professional groups' (#ArtistsofTikTok, #CopsofTikTok, #NursesofTikTok, and #TeachersofTikTok) hashtag usage is compared in order to determine whether different communities use distinct patterns of tags in order to communicate with the platform. I find that while some tags are commonly used across all types of content, the remaining tags create clusters, which represent the differing interests of the four pooled groups. This is an important finding because it indicates creators can use many different community-associated hashtags to spread information to the relevant group. If done successfully, knowledge of events, politics and more can be spread to a target audience very efficiently.

Second, statistical analysis on the previously trending tag #NoNuanceNovember is conducted. This tag represents a trend in which users share their controversial opinions without nuance. As a result, many participating videos gained traction and discourse on the app. Users who want to search through the tag to participate in the discussion, however, would need to scroll through the many irrelevant videos that used the tag simply to capitalize on perceived algorithmic favorability of the trend. By using video posting date, video length, and video description, I create a machine-learning model that is trained predict whether a TikTok video is relevant to the hashtag. Video relevancy to the #NoNuanceNovember challenge can be predicted with 94% accuracy using only these anonymous video attributes. Such prediction of relevancy is valuable both to users who want to be able to view more challenge-participatory videos under the hashtag, as well as to social media platforms that want to hold the attention of users for as long as possible. Additionally, this accurate classification does not use any predictors derived by video analysis software, so it avoids the bias involved in algorithmic face and body recognition.

The first analysis adds to our understanding of user-algorithm interactions on social media platforms by demonstrating that shared hashtags are a driving factor in TikTok community interactions, and the second provides an accurate prediction model that is able to classify content by hashtag relevancy using only anonymous and text-based attributes.

2 Background and Literature

Despite differing levels of user-algorithm interactions across all platforms, other studies' empirical strategies into algorithmic awareness and hashtag usage have served as the inspiration for this thesis. There exists a broad range of literature on the general subject of human-algorithm interaction, spanning many foci as algorithmic suggestion can impact ads, creators, specific groups of users, and even potentially those not using the platform by the type of (mis)information that is spread. Looking at multi-platform human-algorithm interaction, Jonas and Burrell examine the consequences of algorithmic bias by region on more general internet users. [9] In order to create a fairer system, the paper seeks to understand the impact of "good" and "bad" use-pattern classification on those in regions which have been historically discriminated against. The authors find a major need for a reconsideration of existing security systems to remove such industry-wide bias.

While the question this thesis hopes to answer does not concern regional algorithmic classification specifically, the heavy use of community-forming recommendation algorithms on TikTok also has a track record of being inequitable[4]. As this becomes increasingly clear, users' interactions with the platform change significantly. Some communities have created events (usually called blackouts or awareness days) during which all users are encouraged to interact with content created by that community as well as refrain from making content if they are not a part of said community. A recent example of this is International Day of Persons with Disabilities, when the disability community of TikTok encouraged users to boost content made by disabled creators and rallied around the tag #DisabilityAwarenessDay[5]. Under this tag, users outside of the community could find content that explained the event and promoted disabled creators. These grassroots events and their hashtags represent community-algorithm interaction because their purpose is to boost the recommendation of the community's content by getting more user engagement. Research on the use of hashtags following or during major events to centralize information has been done across other platforms [17] [10] but not yet TikTok.

The dynamic functions of hashtags have long been covered across many platforms. [3] Most importantly, researchers conclude that hashtags are not simply helpful search functions for users to categorize or find information but a novel form of metacommunication between users. [5] On TikTok, this metacommunication exists not just as a communicative link from user to user, but as a link between user and algorithm. Because the dissemination of content depends on a recommendation algorithm, a video's hashtags also allow creators to communicate which communities should see certain videos. Creators can tag professional groups like #TeachersofTikTok and expect the algorithm to recommend the content to those in that community.

[4]https://www.bbc.com/news/technology-50645345
[5]https://www.allure.com/story/international-day-of-persons-with-disabilities-celebrate-support

Evaluating topics present within collections of hashtags has long held a place in social network analysis. Topic modelling has been significantly integrated with community detection on Twitter in order to understand how these different concepts can overlap. [7] [19] [2] Many of these methods of text analysis are slightly more complicated in their potential application to TikTok data as, unlike Twitter, the driving force of the TikTok platform is the video format. While this thesis focuses almost exclusively on information gleaned from text-based attributes, a topic modelling project could also analyze the multi-media aspect of the application. Existing research on U.S. political content from the app considers this idea. [13]

As this thesis focuses on user-selected hashtags, a final section of relevant hashtag literature is that which considers the relevancy of a tag to a section of content. On Twitter, this is a mature question. Prediction of hashtags and hashtag suggestion is an ever-evolving area of study. [6] [14] [11] Through the lens of hashtags' search functions, predicting applicable tags should be a valuable ability to social platforms as even content without "proper" tags can be accurately categorized for search. This literature focuses on finding a relevant hashtag for content while the focus of this thesis is the adjacent question of assessing which content already tagged is truly relevant to that hashtag. Other studies take what is essentially the reverse of this question and attempt to predict an existing hashtag's popularity or lifetime. [16] [12] Such studies are applicable to TikTok's trending tags such as the #NoNuanceNovember challenge, however these also do not fully address popularity in terms of tag-relevant and tag-irrelevant content creation.

3 Professional Community Distinctions in Hashtag Use

3.1 Data Collection

This paper takes advantage of the popularity of community referencing tag structures such as "#—ofTikTok" in order to collect content representative of those who self-identify to be in that community. In order to keep selected communities comparable, the focus of this analysis is on professional groups. In December 2020, I explored the TikTok hashtag search function for all major professional groups under hashtags that used the structure "#—ofTikTok", "#—onTikTok", "#—TikTok", and "#—Tok" such that the blank space was a profession or professions. I found the four most popular distinct professional communities were #ArtistsofTikTok (8.1B views), #CopsofTikTok (5.6B views), #TeachersofTikTok (5.1B views) and #NursesofTikTok (1.1B views). Using the TikTok API package created by David Teather[6], I used Python to run code in order to web scrape all available public content under each of these tags. Code files are available on my GitHub[7]. The scraped content represents all videos a TikTok user could view under the hashtag

[6] https://github.com/davidteather/TikTok-Api
[7] https://github.com/ek8terina

if they were to scroll through the searched tag at approximately the same time as the video collection. The earliest observation comes from #ArtistsofTikTok on April 15th, 2019 while the latest comes from #TeachersofTikTok on December 9th, 2020, the approximate date of data collection. There are 4694, 3890, 4343, and 4600 observations under each tag respectively, or 17527 videos when pooled.

3.2 Data Preparation and Method

Upon scraping, data was coded 1, 2, 3, or 4 to represent each community as determined by the tag from which each video was scraped (#ArtistsofTikTok, #CopsofTikTok, and so on). Using R and regex functions, the description of each video was cleaned to create a list of hashtags used in each video, the number of tags used by each video, and the number of non-tag characters used by each video. The character limit of a TikTok description is 150 so the latter two attributes must be 150 or less by definition. The latter variables alone have a limited ability to portray the hashtags used by each group. An observation with the description, "#NursesofTikTok #nurse" and an observation with the description "#ArtistsofTikTok #art" will have the same values. They both have 2 hashtags and no non-hashtag characters. Despite this, the human eye would know to put observation 1 in the nurses' professional group and observation 2 into the artists' professional group.

In order to numerically represent differences in the hashtags themselves, every hashtag that appeared at least once in the pooled data set was converted into a dummy variable {0,1}. A video had a value of 0 if it did not use that hashtag and a value of 1 if it used the tag. This created 16856 dummy variables for the 17527-observation data set with each dummy representing a unique hashtag that appeared in the pooled data. The hashtags #ArtistsofTiktok, #CopsofTikTok, #TeachersofTikTok, and #NursesofTikTok were then taken out as they represent the 1, 2, 3, 4 coding already added to the observations. Additionally, all hashtags that appeared only once in the dataset were also taken out. A hashtag that appeared only once was deemed unlikely to be indicative of any particular group and is thus unhelpful to forming community distinctions. This left 6163 tags that each appeared in the dataset more than once and were not the community tags themselves. Finally, the pooled dataset was also cleaned of the top three tags, #fyp, #foryou, and #foryoupage. The use of these algorithmically referencing tags dwarf the counts of all other tags in the dataset and are used across communities so they do not represent any meaningful community-specific use pattern. In order to achieve the goal of clusters which recreate the 1, 2, 3, 4 group coding, these had to be removed.

Graphical representations greater than 3-dimensions are not easy to interpret to the human eye, however this analysis considers four professions. In order to visualize potential clusters, the pooled observations were

further split into three datasets each of three groups or less.

Dataset 1: All videos of #ArtistsofTikTok and #CopsofTikTok

Dataset 2: All videos of #TeachersofTikTok and #NursesofTikTok

Dataset 3: All videos of #ArtistsofTikTok, #TeachersofTikTok, and #NursesofTikTok

Each dataset was additionally cleaned of tags that did not appear or only appeared once in the respective set. The final data sets come out to have the following hashtag variables, Dataset 1: 3674 tag dummy variables, Dataset 2: 3221 tag dummy variables, Dataset 3: 5048 tag dummy variables.

In order to create reasonable cluster visualizations, these thousands of {0,1} indicators needed to be compressed while preserving as much of the information from all three to five thousand of them as possible. First, for each dataset, a distance matrix using the binary distance metric, the Jaccard Index (Equation 1), was calculated. This distance matrix would represent the dissimilarity of each TikTok video's tag use pattern to every other video's tag use pattern. If tag use really is distinct between professional communities, it would be expected that videos within the same group are fairly similar to each other while those in a different group are more similar to each other than to the other group.

$$J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

Once a distance matrix for each dataset was calculated, multidimensional scaling (MDS) was done on all three datasets. The idea behind MDS is to represent a dataset that is in p dimensions instead in k dimensions (such that $k < p$) while still keeping approximately the original dissimilarity between observations. Equation 2 represents this concept as a pair-wise distance approximation.

$$d_{i,j} = \| z_i - z_j \|_2 \tag{2}$$

where:

$d_{i,j}$ = the distance between 2 observations in p dimensions

z_i = an observation scaled to k dimensions

z_j = an observation scaled to k dimensions

Using classical metric multidimensional scaling (CMDS) (cmdscale in R), the Jaccard distance matrices for each dataset were scaled to lower dimensionality. Datasets 1 and 2 were scaled to two dimensions while Dataset 3 was scaled to three dimensions. Classical scaling seeks to minimize a particular stress function (equation 3). [8] When the dissimilarity matrix used is a Euclidean distance metric, CMDS scaling is equivalent to principal component analysis (PCA) and is therefore a simple linear dimension reduction. As this analysis was conducted using a non-Euclidean distance metric (Jaccard Binary Distance), CMDS leads

to a similar outcome, but with a nonlinear mapping. The scaled dimensions of each dataset were used to create visualizations of the dissimilarities in tag use across all observations.

$$S_C(z_1, z_2, ...z_N) = \sum_{i,i'} (s_{ii'} - \langle z_i - \overline{z}, z_{i'} - \overline{z} \rangle)^2 \tag{3}$$

where:

$s_{ii'}$ = pairwise similarities between 2 observations in p dimensions

z_i = an observation scaled to k dimensions

z_i' = another observation scaled to k dimensions

N = number of observations

3.3 Results

The results yielded three figures (one for each dataset of different pooled communities). Each figure has axes representing the reduced dimensions (referred to as eigenfunctions due to the calculation methodology) and each point is one TikTok video from the pooled set.

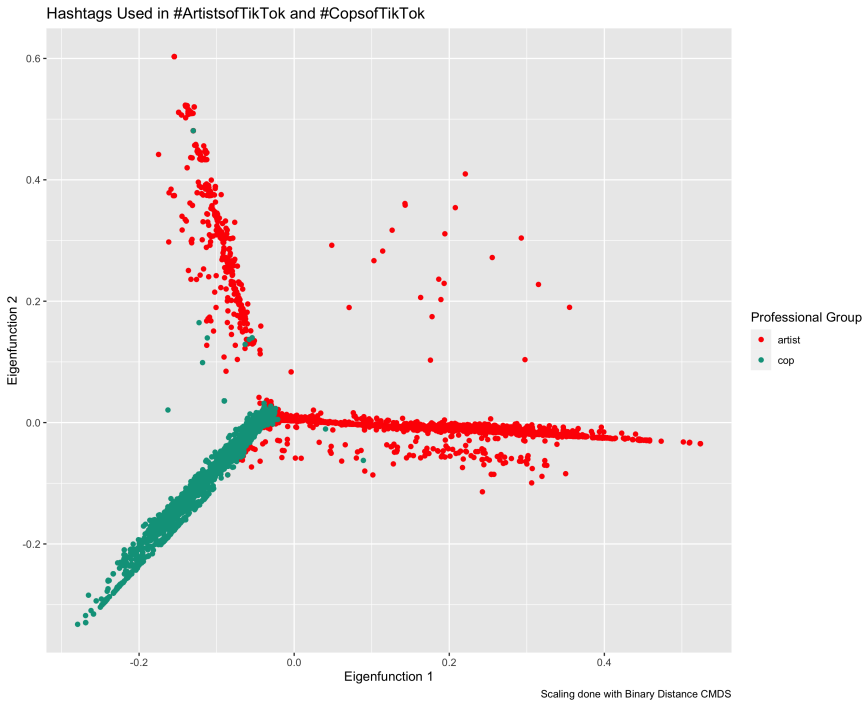

Figure 1: Dissimilarity in hashtags used by #ArtistsofTikTok and #CopsofTikTok professional groups

The first dataset's binary distance CMDS constructed Figure 1. The visualization includes two axes to represent the 3674 tag dummy variables reduced to two dimensions. Each point is a TikTok video, and its

tag use dissimilarity to all other videos is represented in its coordinates within the reduced dimensions. The nature of CMDS is to preserve the dissimilarity in the original variables as much as possible, so it can be interpreted that Figure 1 represents differences in hashtag use pattern across the pooled dataset.

Figure 1 shows three clusters each connected to one another at the (approximate) origin. These three clusters map roughly to the professional groups. Two of the clusters are videos from #ArtistsofTikTok while the remaining cluster consists of videos from #CopsofTikTok. When points are very close to each other, this means that the videos have very similar tag use patterns, while those very far from each other are very dissimilar. The common source point indicates the existence of hashtag use patterns that are used across both communities. These tag patterns have high similarity to each other but also a low similarity to those easily classed into one of the groups. The collection tags (#ArtistsofTikTok, #CopsofTikTok, etc.) were taken out, so this means an example of such an observation could be one whose tags are just "#viral". Figure 1 visualizes why tag pattern "#viral" could not be put into a correct community. This tag pattern would be similar to videos in #ArtistsofTikTok, as well as to videos in #CopsofTikTok. Aside from the common source point, videos clearly cluster by dissimilarities in tag use. A video from the #ArtistsofTikTok community that uses the tag pattern #ArtistsofTikTok, #art, #resin, and #viral should be correctly deemed dissimilar to any video that comes from #CopsofTikTok where hashtags such as #art and #resin appear very infrequently, if at all.

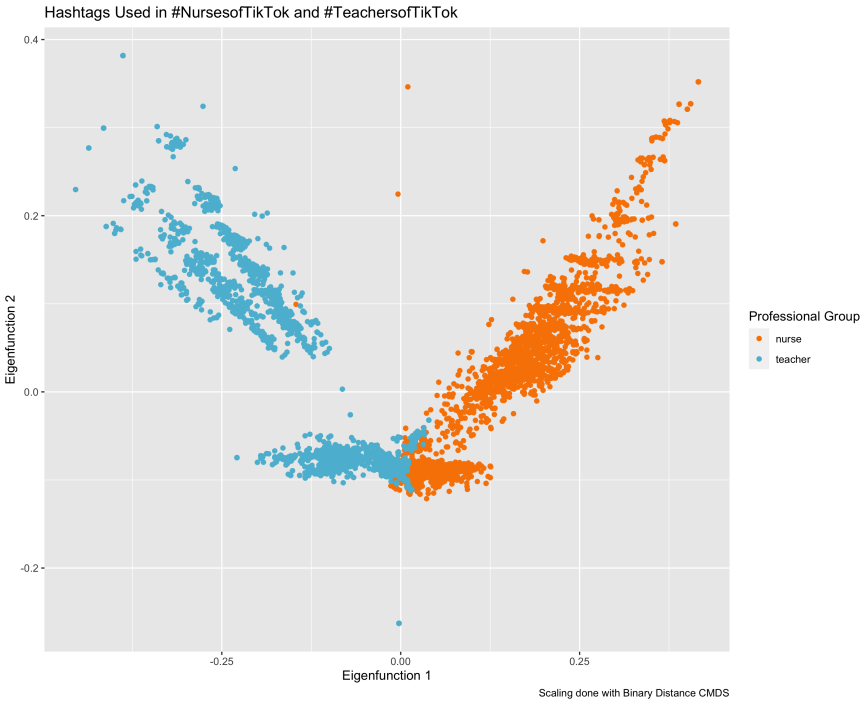

Figure 2: Dissimilarity in hashtags used by #NursesofTikTok and #TeachersofTikTok professional groups

Figure 2 comes from the binary distance CMDS of dataset 2, which is the pooled observations from #NursesofTikTok and #TeachersofTikTok. Axes one and two represent the 3221 tag dummy variables scaled down to two dimensions while preserving dissimilarity

Although the shapes of the clusters seen in Figure 2 differ significantly from Figure 1, the key aspects of the two are the same. Both figures show an area of mixing between the 2 professional groups and, beyond this, can be split quite cleanly into communities. Once again, this implies that even without the community tags themselves, the remaining tag patterns split generally into those that cannot be put into one group (#viral, for example) and those that are present almost exclusively in one group.

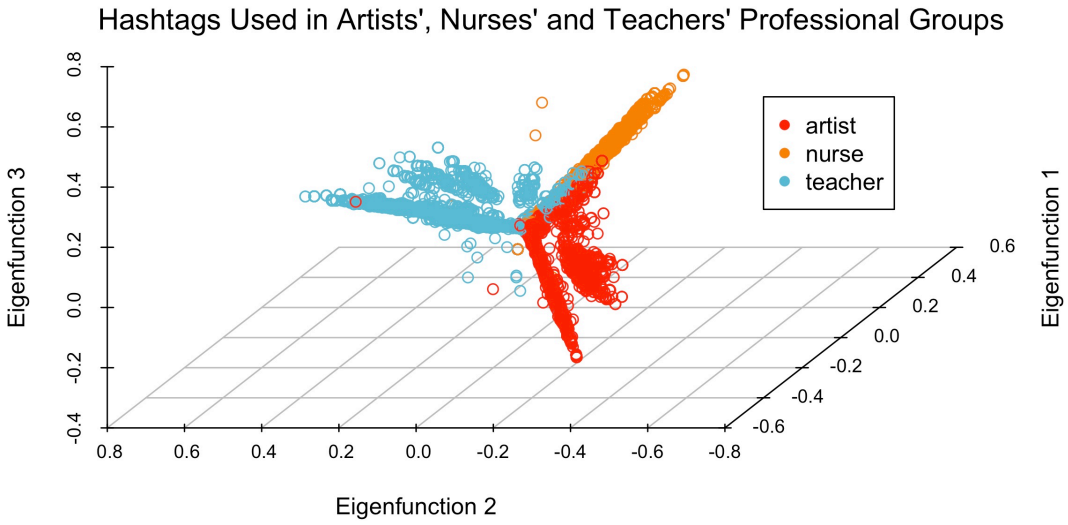

Figure 3: Dissimilarity in hashtags used by #ArtistsofTikTok, #NursesofTikTok, and #TeachersofTikTok professional groups

The third dataset is pooled from #ArtistsofTikTok, #TeachersofTikTok, and #NursesofTikTok and scaled to 3-D from 5048 dummy variables. Unlike Figures 1 and 2, this visualization uses three dimensions in order to enable the three professional groups to visually cluster as much as possible.

Figure 3 shows a shape that looks like a 3-dimensional version of Figures 1 and 2. It has a common source point, from which all clusters spread out and then split into three colors quite purely, even if each color contains multiple clusters. This serves as a confirmation that this scaling method could be applied to more than just two professional groups, and it implies that additional groups can still be separated from each other by tag use dissimilarity. A tag pattern like "#nurse, #viral" is one that would be expected in neither the artist nor the teacher professional clusters because its use of #nurse should make it far more similar to other content in the #NursesofTikTok community.

3.4 Community Tag Discussion

When creators tag one of the community hashtags like #ArtistsofTikTok, this enables the creator to identify content as relevant to that community. Therefore, the creator tells the algorithm that their content should be recommended to artists or those interested in art. The results across these three visualizations point to the fact that within these TikTok professional communities, the remaining hashtags used are also heavily associated with that community. This implies that users can signal the same community relevancy to the algorithm through many different tags and similarly, that a single community has several community-specific interests or tags.

This contributes to our understanding of how communities on TikTok function. Figures 1, 2, and 3 highlight that there are recognizable patterns in hashtag usage within the self-identified professional communities. This means even though users tell the algorithm how their content should be recommended through one tag (#ArtistsofTikTok, #CopsofTikTok, etc.), they often include additional signals saying the same thing. A description may include "#ArtistsofTiktok" and "#artist" even though both of these point the content to the same community. Such patterns of tag use are important to understand because they imply creators are aware of how their tags influence who is the content's audience, and that creators use this to their advantage. They tailor tags such that they will increase the likelihood of getting to the targeted community by using multiple relevant tags.

From an information dissemination perspective, recognizing that both activists and those with mal intent can use this tag awareness to their advantage is critical. Hashtags are not the only attribute used in recommendation, but they are an important one. Awareness of a particular event, for example, can be spread by being associated with large swaths of content that uses several community tags to ensure its recommendation to target groups. The previously mentioned tag, #DisabilityAwarenessDay, can be seen on content with community tags like #DisabledTikTok, #DisabilityTikTok, and #Disability. This method of tag grouping could be used to spread political information, activism, or even hateful rhetoric to specific audiences.

4 Analysis of Relevancy and Promotion of Content

4.1 Data: Collection and Background

In addition to the standard hashtag systems used across all social media platforms, a type of hashtag unique to TikTok is the challenge tag (a.k.a. trending tag). Challenge tags are a special category of hashtags promoted by the platform as a way to directly engage with users by encouraging the creation of a video following a

specific set of instructions. The key aspect of challenge tags, as opposed to tags like #ArtistsofTikTok or #fyp, is that challenge tags are prominently displayed on the platform's discover page. Figure 4 showcases three different challenge tags from the user's perspective.

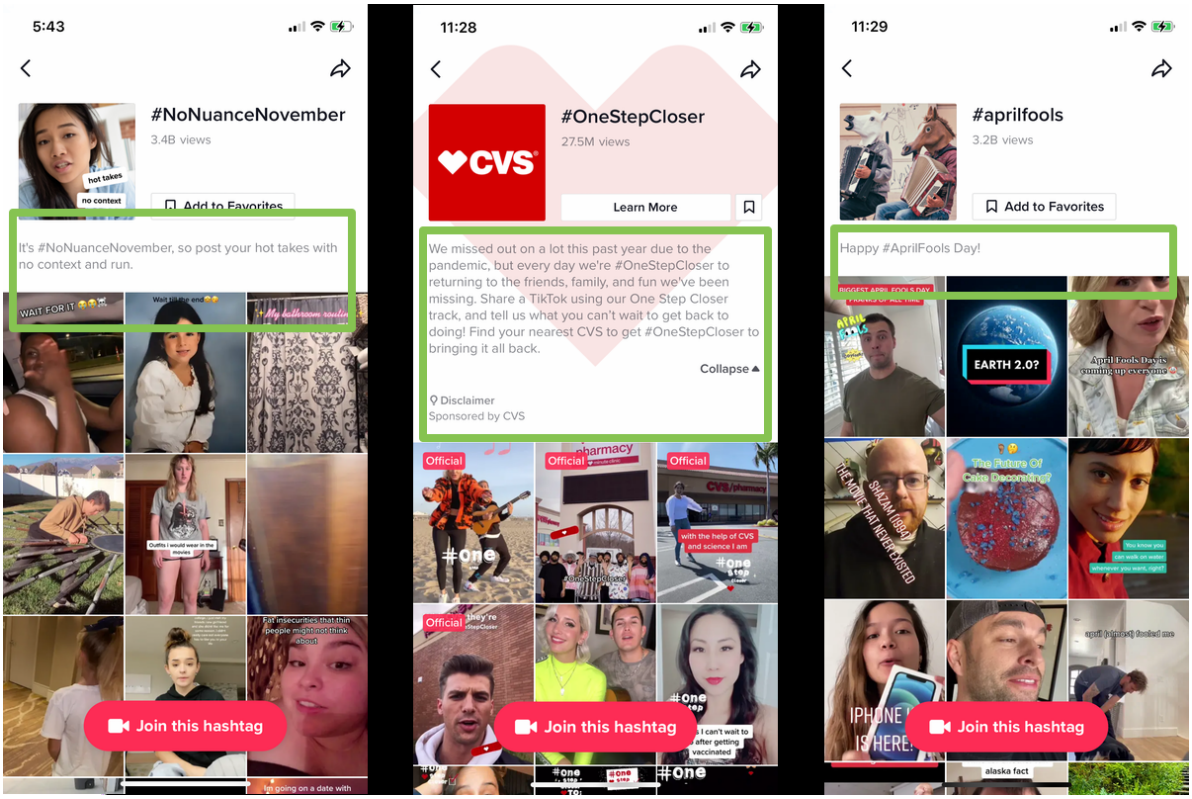

Figure 4: Screenshots of 3 challenge tags: #NoNuanceNovember, #OneStepCloser, and #AprilFools. Descriptions of each challenge are outlined in green.

The three challenge tags of Figure 4 highlight the unique aspects of trending tags. Most importantly, they have a tag description above the scroll of videos (outlined in green). The challenge's description explains the meaning of the tag, how to participate, and the brand sponsorship, when applicable. Normal tags have neither this invitation to participate nor the implication of being "trending". As a result, the use of a challenge tag appeals to users who want to genuinely participate in the trend and also to users who want to use the tag on irrelevant content in hopes of capitalizing on the trend for a boost in recommendation.

Because of these two types of creators, trending tags are often filled with irrelevant content that a user must scroll through in order to watch the genuinely challenge-participatory videos. This analysis focuses on the previously trending tag #NoNuanceNovember (NNN). This trending tag has the description:

> "It's #NoNuanceNovember, so post your hot takes with no context and run."

This instructs users to create a video of their controversial opinions with no room for nuance. Because this

prompt led to many discussions and additional opinions shared, it is likely users may want to scroll through this challenge tag in order to engage with the content (e.g., liking, sharing, commenting). Unfortunately, due to the tag being a trend, many videos have used the hashtag on content that has nothing to do with the instructions.

As users would want to see more challenge relevant content when scrolling through the tag, and the platform would want to keep users' attention as long as possible, it is desirable to be able to predict NNN challenge relevancy based on simple video attributes. In order to train such a model, all content from the NNN challenge tag was scraped in early March 2021 using David Teather's TikTok API[8]. Coding was done in Python and is available on my GitHub[9]. The raw NNN dataset contains 4189 observations, with the earliest video posting date being November 7th, 2020 and the most recent (as of collection) being March 5th, 2021. These observations represent all videos a user could scroll through under the tag as of March 5th.

4.2 Data: Variables Created

The main video attributes collected from the #NoNuanceNovember scrape were video description, time of posting (in seconds), length of video (0-60 seconds), and link. The goal of this analysis is to train a supervised learning method, meaning one with known inputs (video attributes) and outputs (relevancy classification), in order to predict and test. As such, it was necessary that the dataset have the classification of challenge relevant and challenge irrelevant on all observations. In order to create this {0,1} dummy variable, I followed the link on each of the 4189 TikTok videos and watched each one to code it as 0 (challenge irrelevant) or 1 (challenge relevant). For coding, emphasis was placed on the sharing of at least one opinion with little or no explanation. Videos deemed potentially 0 or 1 in a first round of coding were marked for recheck and reevaluated as a group after all others were coded. This coding decreased the dataset to 4145 videos because 44 videos had been deleted or made private since the initial scrape and could not be watched. Challenge relevancy coding revealed a class imbalance in the data with approximately 79% (3284) of videos being challenge irrelevant. While this makes prediction of relevancy slightly more difficult, it only makes it all the more important for the benefit of users who likely would not want to scroll through so many irrelevant videos.

In addition to the creation of a challenge relevancy dummy variable, using R and regex, video descriptions were cleaned in order to create variables representing the number of hashtags used and the number of non-tag characters used. Although both of these numbers capture valuable information about the video, neither are able to fully portray which tags are included. In order to represent this empirically, a dimensionality

[8]https://github.com/davidteather/TikTok-Api
[9]https://github.com/ek8terina

reduction method very to that which was described in 3.2 was conducted.

Every unique hashtag used in the dataset was coded to be a {0,1} variable, 0 if a video used the tag and 1 if that tag was not used. This created 5888 tag dummy variables for the 4145 videos in the #NoNuanceNovember challenge tag. One of these tag dummy variables is the #NoNuanceNovember challenge tag itself, which is a vector of ones as it is used in all videos by definition. It would be unreasonable to use all 5888 tag dummy variables in challenge relevancy prediction as this could increase the computational power needed significantly.

In order to alleviate this issue, the same multidimensional scaling (MDS) method was used on these tag dummy variables as was used in section 3.2. The aim of MDS is to approximate the dissimilarity of variables in very large p dimensions to a much smaller k dimension. Equation 2 (section 3.2) represents this goal. If it is the case that challenge relevant videos have tag use patterns similar to each other and dissimilar to challenge irrelevant videos, MDS will keep this dissimilarity, but it will represent it in two or three dimensions rather than 5888.

A dissimilarity matrix was calculated using the binary distance metric, the Jaccard Index. Equation 1 (section 3.2) shows this distance in set notation. Using classical metric multidimensional scaling (CMDS) (see Equation 3 in section 3.2) this dissimilarity matrix was non-linearly mapped to three dimensions. These three dimensions represent the pairwise dissimilarity between #NoNuanceNovember videos' tag use patterns, but in only three constructed variables rather than all 5888 {0,1} tag dummy variables. Figure 5 is a visualization of the NNN videos across the first two of the three scaled variables.

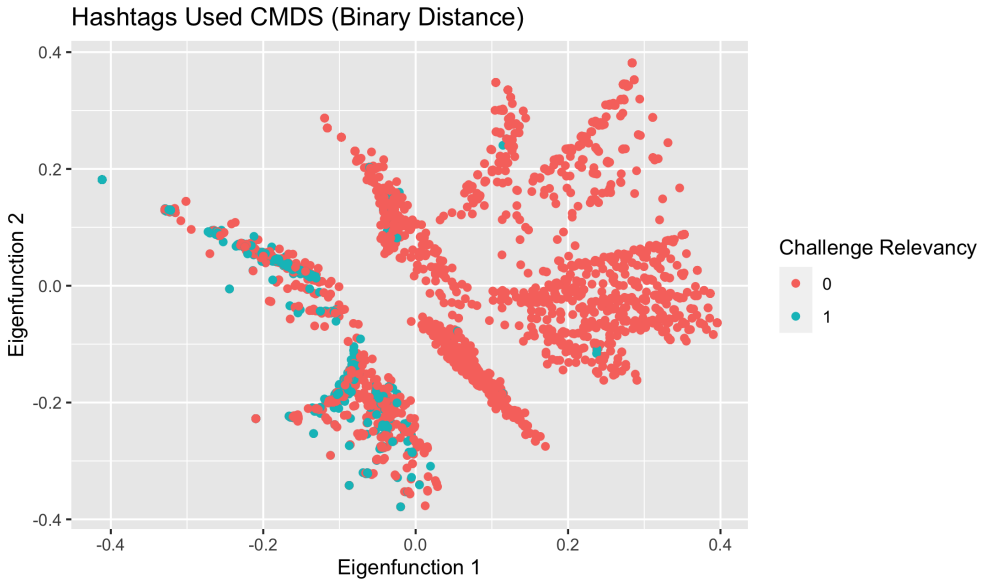

Figure 5: Dissimilarity in hashtags used by videos in the #NoNuanceNovember challenge tag

As Figure 5 highlights, quantifying tag use dissimilarity across all #NoNuanceNovember videos reveals that it is possible to create several clusters that are almost purely challenge irrelevant. There is also a set of two clusters which contain almost all the challenge relevant videos, but they are not very pure. This implies there are many tag patterns, which are used almost exclusively by irrelevant videos, while the tag patterns used by challenge relevant videos are similar to a selection of challenge irrelevant videos. The distinction this set of three scaled variables creates is important for supervised learning prediction methods as it enables a large proportion of videos to be easily and accurately classified as challenge irrelevant.

4.3 Methodology: Four Machine Learning Methods

After creating new variables and cleaning the #NoNuanceNovember dataset, there were 4145 videos representing all videos a user could view under the tag at approximately the day of collection in early March 2021. This means the videos' creation dates spanned November 7th, 2020 through March 5th, 2021. On TikTok, videos can be deleted, made private, or posted, but they cannot be edited. This means that since collection, the true set of videos in the tag could increase or decrease slightly, but as the November-based trend has ended, these changes are unlikely to be by much. For this reason, the set of videos represented in this paper's dataset is the *population* (as of collection) of all videos under the NNN tag, instead of a sample.

Each video's important variables are the time of posting, length of video, number of hashtags, number of non-tag characters, the three scaled variables representing tag use dissimilarity, and the coded challenge relevancy classification. The former seven variables were used to train a model that could predict challenge relevancy. Four different models were attempted to achieve the best prediction: a weighted logistic regression model, a classification and regression tree (CART), and two different Random Forests (one which grew from the original sample and one which used up-sampling to balance classes). All models were measured on the same random 1145 video test set. This section will explain and compare the methodology of each model constructed and section 4.4 will compare each model's predictions.

A logistic regression model was constructed using the time of posting, length of video, number of hashtags, number of non-tag characters, and the three scaled variables representing tag use dissimilarity as predictors to classify a video as 0 (challenge irrelevant) or 1 (challenge relevant). Equation 4 shows the function used in order to calculate probability of challenge relevancy. Videos with a probability larger than .5 were classified as challenge relevant, while those with a probability lower were classified as challenge irrelevant.

$$P(Y = 1|X = x) = \frac{exp(\mathbf{X}\beta)}{1 + exp(\mathbf{X}\beta)} \tag{4}$$

where:

$\mathbf{X}\beta = \beta_0 + \beta_1 tagNumber + \beta_2 charNumber + \beta_3 length + \beta_5 time + \beta_6 tagVec1 + \beta_7 tagVec2 + \beta_8 tagVec3$

$tagNumber$ = Number of hashtags

$charNumber$ = Number of non-tag characters

$length$ = Length of video

$time$ = Posting time

$tagVec$ = Constructed tag dissimilarity variables

To accommodate for the class imbalance of approximately 79% challenge irrelevant videos, weights were included in the logistic regression. Ideal weights were calculated through 5-fold cross validation. It was found the weights .19 and .81 (tag irrelevant and tag relevant respectively) gave the best prediction. This weighting mirrors the dataset's true class imbalance very closely because 3284 of 4145 (79.2%) videos are challenge tag irrelevant. The logistic model's regression on the training set is shown in Table 1.

Table 1: Weighted Logistic Regression Model

	Dependent variable:
	Challenge Relevancy
Length of Video	0.058^{***}
Time of Posting	0.00000
Number of Tags	-0.016
Number of non-Tag Chars	-0.003
Tag e-vector 1	-10.257^{***}
Tag e-vector 2	-4.637^{***}
Tag e-vector 3	-4.698^{***}
Constant	-102.586
Observations	3,000
Log Likelihood	-187.680
Akaike Inf. Crit.	391.360
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1 notes that length of video and the three constructed hashtag disparity variables have a p-value $< .01$. Through the lens of inference, this should be interpreted cautiously, as this model is being used for prediction and does not involve a rigorous test of regression assumptions. Instead, these significances should be seen for their indication that these 4 variables are valuable for their predictive power. For the logistic model, a video's length and the tag use patterns in the description are the important predictors of challenge relevancy. The second model trained on the NNN dataset was a classification and regression tree (CART). Whereas a logistic regression model is a linear classifier (i.e., it creates a linear decision boundary in p dimensions) a CART method trains a model that creates splits in the predictor variables. The algorithm calculates splits for predictor variables at values that will minimize the squared residuals when a constant function is fit on either side of the split. As a result, a CART split is similar to finding the best point for

regression discontinuity, but with the added requirement that the regression on either side of the discontinuity is a constant. If a predictor variable has 3 such splits, the observations in the lower third could be predicted as relevant, those in the middle third predicted to be irrelevant, and those in the upper third once again predicted to be relevant. As a result, CART could be more accurate at prediction than logistic regression if there are non-linear relationships between a predictor and the log-odds. CART was run on the same training set as the logistic regression model using the rpart package in R. Figure 6 shows the resulting tree.

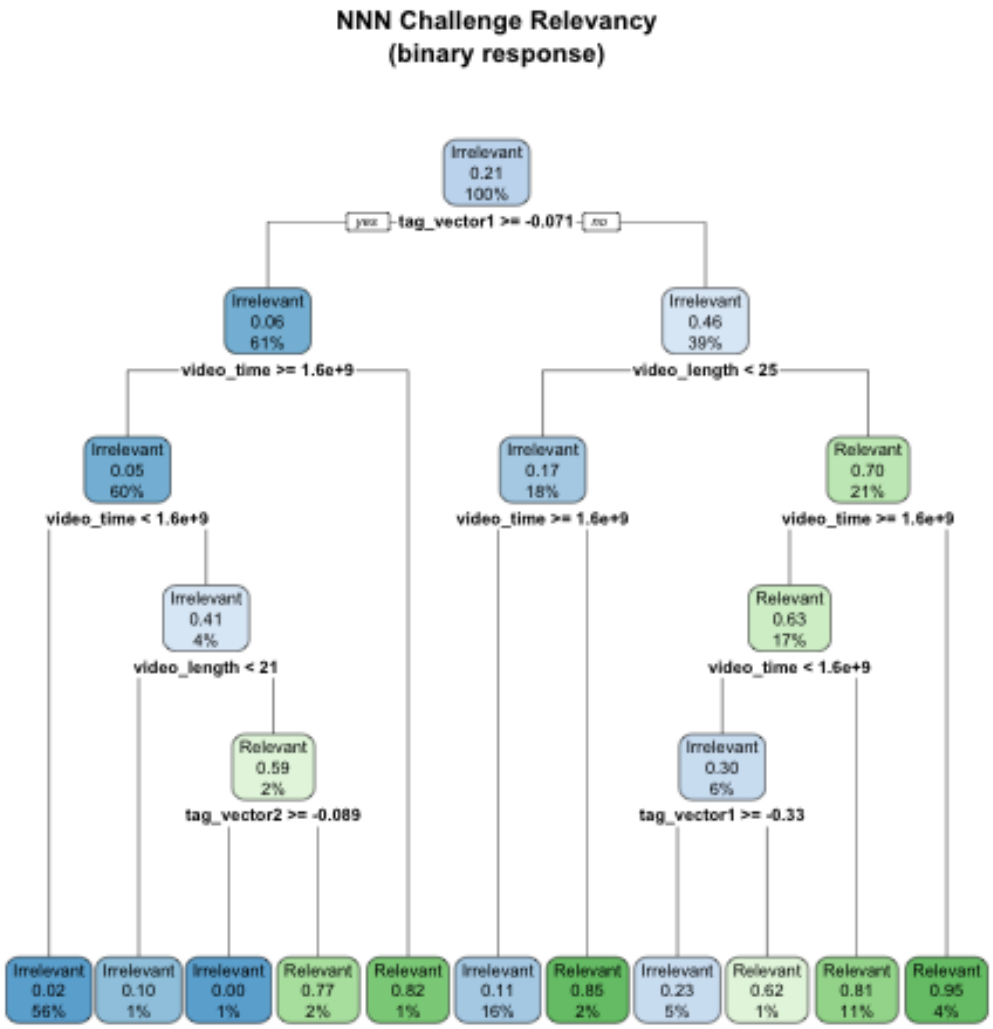

Figure 6: CART Tree for #NoNuanceNovember videos, predicted classifications by challenge relevancy

Each colored box, or node, of the tree diagram in Figure 6 has three lines of text and underneath, a split. The three labels on each colored box represent the predicted class of that node (relevant or irrelevant), the proportion of relevant content in the node, and the percent of total observations in the node. Looking at

the first node, it predicts that content is irrelevant, says 21% of all observations are relevant, and 100% of all observations are under this node. This is just the overall dataset without any splits, so it is only natural to predict content to be irrelevant because of the class imbalance. As more splits are added, the smaller portions of videos in each section can be predicted more accurately. The last row of Figure 6 shows eleven leaves. Those which are a darker green or a darker blue create very pure groupings of content. The rightmost leaf predicts its section of videos to be relevant, and 95% of videos in the leaf indeed are.

Similar to logistic regression, the visualization of the CART method can shed some light on predictors which were most valuable to creating meaningful splits. Splits are calculated iteratively (from the highest to lowest). The first split made in the data is in one of the constructed tag use pattern variables. From there, the video posting time, video length, and another constructed tag use variable appear. This implies these 4 variables represent those which are most important in CART's classification. Notably, video posting time does not appear as important in the logistic regression model. This appearance indicates CART is able to harness the predictive power of an extra variable.

Both of the final two models are based in the Random Forest machine learning technique. Random Forest is similar to CART, but rather than using a single decision tree, the method creates n trees and averages across all of them. A single tree, as in CART, can be a method that has low bias (i.e., with many splits, it can follow the true classifications of the training set very well) but high variance. As a result, a single tree may have great classification ability specifically for the training data, but when applied on a separate test set, these splits are not helpful. A method that uses an ensemble of trees such as Random Forest can alleviate some of these issues because it involves averaging over many trees.

This paper conducts two slightly different Random Forest methods both using the randomForest package in R. The first is on the training dataset used by all previous methods. 100 trees were grown based on the class imbalanced (approximately 79% irrelevant) data. The second Random Forest of 100 trees was grown from up-sampled training data. Up-sampling was done by randomly sampling challenge-relevant videos with replacement from the training set until the set had the same number of challenge relevant videos as challenge irrelevant. This was done in hopes of improving the issues that may come from class imbalance just as the weights of the logistic regression did. The caret package in R was used in order to conduct the up-sampling process.

Because Random Forest is an ensemble of trees method, it is much more difficult to visualize the methodology. In order to investigate the variable importance, as has been done on both CART and weighted logistic regression, a variable importance plot was constructed using the mean decrease in Gini coefficients for each variable. Equation 5 shows the equation for Gini impurity with only 2 classes (relevant and irrelevant) while

Equation 6 shows the Mean Decrease in Gini (MDG) calculation across the forest.

$$GI = 1 - \sum_{i=1}^{2} p^2 \tag{5}$$

where:

GI = Gini Impurity

p = Probability of observations being in class i

$$MDG_v = \frac{1}{n} \sum_{i=1}^{n} (GD_{v,i}) \tag{6}$$

where:

GD = Decrease in Gini impurity by variable v in tree i

n = Number of Trees

Mean Decrease in Gini across variables represents how well splits in the variables improve classification purity on average. A variable with high predictive power should be able to decrease the Gini Impurity coefficient significantly across many trees. Most important are the relative differences in MDG between variables rather than the absolute values themselves. Figure 7 shows in panels A and B the MDG plots of both Random Forest methods.

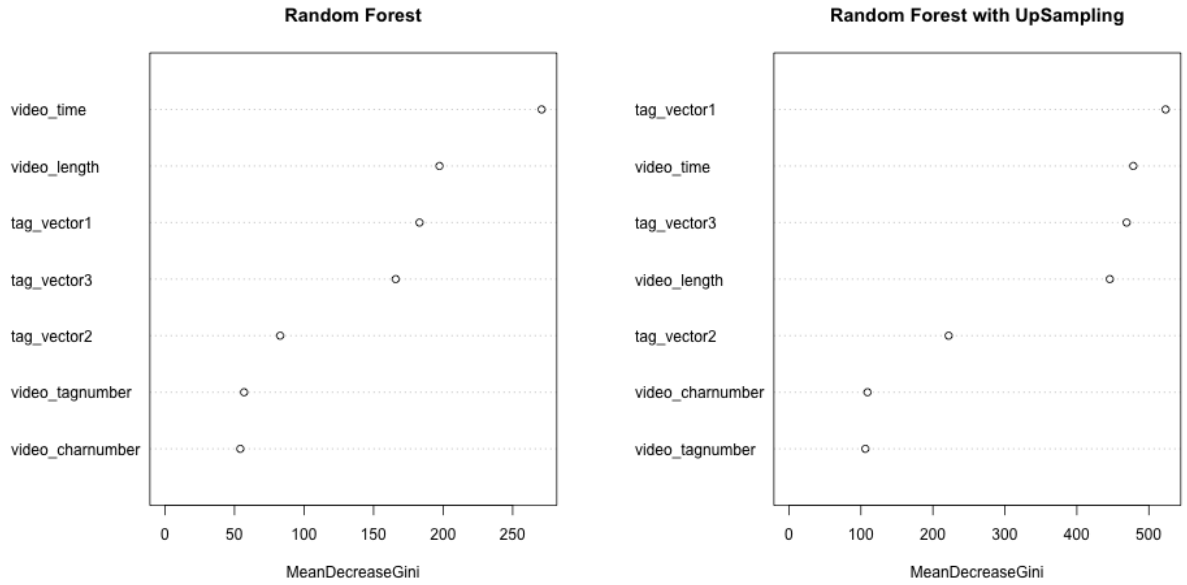

Figure 7: A: Variable Importance for Random Forest, B: Variable Importance for Up-sampled Random Forest

From both Figures 7A and 7B, a major takeaway that mirrors other methods is that the number of

tags and the number of non-tag characters have very low relative predictive power. Both variables cannot create pure groupings using any ML models trained on the NNN dataset. Beyond this, Panels A and B show slightly differing results, though not to a large degree. For both Random Forest on the normal dataset and on the up-sampled one, posting time, video length, and at least two constructed tag usage variables are all relatively highly predictive. Interestingly, while this generally aligns with CART, this differs highly with logistic regression, which was unaffected by the time of a video's posting. Once again, it implies that this variable's predictive power cannot be harnessed with logistic regression. The second tag use variable seems relatively less predictive in both panels A and B. This points to the fact that it does not add much relative predictive power when there are already splits created in two other tag pattern variables. Reducing the 5888 indicators to just two dimensions would likely still yield good predictive results. Additionally, reducing the 5888 indicators to four or more dimensions instead of three would likely have a marginally decreasing benefit.

4.4 Results

Once trained, all four methods were tested on the same 1145 video dataset in order to compare prediction abilities. To assess models' predictions on the test set, two measurements were considered: the standard accuracy measurement and the area under the receiver operating characteristic (ROC) curve. Equation 7 represents the standard accuracy measurement. The ROC curve is a graph that depicts a model's classification performance at all classification thresholds. In order to do this, it plots the true positive rate (aka TPR, recall, or sensitivity) on the x-axis and the false positive rate (FPR) on the y-axis. Equations 8 and 9 represent these two respectively. Note: Positives are observations predicted to be relevant and negatives are observations predicted to be irrelevant. True implies the prediction is correct and false implies the prediction is incorrect.

$$Accuracy = \frac{TP + TN}{P + N} \tag{7}$$

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{TN + FP} \tag{9}$$

where:

TP = True Positives

TN = True Negatives

P = Positives

N = Negatives

The ROC curve shows how well a model can predict a higher probability for a video that is truly relevant

as opposed to a video that is truly irrelevant. This is a useful tool because a perfect model would always give higher predicted probabilities to observations which are relevant than to those that are not. Thus, the ROC curve highlights how well a model discriminates between relevant and irrelevant observations (regardless of the inherent class imbalance). Figure 8 Panel A, B, C, and D show the ROC curves for all 4 models.

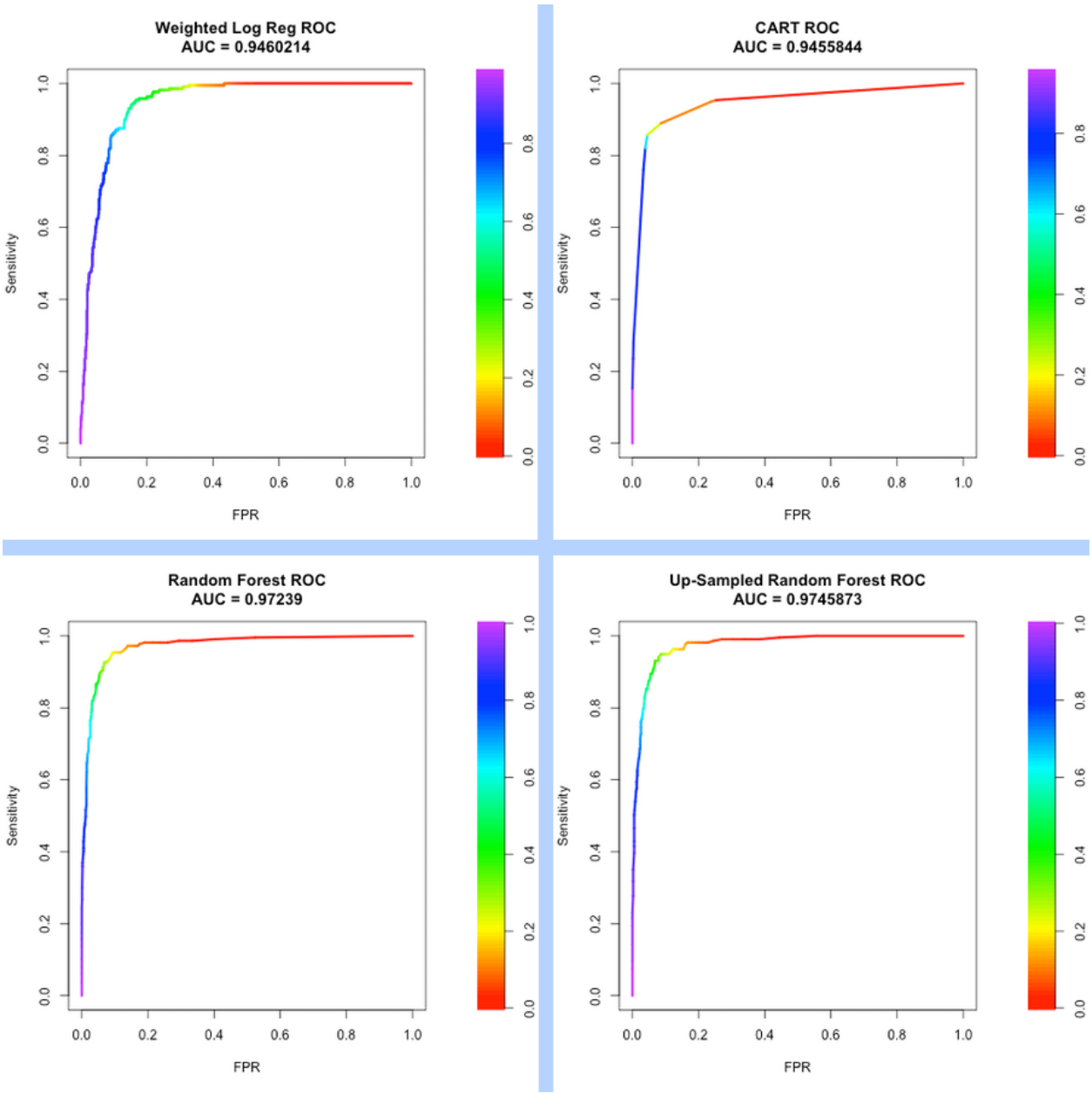

Figure 8: A: Weighted Log Reg ROC, B: CART ROC, C: Random Forest ROC, D: Up-Sampled Random Forest ROC

On each ROC curve, small values in the FPR axis mean low rates of false positives and high rates of true negatives, while big values in the Sensitivity (TPR) axis are high rates of true positives and low rates

of false negatives. This means a perfect model will touch (0.0, 1.0) with a perfect 90-degree angle. A model which cannot discriminate between classes any better than random guessing would be a line which is TPR = FPR (i.e., a 45-degree angle). The area under the curve is labeled AUC, and it follows that the perfect AUC is 1, while the no-better-than-random AUC is 0.5. Figure 8 shows that all models are very good at discriminating between relevant and irrelevant content. Weighted logistic regression and CART have similar ROC-AUC scores of approximately .95, while both Random Forest models improve on this with scores of approximately .97. As a result, the ROC-AUC indicates that if we want a model that best discriminates between the two classes generally, the best bet is to choose either of the Random Forest models.

Accuracy is another way to look at how good a model's prediction is. Accuracy and ROC-AUC are naturally highly correlated (good ROC-AUC will usually mean good accuracy), but sometimes the two measures can select two different models from a set. Accuracy is just the number of correct classifications divided by the total observations. As such, it is important to note that accuracy is quite specific to that set and prone to change, especially if the class imbalance of the test changes. Additionally, in a class imbalanced set, accuracy can be a bit misleading. In NNN, for example, for a classifier to be better than always predicting irrelevant, it must be better than .79 accuracy. That being said, the NNN dataset is not a sample of all of videos under the #NoNuanceNovember hashtag. Instead, it is the population of all user-viewable videos (as of collection). This means the observations of the dataset will not change with a random resampling. For this reason, accuracy would be a useful metric. The models do need to have the general power to discriminate between random relevant and irrelevant observations, but they should also give a very high proportion of correct classifications on this dataset specifically. Table 2 shows each model along with its ROC-AUC, as well as its accuracy.

Model	Accuracy	ROC AUC
Weighted Log Reg	0.857	0.946
CART	0.936	0.946
Random Forest	0.938	0.972
Up-Sampled Random Forest	0.935	0.975

Table 2: 4 Model's Predictive Abilities

Going by pure accuracy, Random Forest (no up-sampling) is the best by a very small margin. Nonetheless, CART, Random Forest, and Up-Sampled Random Forest all give an accuracy of approximately 94%. As Random Forest (no up-sampling) also has an extremely high ROC-AUC (only .003 less than its up-sampled counterpart), this thesis concludes that a Random Forest (no up-sampling) model is the best classification method out of those constructed.

4.5 Challenge Relevancy Discussion

Using the following variables: number of tags, number of non-tag characters, time of posting, video length, and three additional constructed variables representing tag dissimilarity, this thesis constructs a Random Forest model that is able to predict #NoNuanceNovember challenge relevancy with 94% accuracy. Additionally, this most accurate model is one whose most important predictors are time of posting, video length, and tag use patterns. With only this very basic information, a model predicts if videos within this trend satisfy the instructions,

> "It's #NoNuanceNovember, so post your hot takes with no context and run."

This has important implications for users, as well as for the platform. On the user side, if one chooses to scroll through a tag with such a conversation-starting background, it stands to reason that they would prefer to see mostly (if not only) those videos which are pertinent. The current of videos under challenge tags seem to be ordered approximately by views; however, this does not separate relevant and irrelevant videos for the users in any meaningful way. On the other hand, the Random Forest method of this thesis (as measured by the ROC-AUC) is able to do so extremely effectively. This separation of content would improve the user experience. On the platform side, a major goal of any social media site is to hold users' attention for as long as possible. By continuously providing challenge-relevant, discourse-starting videos under the #NoNuanceNovember tag, they would be likely to do so.

Additionally, this Random Forest is trained to predict challenge relevancy using exclusively anonymous video attributes. None of the predictors are derived from video analysis software, so bias involved in algorithmic facial or body recognition can be avoided. AI researcher from the UC Berkeley School of Information, Marc Faddoul, found that the TikTok recommendation system tended to recommend users to follow those who looked like creators they already follow[10]. A classification that uses no predictors gleaned from video or other user-specific info would hopefully avoid this problem, at least in #NoNuanceNovember challenge relevancy prediction specifically. The simple fact that challenge relevancy can be predicted accurately from these attributes alone is an important contribution to TikTok literature. Further research is needed in order to better understand when social media algorithms can or should choose to exclude non-anonymous and biased variables.

[10]https://www.ischool.berkeley.edu/news/2020/alumnus-marc-faddoul-discovers-racial-biases-tiktoks-algorithm

5 Conclusion

In this thesis I analyze how users interact with the TikTok algorithm, what methods they use in order to make content algorithmically favorable, and how to improve the user experience through content relevancy prediction in trending discussion hashtags. I find that users can interact with the algorithm through choice of hashtags, telling the algorithm to whom content should be recommended. Users can self-identify their content as community-relevant in several ways because the same communities have many different accepted tag-use patterns. This is an important finding that sheds light on the context of information dissemination on TikTok. By attaching content to several hashtags that are community relevant, event, political, and general information can be hyper-directed to relevant communities.

Additionally, I train a Random Forest model that can predict, with 94% accuracy, the relevancy of a video to the previously trending challenge tag, #NoNuanceNovember. As this tag is home to many controversial and generally engaging videos, the experience of users searching through the tag can be improved by providing more challenge relevant content than challenge irrelevant. This model is also trained on the basic video attributes of posting date, video length, number of hashtags, number of non-tag characters, and hashtag pattern. As a result, this algorithm would avoid any biases that may come from prediction involving user/video characteristics and is completely anonymous. This paper contributes to existing social network analysis literature by portraying the use of community tag patterns in a novel visualization and constructing an accurate content relevancy prediction model that could improve user experience on the platform.

6 Ethical Considerations

In investigating the use of hashtags across groups and the power of predictive modeling on challenge relevancy, I have looked at and collected large amounts of TikTok videos both in many professional communities, as well as more widely across challenges. While all of these videos are public and can easily be viewed by anyone on the app, the aggregation and systematization of large amounts of this data could be seen as an overstep in privacy. In order to maintain data protection, none of the identifiable data has been included in this paper or shared with others.

References

[1] Crystal Abidin. Mapping internet celebrity on TikTok: Exploring attention economies and visibility labours. *Cultural Science Journal*, 12(1):77–103, 2020.

[2] A. Anusha and Sanjay Singh. Is that twitter hashtag worth reading. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, WCI '15, page 272–277, New York, NY, USA, 2015. Association for Computing Machinery.

[3] Ekaterina Budnik, Violetta Gaputina, and Vera Boguslavskaya. Dynamic of hashtag functions development in new media: Hashtag as an identificational mark of digital communication in social networks. In *Proceedings of the XI International Scientific Conference Communicative Strategies of the Information Society*, CSIS'2019, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. When users control the algorithms: Values expressed in practices on twitter. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[5] Alice R. Daer, Rebecca Hoffman, and Seth Goodman. Rhetorical functions of hashtag forms across social media applications. In *Proceedings of the 32nd ACM International Conference on The Design of Communication CD-ROM*, SIGDOC '14, New York, NY, USA, 2014. Association for Computing Machinery.

[6] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 593–596, New York, NY, USA, 2013. Association for Computing Machinery.

[7] Obaida Hanteer, Luca Rossi, Davide Vega D'Aurelio, and Matteo Magnani. From interaction to participation: The role of the imagined audience in social media community detection and an application to political communication on twitter. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '18, page 531–534. IEEE Press, 2018.

[8] Trevor Hastie, Jerome Friedman, and Robert Tisbshirani. *The Elements of statistical learning: data mining, inference, and prediction*. Springer, second edition, 2017.

[9] Anne Jonas and Jenna Burrell. Friction, snake oil, and weird countries: Cybersecurity systems could deepen global inequality through regional blocking. *Big Data Society*, 6:205395171983523, 03 2019.

[10] Muhammad Haseeb UR Rehman Khan, Kei Wakabayashi, and Satoshi Fukuyama. Events insights extraction from twitter using lda and day-hashtag pooling. In *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications Services*, iiWAS2019, page 240–244, New York, NY, USA, 2019. Association for Computing Machinery.

[11] Eric Krokos, Hanan Samet, and Jagan Sankaranarayanan. A look into twitter hashtag discovery and generation. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '14, page 49–56, New York, NY, USA, 2014. Association for Computing Machinery.

[12] Zongyang Ma, Aixin Sun, and Gao Cong. Will this hashtag be popular tomorrow? In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 1173–1174, New York, NY, USA, 2012. Association for Computing Machinery.

[13] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the partisan beat: A first analysis of political communication on tiktok. In *12th ACM Conference on Web Science*, WebSci '20, page 257–266, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Eriko Otsuka, Scott A. Wallace, and David Chiu. Design and evaluation of a twitter hashtag recommendation system. In *Proceedings of the 18th International Database Engineering Applications Symposium*, IDEAS '14, page 330–333, New York, NY, USA, 2014. Association for Computing Machinery.

[15] Ellen Simpson and Bryan Semaan. For you, or for"you"? everyday lgbtq+ encounters with tiktok. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), January 2021.

[16] Oren Tsur and Ari Rappoport. What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 643–652, New York, NY, USA, 2012. Association for Computing Machinery.

[17] Ian D. Wood, John Glover, and Paul Buitelaar. Community topic usage in online social media. *Trans. Soc. Comput.*, 3(3), May 2020.

[18] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. Agent, gatekeeper, drug dealer: How content creators craft algorithmic personas. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[19] Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Trans. Intell. Syst. Technol.*, 3(4), September 2012.