Predicting Audience Tweet Engagement

by

Julia Wu

S.B., Computer Science and Engineering, Mathematics (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

Author .....
Department of Electrical Engineering and Computer Science
January 14, 2022

Certified by .....
Deb Roy
Professor
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Predicting Audience Tweet Engagement

by

Julia Wu

Submitted to the Department of Electrical Engineering and Computer Science
on January 14, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Social media has become the ubiquitous infrastructure through which the world is connected. It allows people to interact not only with family members and friends but also with prominent figures like movie stars, presidential candidates, and even royalty. These celebrities have immense presences on social media, and each post they share has the potential to reach millions of people. As the sphere of social media influence grows increasingly large, it also becomes increasingly important to be able to understand how influencers on social media affect their audience. However, it is difficult for individuals with large social media platforms to gain insight into how their posts influence their followers. While social media platforms do provide influencers with some audience breakdowns and statistics, they are often not granular enough to be useful. In this thesis, we present methods to analyze an influencer's tweets and audience. We then use these results to predict which segments of an influencers audience will interact with different types of posts. These insights can help determine which areas an influencer has the greatest potential to make an impact in and thus guide the direction and content of influencer campaigns.

Thesis Supervisor: Deb Roy
Title: Professor

Acknowledgments

First, I would like to thank my advisor, Professor Deb Roy, for all his guidance and support throughout this thesis. This opportunity to research with him and CCC over the past summer and fall semesters has been an amazing and valuable learning experience.

Next, I would like to thank Brandon Roy for the time and effort he has put into helping me develop and implement ideas. He has been beyond helpful and has always been happy to have long discussions and brainstorming sessions, which have all been instrumental to this work. It has been wonderful to collaborate with him throughout this journey. I would also like to thank Bridgit Mendler for providing much of the motivation for this thesis, particularly in the earlier days.

I am eternally grateful to MIT for all the incredible memories, opportunities, and friends that have shaped me both as a student and as a person, and have made the past 4.5 years so fulfilling. To Angel, Elizabeth, Katherine, Seungweon, and Steven: thank you for making MIT feel like home.

Finally, I want to thank my parents and older brother Michael for their unwavering love, advice, and encouragement. Without them, I would not be who or where I am today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In an era of social media dominance, platforms such as Twitter have become a major method of sharing and consuming information. These social media platforms have facilitated the formation of social networks that reflect and extend those in the real world.

In particular, social media differs from previous modes of communication in that it supplies each user with the equivalent of an online megaphone. Users can broadcast news and opinions to their followers and beyond with the use of tools like hashtags and mentions. Even the average person who plays their cards right can witness their posts go viral and reach a much greater viewership than just their follower base.

On the other end of the spectrum, celebrities start off with huge followings on social media. By default, their social media posts are exposed to a large-scale audience. In addition to having sheer numbers of followers, many celebrities also tend to hold a higher degree of trustworthiness, especially when their messages align with their fields of expertise. For example, a post about cooking would likely hold much more merit if it was made by Gordon Ramsay rather than by a standard Twitter user. As a result, a celebrity has significantly greater potential than the average user to impact a large number of people. A natural question that arises is how those with substantial social media presences can utilize their platforms more effectively to spread messages.

As this is an inherently broad problem, we seek to help answer it by narrowing its scope and considering how an audience responds to a post on Twitter. We focus

on predicting who in an audience will interact with a tweet. If we can predict which parts of an audience will interact with a post, we can choose which posts to share to capture the audience's attention, as well as help craft messages to maximize audience engagement. We would also be able to select specific people who are best suited to reach a target audience or spread a certain message.

The ability to make these tweet engagement predictions has the potential to make a far-reaching impact in the realm of social media communication, beyond the extent of a single Twitter user's audience. We can create campaigns to help combat an information imbalance across divides, or draw attention to important issues among underexposed groups. We can utilize these predictions to determine the effect a tweet's word choice can have on who chooses to interact with the tweet, and then to modify the word choice to resonate with specific communities of people. There are a multitude of applications that have direct impact on how information is spread on Twitter.

There are three main factors that affect tweet engagement: the message (tweet content), the audience, and the messenger (person who tweets). In this thesis, we restrict our attention to a single influencer, American actress Kerry Washington[1], and her tweets and audience. By keeping the messenger constant, we focus on using the first two factors, the tweet content and the audience, in order to predict tweet engagement. In particular, we will approach the audience from two perspectives: the cluster level, where we aggregate users according to their interests and thus predict which interest clusters will respond to tweets; and the individual level, where we consider each individual user and predict which users will respond to tweets.

Although this work centers on a specific influencer, the techniques used and insights gained can be applied to any Twitter user and their tweets and followers. The goal of this thesis is to develop generalizable results to better inform influencers on how they can maximize their impact on social media.

[1]We chose to study Kerry Washington because she balances both her work and her Twitter content across her entertainment and social activism projects. Additionally, both Kerry Washington and the team at her initiative Influence Change were interested in pursuing the underlying research questions of this thesis, and so we had access to her Influence Change team to discuss the trajectory of this project as it developed.

1.1 Related Work

This research draws inspiration and ideas from several works in the realm of Twitter analysis. Previously, works such as Kupavskii et al. [4], Vosoughi et al. [13], and Cheng et al. [2] introduced the concept and construction of retweet cascades, which helped prompt our investigation into the spread of information on Twitter and which we will borrow to visualize the retweet structure of a tweet as a tree.

Zhang et al. [17] perform an in-depth analysis on the problem of identifying communities on Twitter by utilizing metrics such as text, hashtag, URL, retweeting, and following similarities to compute user similarity, which they then use to cluster users. For simplicity, we will use only following relationships to group users into interest clusters, as Weng et al. [14] demonstrate follow links on Twitter are correlated with user interests.

There are a wide variety of papers centered on predicting the various interactions that take place on Twitter. Zaman et al. [15] take a simpler approach to predicting whether a user will retweet a tweet by looking at basic user and tweet features such as number of followers and following. Petrovic et al. [7] attempt to solve the same problem by adding more nuanced features, such as the user's number of favorites and statuses and the tweet's number of hashtags, mentions, and URLs. At a user level, Sotiropoulos et al. [10] seek to predict the interactions between pairs of users, as well as study the correlation between different types of user interactions. Shugars and Beauchamp [9] also predict user engagement in conversations. Other works are primarily concerned with the popularity of tweets, namely predicting the number of retweets a tweet will receive [16] [4]. In a similar vein, we pursue a model that will predict retweet relationships, both at a community and individual level.

Detailed research has also been done on topic modeling on Twitter. Sanandres et al. [8] and Alvarez-Melis and Saveski [1] both use Latent Dirichlet Allocation (LDA) models to extract topics from tweets. Similarly, we derive the most informative words from tweets using a variant of the simpler term frequency-inverse document frequency (TF-IDF) measure to help provide our model with more information about what the

user interest clusters are interested in.

1.2 Thesis Organization

The remainder of the thesis is organized as follows:

- Chapter 2 describes the tweets we work with and how we extract feature representations.
- Chapter 3 breaks down the audience and the different perspectives from which we analyze the audience, as well as develops audience representations at both the cluster and individual levels.
- Chapter 4 details the cluster level prediction task, presents the cluster prediction model, and analyzes its results.
- Chapter 5 details the individual level prediction task, presents the individual prediction model, and analyzes its results.
- Chapter 6 summarizes the thesis contributions and discusses any limitations and possible directions for future work.

Chapter 2

Tweets

We first examine Kerry Washington's tweets. Our aim is to develop a fine-grained representation of the content of her tweets.

2.1 Data Overview

Using Twitter's Powertrack and Historic Powertrack APIs, we retrieved all of Kerry Washington's tweets from January 1, 2021 to November 18, 2021, ignoring any retweets. This yielded a total of 709 tweets, which we will refer to as KW tweets. Using the same Twitter APIs, we pulled all retweets, replies (including replies of replies), and quote tweets of these KW tweets. We call each of these a response to a KW tweet. Specifically, we need the users who posted these responses. In total, there are 187,609 unique and non-private users who have responded to any of the KW tweets. Throughout the rest of this thesis, we will be working with this set of 709 KW tweets and 187,609 users who posted the responses to these KW tweets.

Observing this collection of KW tweets, we conclude that her tweets are primarily about social issues (e.g. racial, gender, and political matters), entertainment-related topics, and personal content. Table 2.1 provides several examples of her tweets, along with the topics of the tweets.

Note that the KW tweets we categorize as "personal content" are rather miscellaneous and have no unifying theme. The KW tweets that we determine to be about

Kerry Washington Tweet	**Topic**
This emotional rollercoaster of Black mourning and injustice is treacherous. We get a tiny window of accountability and then more devastation. The need for reckoning and reform is undeniable.	Social issues
ALL elections are important. All of em. Not just every four years. Check your registration, register yourself (and friends and family too!) and then get out to vote! #NationalVoterRegistrationDay	Social issues
These Senators work for us. WE have the power. If you believe voting should be more accessible to all, keep the momentum going. #CallOutYourSenators tell them to Vote YES on the #ForThePeopleAct	Social issues
BIG NEWS! @SimpsonStreet is making a show!!!!! This project is such a labor of love, I cannot wait to direct and work with this badass writers room led by @dramaraamla and the INSANELY talented Emayatzi Corinealdi. More to come! #ReasonableDoubt https://t.co/Eka36Oayhl	Entertainment
Through the stories of doctors, parents, friends, & families, #InTheSunFilm teaches us how to protect & cherish our skin. It was an honor to produce this film with @SimpsonStreet, #NeutrogenaStudios & @Neutrogena. Can't wait for you to see it on April 27. http://inthesunfilm.com https://t.co/V8bDUD42O3	Entertainment
😱😱😱😱😱😱 These were some of the most BRILLIANT performances I have seen on a television screen...EVER. To be featured in the same category as you ladies is such an honor. I am so deeply grateful to the #SAGAwards for this nomination! https://t.co/aNhXCPrp9H	Entertainment
Tweeps...it's been a WEEK. I need some JOY. Send me some funny videos?	Personal
"To me self care...is learning to say no. It's knowing yourself so you can make choices that are an expression of you" - @TraceeEllisRoss with my #MondayMantra	Personal
Happy Birthday @theebillyporter!!!!!! I looooooove you!!!! Xoxoxoxo https://t.co/0qytbwUp2G	Personal

Table 2.1: Examples of Kerry Washington's tweets and their topics.

social issues or entertainment are well-contained within their respective areas and have underlying similarities. Even from the few examples displayed, we can see how the personal tweets vary much more in topic than the other two types of tweets.

2.2 How Tweets Spread

With the KW tweet and response data we have assembled, we can observe how the KW tweets spread. As retweets correlate most directly with pure information spreading, we focus on the retweets for now. We construct retweet cascades as in [13] to visualize the propagation of information throughout an audience. Figure 2-1 shows two examples of retweet cascades for two corresponding KW tweets. Figure 2-1a depicts the retweet cascade corresponding to a KW tweet about entertainment, while Figure 2-1b depicts the retweet cascade corresponding to a KW tweet about politics.

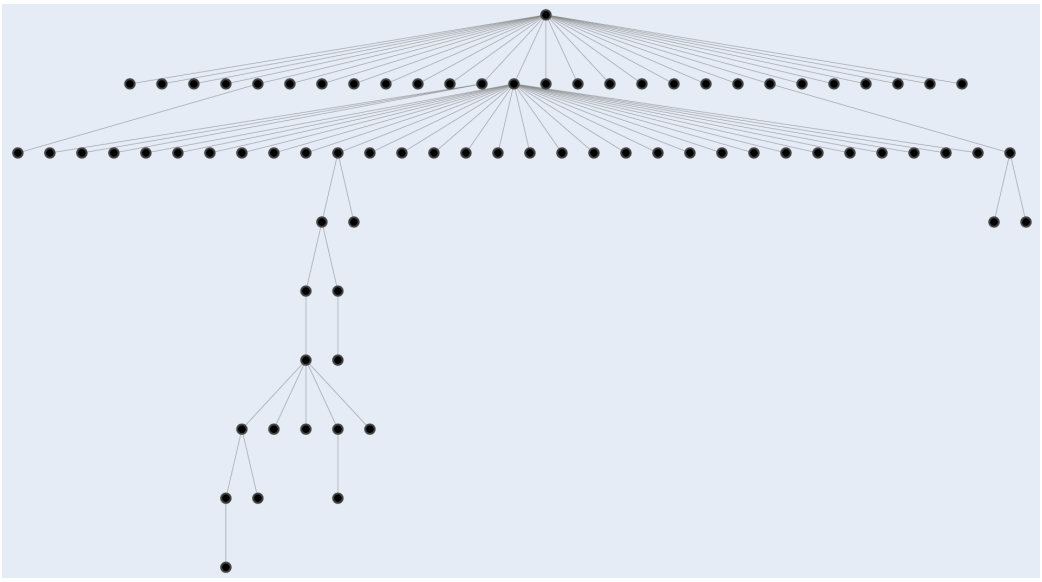
Both retweet cascades illustrate that the majority of users simply retweet the tweet from someone and do not spark further retweets, i.e. no one else retweets from them. We can see these users are most common, as most nodes have either 0 or 1 child nodes. Some users, however, are able to generate subtrees of retweets. In Figure 2-1a, one user in the first level of retweets generates a wide second level of retweets, which then spawns a deep subtree of retweets. In Figure 2-1b, we see a large subtree of retweets that extends down through multiple levels of the audience.

For the most part, the spread of the KW tweets is facilitated by a few users in each retweet cascade. This suggests the presence of various roles within the social networks on Twitter. For example, the users who generate subtrees of retweets may be influencers within their Twitter communities. If we can identify these users, we can more efficiently spread information across specific communities by targeting them.

2.3 Tweet Content

2.3.1 Tweet Text

We need the text of each KW tweet in order to construct a tweet representation. We can get the text from the tweet objects returned by the Powertrack and Historic Powertrack APIs. However, there are some tweets that we must pay closer attention to. Replies and quote tweets are two particular types of tweets that are responses to



(a) OK so I already loved @MsSarahPaulson, but after this episode of #StreetY-ouGrewUp on I looooooooooooooooove her. Go watch, youll see why too! 👉🏽 https://t.co/ecWhfLqHfw https://t.co/R7Kl8r93kQ

(b) California!!!! We KAM do this!!!! (see what I did there?😂)....But for real, make sure youre registered and vote "no" by September 14th!!!! https://t.co/QskbikDuoM

Figure 2-1: Examples of retweet cascades of KW tweets. Note that the root node of the retweet cascade is the original KW tweet, every other node is a retweet (or more accurately, the user who retweeted), and an edge corresponds to where the retweet was derived from.

other tweets. When Kerry Washington writes a reply or quote tweet, she sometimes does not include much text and simply expresses her agreement with the original tweet's message. Table 2.2 displays examples of Kerry Washington's responses and the tweets she responded to. The examples demonstrate how sparse a response from Kerry Washington can be, and how the text from the original tweet contains all the necessary information and context.

Kerry Washington Response	**Original Tweet**
Amen. So true.	Winning is fun, but if you find a family along the way, you can't lose.
Thank you @airbnb	Starting today, Airbnb will begin housing 20,000 Afghan refugees globally for free.
Amen @coribush. #BanOffOurBodies	I'm thinking about the Black, brown, low-income, queer, and young folks in Texas. The folks this abortion health care ban will disproportionately harm. Wealthy white folks will have the means to access abortion care. Our communities won't.

Table 2.2: Examples of Kerry Washington's responses and the original tweets.

Approximately 40% of the 709 KW tweets are quote tweets and 5% are replies, almost all of which do not contain sufficient context. For these tweets, it is nearly impossible to infer anything about the subject of the original tweet or Kerry Washington's response just by looking at the text of her tweet alone. It would likely also be almost impossible to predict who would respond. To add the necessary contextual information, we concatenate the text of the original tweet with the text of Kerry Washington's tweet.

2.3.2 Tweet Content Representation

Once we have the full tweet text for each KW tweet, our objective is to represent the content of the KW tweets in a way that is optimal for classification. We use the Twitter-based transformer BERTweet [5], which is based on BERT-base and has been pre-trained on a corpus of 850 million English tweets, to generate tweet representations. For text classification tasks, BERT uses the final hidden state of the CLS

token as the representation of the input sequence [12]. As our task is related to text classification, we take BERTweet's final CLS embedding as our representation of the tweet. Furthermore, we will be fine-tuning BERTweet in order to help optimize the representations for our task [12].

Aside from word embeddings, Suh et al. [11] show that the hashtags, URLs, and mentions of users (i.e. tagging a user's username) in a tweet are highly correlated with the tweet's retweetability. So, we add 3 features to represent the number of hashtags, URLs, and mentions of users that the KW tweet contains. We also add 3 binary features that represent whether the KW tweet contains media (attached photos or videos), is a reply, or is a quote tweet. Replies are not shown in a user's Twitter timeline unless the user follows both the original tweeter and the replier, and so generally replies tend to receive fewer interactions simply because they are not surfaced as much in followers' timelines. We will call these 6 features the selected tweet features. Our final representation of each KW tweet will be the concatenation of these 6 selected tweet features and the fine-tuned embedding of the CLS token from BERTweet.

Chapter 3

Audience

Next, we will be analyzing Kerry Washington's Twitter audience. Intuitively, an audience will interact with a tweet if they are interested in the content of the tweet. As our goal is ultimately to predict which users would interact with a tweet, we want to capture information about their interests in our representation of the audience.

3.1 Audience Identification

We consider a Twitter user's audience to be the user's followers. Since Kerry Washington's Twitter account was created in 2010, we assume that her follower base is relatively stable. For example, we would not expect the influx of thousands of new followers each day that we might expect when a celebrity initially creates a Twitter account. This assumption of a stable follower base also requires that Kerry Washington's Twitter behavior has not recently undergone any major shifts. We essentially take a snapshot of Kerry Washington's audience by fetching her followers at one point in time (July 2021).

In order to pare down the set of users we work with as well as ignore any dormant or infrequent users, we identify a subset of Kerry Washington's 5.5 million followers that are active on Twitter. We pulled the user information objects of all of her non-private followers using the Twitter API. The user information object of a user contains detailed information about the user, such as the user's display name, number

of favorite tweets, and number of statuses.

Then, we use two kinds of filtering methods to screen out any inactive users. The first looks only at the state of a user's profile at one point in time and thus filters users out based on the overall activity level of the user. We consider followers who satisfy any of the following criteria to be inactive:

1. Fewer than 5 followers and fewer than 5 friends
2. 0 friends and 0 statuses
3. Fewer than 20 favorites and most recent status is from before January 1, 2020
4. Default profile image, default background image, and most recent status was posted before January 1, 2020
5. Fewer than 20 friends and most recent status is from before January 1, 2020
6. Default profile image, default background image, fewer than 10 statuses, fewer than 20 favorites, and fewer than 20 friends
7. Default profile image, default background image, fewer than 10 statuses, fewer than 20 favorites, and fewer than 20 followers
8. Fewer than 5 statuses, fewer than 10 favorites, and fewer than 15 friends
9. Fewer than 5 statuses, fewer than 10 favorites, and fewer than 15 followers

This method marks 2,053,015 users as active.

The second filtering method compares a user's information object at two points in time and filters users out based on whether they have made changes to their profiles over that time period. We pulled the user information objects of her followers from Twitter at two different times, once in April 2021 and once in July 2021. If there are any differences between a user's information objects at these two times, then we consider that user to be active. We also mark any user who has tweeted after January 1, 2021, as active. This method marks 1,153,338 total users as active.

Combining the users deemed active by at least one of these two methods yields a population of 2,083,496 users. Moving forward, we will treat this community of 2 million active users as Kerry Washington's audience.

3.2 Cluster Level Audience Characterization

We now discuss several lenses through which we can view an audience at the cluster level and work towards a representation of the interest clusters.

3.2.1 Who Users Follow

As demonstrated by Weng et al. [14], Twitter follow relationships are correlated with user interests, even when they are not reciprocated. So, although we do not have direct access to a user's interests, we can use the Twitter accounts they follow as a proxy. As a result, we cluster Kerry Washington's audience based on the accounts that they follow.

Over the 2 million users in Kerry Washington's audience, we identify the approximately 60 million accounts followed and construct a 2M $\times$ 60M binary matrix that represents the accounts followed by each user in the audience. Then, we perform a truncated singular value decomposition (SVD) on this matrix and take the left singular vectors as representations of the users. Finally, we perform k-means clustering on these user vectors to get our interest clusters.

Cluster Evaluation

As with any clustering problem, a major area of consideration is how to select the number of clusters, or the value of k since we are using k-means as our clustering algorithm. We propose a method of evaluating different clustering runs that use different values of k, where we essentially compare the differences in the topics of the recent tweets from the users in each cluster.

We collect the tweets posted by Kerry Washington's audience. To help reduce the amount of data, we sampled approximately half of the audience, resulting in roughly 1

million users, and pulled the 100 most recent tweets from each user using the Twitter API. From these tweets, we take those that were posted on and after January 1, 2021 that have an English language code. The most recent tweet is from October 28, 2021. This gives us a collection of 38,406,196 tweets across 10 months.

To distill the topics that the users in the audience are tweeting about, we run LDA on these 38 million audience tweets. Since we are concerned about the topics that each user is interested in rather than the topic of each tweet, we combine each user's tweets into a single document. This produces a set of approximately 1 million documents. We preprocess these documents by removing any URLs, tokenizing the text, removing any stopwords, numbers, and punctuation, and lemmatizing and stemming the tokens. Then, we train an LDA model with 100 topics on these documents and use this model to generate topic probability distributions for each user. A user's topic distribution is a 100-dimensional vector of probabilities, where each value is the fraction of words in the user's consolidated tweet document that belong to the corresponding topic.

We define the topical difference between two users as in [14].

Definition 3.2.1 *Given two users i, j and their respective topic probability distributions p_i, p_j, the topical difference between i and j is*

$$dist(i,j) = \sqrt{2 * D_{JS}(i,j)}.$$

Note that D_{JS} is the Jensen-Shannon divergence between the probability distributions p_i and p_j:

$$D_{JS}(i,j) = \frac{1}{2}\Big(D_{KL}(p_i||M) + D_{KL}(p_j||M)\Big),$$

where $M = \frac{1}{2}(p_i + p_j)$ (to ensure that this measure is symmetric), and D_{KL} is the Kullback-Leibler divergence:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

Now that we have a metric to compare users with, we evaluate the clusters using the following procedure:

1. For each cluster c_1, sample $5 * |c_1|$ (where $|c_1|$ is the number of users in c_1) pairs of users and calculate the topical difference between each pair. Let d_1 be this set of $5 * |c_1|$ topical differences and μ_1 be the true population mean of the distribution of the topical differences between users in c_1 (note that μ_1 does not refer to the empirical mean of d_1). Since d_1 is composed of topical differences between users in the same cluster, we ideally want μ_1 to be small.

   (a) For a different cluster c_2, sample $5 * \max(|c_1|, |c_2|)$ pairs of users (u_1, u_2), where u_1 is a user from c_1 and u_2 is a user from c_2. Calculate the topical difference between each pair. Let d_2 be this set of $5 * \max(|c_1|, |c_2|)$ topical differences and μ_2 be the true population mean of the topical differences between users in c_1 and c_2 (note that μ_2 does not refer to the empirical mean of d_2). Since d_2 is composed of topical differences between users from different clusters, we ideally want μ_2 to be large.

   (b) Conduct a two-sample t-test on d_1 and d_2 to determine if these two sets of topical difference samples are drawn from the same population. Our null hypothesis is $H_0 : \mu_1 = \mu_2$, and our alternative hypothesis is $H_1 : \mu_1 < \mu_2$. We choose the significant level to be $\alpha = 0.05$. Note that if we reject H_0, this basically says that the topical differences between users in c_1 are smaller than the topical differences between users in c_1 and users in c_2. In other words, users in c_1 tweet about and thus are interested in more similar topics. Ideally, we want to reject H_0, as this would demonstrate that the users in c_1 are distinct from those in c_2, i.e. our clustering has created distinguishable clusters of users.

   (c) Repeat steps (a) and (b) for each cluster $c_2 > c_1$.

2. Repeat step 1 for each cluster c_1.

Using this evaluation scheme, we tested out $k = 50, 60, 70, \ldots, 500$. For each value of k, we perform $k(k-1)$ t-tests. We choose the value of k that has the highest proportion of rejected t-tests: 50. So, we divide our audience into 50 interest clusters.

We choose to evaluate our clustering based on the topical distributions as opposed to something like the distribution across accounts followed (which is what the clustering is actually performed on) because we care more directly about representing user interests. The users' tweets express their interests in a more nuanced and thorough way.

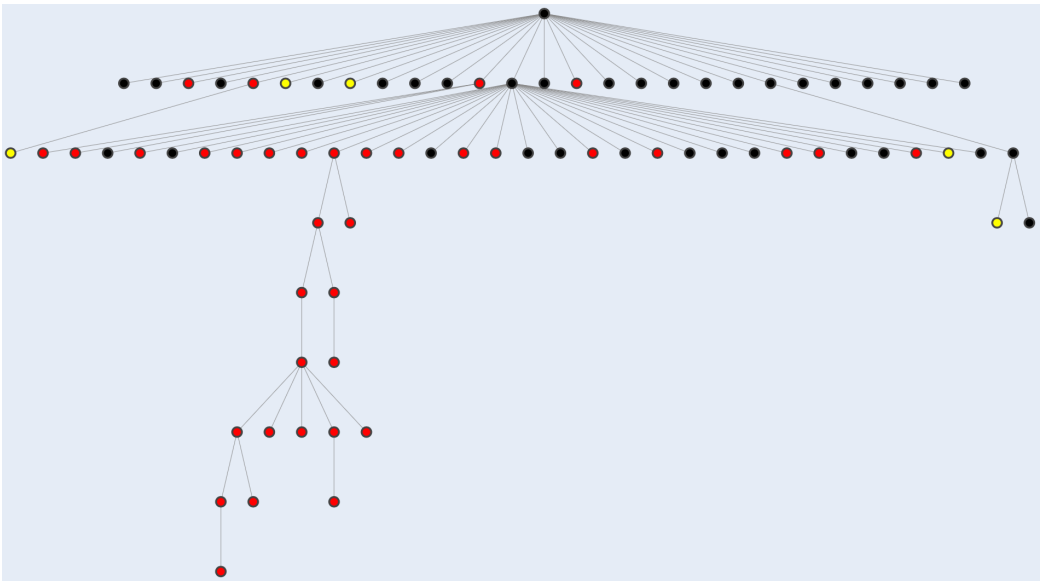
User Projection

Since Kerry Washington is a public figure and has a non-private Twitter account, users do not have to follow her in order to respond to her tweets. Of the 187,609 unique users who responded to the 709 KW tweets, 120,078 users are not included in our interest clusters. That is, these 120,078 users are either not following Kerry Washington or were not considered active users during our audience identification step in Section 3.1. Because a significant portion of the responders are not in our interest clusters, we project them into our clusters to acquire more accurate data about which clusters have responded to tweets. We project users according to our clustering method: we derive the SVD user representation and then choose the cluster closest to this representation. Note that this generates a decent cluster projection as long as the set of accounts followed by the user has significant overlap with the 60 million accounts followed by the KW audience.

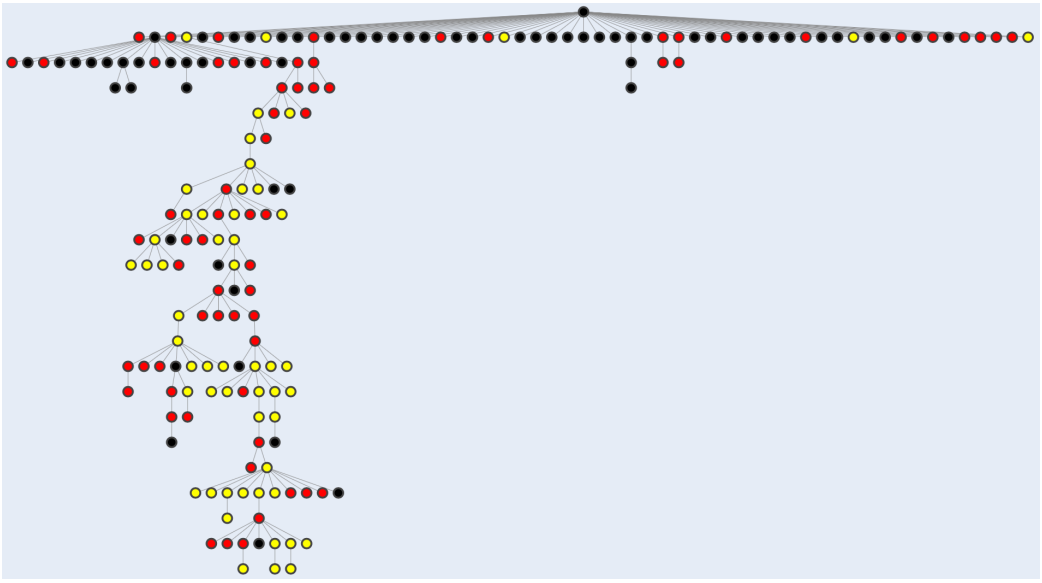
How Tweets Spread Through Interest Clusters

Once we have the audience sectioned into interest clusters, we revisit the retweet cascades introduced in the previous section. We can now contextualize these retweet cascades with respect to the overall audience and visualize how the KW tweets spread throughout the various interest clusters in the audience. Figure 3-1 shows the same retweet cascades as Figure 2-1, and additionally depicts the two clusters that contribute the most retweets.

These retweet cascades visualize which clusters are "activated" by the KW tweets. In Figure 3-1a, the second level of retweets is mostly red and the deep subtree is entirely red, demonstrating that the KW tweet reached and resonated with users in



(a) Red and yellow nodes correspond to clusters 33 and 2 respectively. Almost 50% of the total retweets this KW tweet has are from cluster 33.



(b) Red and yellow nodes correspond to clusters 8 and 40 respectively. Roughly 35% and 30% of the total retweets this KW tweet has are from clusters 8 and 40 respectively.

Figure 3-1: Examples of retweet cascades intersected with interest clusters. Colored nodes indicate users in clusters; we show only the 2 clusters who have retweeted the KW tweet the most. Red and yellow nodes indicate users in the first and second most common clusters respectively.

cluster 33. In Figure 3-1b, the deep subtree consists primarily of red and yellow nodes, illustrating how the KW tweet is circulated within clusters 8 and 40. We will further interpret these observations in the next section, where we will gain some insight into the clusters themselves.

As mentioned in Section 2.2, there may be users who are influencers within their interest clusters. Their retweets generate wide and deep subtrees of retweets, spreading the information in the tweet. Figure 3-1 illustrates how the KW tweets reach particular clusters and stay within those clusters. Clearly, users tend to react and therefore interact more with tweets that are spread by other users in the same cluster as they share similar interests. So, if we can pinpoint users who have influence within their interest clusters, we can then aim to elicit a response from these users to more effectively spread tweets throughout these clusters.

3.2.2 How Users Describe Themselves

Each Twitter user has the option to describe themselves in their Twitter bios. How people describe themselves often provides a great deal of insight into who they are and what they are interested in. In particular, we want to leverage user Twitter bios to qualitatively differentiate the interest clusters obtained in the previous section.

In order to interpret these interest clusters, we summarize each cluster using their Twitter bios. First, we preprocess the text from the user Twitter bios by tokenizing each bio, removing stopwords, digits, and punctuation, and stemming tokens. Next, we will use a method inspired by TF-IDF to assign scores to the words in each cluster's user Twitter bios. Since we want information at the aggregated cluster level, we group the user bios by cluster and look at the cluster's bios as a whole. Essentially, this TF-IDF variant compares the frequency of a word in each cluster's bios with the background frequency of the word over all clusters' bios. The score for a word in a cluster calculated by this method intuitively corresponds to how relevant the word is to the cluster relative to how relevant the word is to the entire audience (e.g. all clusters). For each cluster, this method surfaces the words that are unique to each cluster relative to all other clusters.

More formally, we adopt the following scoring method. Given a cluster c and a word w, we define $N_c(w)$ as the number of occurrences of w in cluster c's Twitter bios and N_c as the total number of words in cluster c's bios. By definition,

$$N_c = \sum_w N_c(w).$$

We can then write the probability of w occurring in cluster c's bios as

$$P(w \mid \text{cluster } c) = \frac{N_c(w)}{N_c}.$$

Similarly, we define $N(w)$ as the total number of occurrences of w in all users' Twitter bios and N as the total number of words in all users' bios. Then, we have the corresponding equations:

$$N = \sum_w N(w)$$
$$P(w) = \frac{N(w)}{N}$$

Now, we define the score of word w for cluster c as

$$s(w \mid \text{cluster } c) = P(w \mid \text{cluster } c) * \log\Big(\frac{P(w \mid \text{cluster } c)}{P(w)}\Big).$$

So, if a word occurs at roughly the same frequency in cluster c as it does overall across all clusters, then the ratio $\frac{P(w|\text{cluster } c)}{P(w)}$ will be close to 1, and its log will be close to 0. For example, if the word "person" is used by exactly half the users in every cluster to describe themselves, then we expect its score $s(w \mid \text{cluster } c)$ for every cluster c to be 0. If a word occurs at a much higher frequency in a cluster c than it does across all other clusters, then $\log(\frac{P(w|\text{cluster } c)}{P(w)})$ will be large. The score is weighted by $P(w \mid \text{cluster } c)$, so rare words will not get surfaced even if they occur more frequently in some clusters than others. For example, if the word "student" is used frequently in cluster c and not in any other, then we would expect its score

for cluster c to be large. On the other hand, if the word "zookeeper" is used more frequently in cluster c than any other cluster but is used infrequently even in cluster c, then its score for cluster c will not be large. We then look at the highest-scoring words for each cluster.

Since this method yields words that are most relevant and distinctive for each cluster, we call these the most informative words. The most informative bio words for each cluster give us insight into the composition of the interest clusters and the different types of people within the audience. Table 3.1 provides examples of several clusters and their most informative bio words. Note that these bio words have been stemmed, causing many to appear differently than they would in the actual bios.

Cluster	**Most Informative Twitter Bio Words**
2	writer, film, actor, black, queer, blm, #blacklivesmatt, matter, tv, nerd, filmmak, screenwrit, comic, write
4	game, gamer, video, youtub, twitch, wrestl, anim, comic, movi, streamer, fan, play, like, i'm, nerd
7	market, founder, design, travel, pr, digit, communic, social, media, women, tech, tweet, communiti, strategist, opinion
8	polit, mom, democrat, liber, former, proud, opinion, dog, progress, wife, #resist, retir, vote, justic, junki
19	indian, india, student, engin, cricket, bangladesh, bollywood, simpl, studi, hai
33	grey, anatomi, account, justin, fan, follow, demi, ela, belieb, stan, scandal, bieber, @justinbieb, dela, lovat
40	#resist #blm #fbr democrat trump resist #theresist dms liber vote proud dm biden #voteblu #bidenharris2020
48	view, justic, tweet, phd, opinion, health, black, feminist, professor, advoc, alum, former, social, research, polici

Table 3.1: Examples of the most informative bio words from some clusters.

Since these words are chosen by the users to describe themselves, they reveal something about the identities of the people in these clusters. So, we can make deductions about the clusters listed in Table 3.1. Cluster 2 appears to be composed of creatives such as filmmakers, writers, and actors. Cluster 4 is a gamer community. Cluster 7 seems to be a group of people in fields such as marketing, design, PR, and communications. Clusters 8 and 40 consist of people who are interested in politics,

and in particular are left-leaning and liberal. Cluster 19 is a population of Indian people, primarily engineering students. Cluster 33 appears to be a collection of pop culture and entertainment fans; fans of the show Grey's Anatomy, the singer Justin Bieber, Kerry Washington's show Scandal, etc. Cluster 48 seems to be comprised of academics or highly-educated people with liberal views. The most informative bio words of all 50 clusters are presented in the appendix.

Returning to the retweet cascades intersected with interest clusters in Figure 3-1, we now know that the entertainment-related KW tweet primarily spreads to cluster 33, a cluster of entertainment fans, and the politics-related KW tweet mostly stays within clusters 8 and 40, both very political clusters. This shows the correlation between how clusters describe themselves and the tweets they respond to.

Once we have these descriptions of users in the various interest clusters, we can use them to interpret any future predictions we make. We can also use these qualitative measures to make generalizations about the clusters.

3.2.3 What Users Tweet About

Another informative way of discovering a user's interests is to look at their tweets. Users tweet about things that they are interested in, so we can access user interests by studying their tweets.

We use the same recent tweets obtained in Section 3.2.1. To determine the interests of the clusters, we group these tweets by cluster. Using the same TF-IDF variant outlined in the previous section, we extract the most informative words from each cluster's recent tweets. Specifically, we apply the same basic scoring method outlined in Section 3.2.2 to the clusters' recent tweets instead of their Twitter bios. Using this, we get the most informative words from each cluster's recent tweets. We also use similar preprocessing steps: tokenize each user bio and remove any stopwords, digits, punctuation, user mentions, and URLs.

Table 3.2 displays examples of clusters with the most informative words from their tweets from 2021. Comparing Tables 3.1 and 3.2, we can see that the most informative words from each cluster's tweets seem to align well with the most informative words

Cluster	Most Informative Recent Tweet Words
2	rt, movie, trans, film, people, think, black, white, asian, gay, episode, movies, made, character, art, police, writing, thinking, queer, read, weird, horror
4	rt, game, #wrestlemania, movie, playstation, wwe, anime, stream, xbox, games, twitch, trailer, #ps4live, ps5, marvel, batman, series, wrestling, match, broadcast, play, gaming
7	join, learn, women, community, virtual, business, event, health, ceo, team, work, marketing, program, tech, leaders, data, digital, companies, founder, future, leadership, global, tips, experience
8	trump, rt, republicans, biden, president, gop, senate, capitol, vote, police, congress, election, white, state, bill, gaetz, america, gun, covid, texas, cruz, georgia, house, democracy, rights, democrats, vaccine, news, breaking, insurrection, law, senator, officer, rep
19	india, indian, rt, modi, bjp, delhi, #modi_job_do, govt, farmers, mumbai, #farmersprotest, cricket, bengal, hindu, #modi_rojgar_do, #ipl2021, pm, exam, minister, tamil, covid, hospital, holi, ipl, #teamindia, #indveng
33	rt, #greysanatomy, #wynonnaearp, #iheartawards, #bringwynonnahome, #station19, album, episode, taylor, harry, #bestfanarmy, song, bts, swift, stan, anatomy, sarah, paulson, grey, scene
40	trump, gaetz, republicans, gop, biden, police, president, capitol, gun, rt, senate, congress, white, america, vote, election, cruz, bill, officer, fox, insurrection, party, racist, news, greene, democrats, state, mcconnell, breaking, texas, house, russia, law, cops, rep
48	black, police, white, people, women, community, health, justice, violence, public, trans, students, asian, workers, communities, pandemic, racism, state, rights, covid, vaccine, research, american, racial, climate, supremacy, history, system, equity, policy, abortion, access, housing, law, data, families, dr, school

Table 3.2: Examples of the most informative recent tweet words from some clusters.

from each cluster's bios. For example, users in clusters 8 and 40 tweet a great deal about political matters and events, which matches how political their Twitter bios are. Similarly, users in cluster 33 tweet mostly about shows and music, and users in cluster 4 tweet predominantly about gaming and gaming-adjacent topics such as streaming. This matches how the users in clusters 33 and 4 categorize themselves as entertainment fans and gamers respectively.

3.2.4 Cluster Representation

For each cluster, we calculate the 1000 most informative words from their recent tweets as in Section 3.2.3. Then, for each of these 1000 words, we take the 200-dimensional GloVe Twitter word embedding, which are generated from the GloVe model [6] trained on 2 billion tweets, and average them. Any word that does not have a corresponding GloVe Twitter word embedding is ignored. This average word embedding will be the representation for the cluster.

This cluster representation implicitly carries information about the structure of the follow graph of the audience, since the clusters are formed based on the audience's follow graph. By using the most informative words from each cluster's tweets, we add information about what each cluster is directly interested in.

Note that we explicitly do not use BERTweet or any other transformer model for the word embeddings of the most informative words. This is because the most informative words for each cluster are a collection of words that do not form a sentence, so there is no inherent order or context to be gained from treating these words as a sentence. So, all of the positional and contextual information that BERTweet and other transformer models would learn, while typically a strong suit of transformers, would be at best irrelevant and at worst detrimental for our task. Thus, we use the GloVe Twitter word embeddings to encode the most informative words extracted from the recent tweets of each cluster.

An advantage of using the averaged GloVe Twitter word embeddings of the most informative words form each cluster's recent tweets as a cluster representation instead of something simpler like one-hot encoding is that this cluster representation is in a continuous space, making it more flexible. What users are interested in and therefore what users tweet about will change over time. Similarly, the types of tweets that users choose to respond to will differ over time. This cluster representation can be adapted to represent the same cluster at different times, thus representing the evolving interests of the same cluster. In addition, this cluster representation is generalizable. Clusters that are interested in similar topics will have similar representations, allowing

us to generalize to clusters beyond those we currently have. If we want to work with a different audience without completely restarting our analysis, we can still represent interest clusters in that audience, even if they are completely different from the interest clusters we have derived from Kerry Washington's audience. Thus, we could still make predictions about entirely new interest clusters.

3.3 Individual Level Audience Characterization

Finally, we develop representations for the individual users in Kerry Washington's audience.

3.3.1 What Users Respond To

We examine the relationship between the users and which KW tweets they choose to respond to in order to generate useful user representations. First, we gather all the users who have responded to any of the KW tweets. Using sklearn, we bicluster these users and the KW tweets [3], which simultaneously clusters rows and columns of a matrix. In this case, we use biclustering to simultaneously group users and KW tweets into clusters. Then, we identify all the unique accounts that these users follow and the unique words in their user bios. We treat these accounts and bio words as the features and use sklearn's feature selection to find the 128 features that are most useful in predicting which bicluster users belong to. In total, we considered 27,518 features. The selected 128 features correspond to the 128 accounts or bio words that are most indicative of which tweets users will respond to.

The 128 features that we select are actually all Twitter accounts followed by users in the audience. This suggests that the following relationships are more indicative of which KW tweets users will respond to than the words in the user bios. Almost all of the 128 accounts are political figures or people who are interested and involved in politics. Specifically, they are all liberal and left-leaning. Table 3.3 shows a few examples of the accounts that were selected as features and their bios.

Feature	Twitter Bio
@RepAdamSchiff	Representing California's 28th Congressional District. Chairman of the House Intelligence Committee (@HouseIntel).
@tedlieu	Husband of Betty, the love of my life. Father of two great kids. USAF veteran. Member of Congress. In that order. Also, empathy is good.
@PreetBharara	Patriotic American & proud immigrant. @Springsteen fan. Banned by Putin, fired by Trump. Former US Atty, SDNY. Host of "Stay Tuned" http://doingjusticebook.com
@MeidasTouch	Producing the best pro-democracy political videos and content. Meidas Media Network. MeidasTouch PAC. Meidas Merch. Because TRUTH is golden.

Table 3.3: Examples of the 128 accounts that were selected as features for individual user representations.

The political nature of nearly all 128 selected accounts implies a strong divide between political and non-political content in Kerry Washington's tweets and audience.

3.3.2 Individual User Representation

For each user, we create a 128-dimensional binary vector representing which features this user has, e.g. which out of the 128 selected accounts the user follows. To construct these vectors, we fetch the full list of friends (accounts followed) of each user using the Twitter API. We will use these binary user vectors as the individual user representations.

This user representation makes use of each individual audience member's following relationships. A downside to this user characterization is that any users who follow the same accounts out of the 128 selected accounts look the exact same, even though they may actually be very different based on the other accounts that they follow.

Chapter 4

Cluster Level Prediction

In our cluster level prediction task, our goal is to predict which interest clusters will respond to a given tweet. An advantage in dissecting the audience into interest clusters is that we can make very interpretable predictions, such as "We predict a politically left-leaning cluster will respond to a tweet regarding the COVID vaccine."

Besides interest in the tweet content, there are multiple factors that affect whether an individual user will engage with a tweet: whether they saw the tweet, how likely they are to engage with tweets in general, etc. If we see that a user has not responded to a tweet, we have no way of distinguishing the user's motivation to not respond; we would not know if it was an intentional choice the user made due to a lack of interest in the tweet, or if it was because the user simply never saw the tweet or never responds to anything on Twitter. As we do not want to model an individual's general behavior on Twitter, we instead aggregate users into clusters based on their interests. Grouping the audience into clusters of similar users would make our predictions more robust to the noise in a non-response signal.

4.1 Data Generation

Our data consists of a feature representation of a KW tweet concatenated with a feature representation of an interest cluster, which we then feed into a classifier to predict whether the cluster will respond to the tweet.

We partition our collection of 709 KW tweets into train and test sets by tweet with a 90-10 split. So, our train set will have 639 KW tweets and our test set will have 70 KW tweets. Every KW tweet has 50 data points, each of which corresponds to an interest cluster. Each of a KW tweet's corresponding 50 data points is a concatenation of the tweet representation and a cluster representation, and has a corresponding binary label indicating whether the cluster responded to the tweet. We further partition the training set into train and validation sets by a 90-10 split.

4.1.1 Label Generation

We must now generate labels for our data points, so we must determine what it means for a cluster to respond to a tweet. While it is very straightforward to determine whether an individual user responds, there is no directly analogous criteria for classifying a response at the cluster level. Clearly, at least one individual in a cluster must respond to a tweet in order for us to say that the entire cluster has responded. So, we will set some threshold: if the number of users in cluster i who respond to a tweet j exceeds this threshold, then we claim that cluster i has responded to tweet j; otherwise, we claim that it has not.

The primary source of ambiguity in such a claim lies in the threshold that we set. Since the different interest clusters are of varying sizes, it seems inaccurate to fix a constant number of users as a threshold. For example, if 20 people in a cluster respond to a tweet, we might want to consider this a cluster response if the cluster contains 100 people but not if the cluster contains 100,000 people. Choosing a percentage of users in a cluster as a threshold would account for varying cluster sizes, but selecting the percentage threshold would be rather arbitrary because we have no way of establishing what a "good" threshold is.

Another source of variability between different interest clusters is the cluster's base rate of response. Members of different clusters may have different tendencies to respond to tweets in general. For example, users in some clusters may retweet every tweet they see, while users in other clusters may never retweet any tweet. Then, just as we did with the issue of varying cluster sizes, we might want to consider 20

responses in a cluster to be a cluster response if the cluster typically never responds to any tweets but not if the cluster regularly responds to tweets.

To resolve both of these ambiguities, we calculate the median number of individual responses in each cluster across the tweets in our training set, and set the threshold for each cluster to be its median number of individual responses. Since this method yields a cluster-specific threshold, it embeds both the size of the cluster and the base response rate of the cluster into the threshold value, thus addressing both causes of confusion. Note that we choose to use the median number of responses rather than the mean because the mean is easily skewed by outliers, and we want our cluster-specific threshold to incorporate information about the cluster's typical number of responses. Examples of clusters with their respective sizes and number of median responses are displayed in Table 4.1.

Cluster	**Cluster Size**	**Number of Median Responses**
2	50,725	9
4	39,752	3
7	31,188	2
8	82,619	15
19	45,854	0
33	24,229	7
40	26,828	7
48	40,993	7

Table 4.1: Examples of the cluster sizes and number of median responses.

Despite how large cluster 19 is, the median number of responses indicates that cluster 19 rarely responds to Kerry Washington's tweets. Similarly, clusters 4 and 7 are also relatively big and respond relatively sparsely. On the other hand, clusters 33 and 40 are almost half the size of cluster 19, and these clusters respond much more frequently.

The overall small magnitude of the median number of responses suggests that each cluster typically does not respond much to most of Kerry Washington's tweets. However, for each cluster, there are likely a few KW tweets that the cluster responds to much more than usual. For example, from Figure 3-1 we can see that the displayed KW tweets elicit many responses from 1 or 2 clusters who are interested in the tweet

topics. So, it is reasonable to think that each cluster has a set of unique KW tweets or certain types of KW tweets that the cluster is excited to see and interact with, and has little motivation to interact with the other KW tweets.

4.2 Baseline

We create two baselines to provide a benchmark for evaluation.

4.2.1 Random Guessing

Our first baseline simply guesses randomly whether each cluster responds to a KW tweet. In other words, this approach assumes each cluster has a 50% probability of responding to a KW tweet.

4.2.2 Naive Frequency-Based

Our second baseline naively uses each cluster's base rate of response as the probability that the cluster responds to a tweet. We define a cluster's base rate of response as the fraction of the training KW tweets that the cluster responds to. This approach assumes that each cluster responds to future KW tweets at the same rate that it has responded to past KW tweets, i.e. each cluster maintains its rate of response to KW tweets over time.

4.3 Simple Model

We construct a model that uses simple representations of the tweet and cluster. To represent the KW tweet, we tokenize the tweet using Tensorflow's text preprocessing tokenizer. Then, we take the 200-dimensional GloVe Twitter word embedding of each token in the tweet and average them to get a final 200-dimensional vector. Any token that does not have a corresponding GloVe Twitter word embedding is ignored.

Note that other than tokenization, we do not perform any preprocessing on the tweet. We do this to be consistent with the preprocessing in the next section. We

also want to preserve as much information from the KW tweet as possible, since the tweet representation will provide most of the information.

To represent the clusters, we use a 50-dimensional one-hot encoding scheme. The input to this model is a 250-dimensional vector formed by concatenating the tweet and cluster representations.

The classifier itself is comprised of two dense layers with ReLU activation, the first with 128 units and the second with 64 units, and an output layer with sigmoid activation. We train the model for at most 50 epochs using early stopping by monitoring validation loss with a patience of 10.

4.4 Fine-Tuning BERTweet

4.4.1 Data Generation

As discussed in Section 2.3, we use BERTweet to generate embeddings for the KW tweets. We use the corresponding BERTweet tokenizer to pad, truncate, and encode each KW tweet, yielding valid input sequences to BERTweet. BERTweet takes in sequences of length no greater than 128, so we pad and truncate our input sequences to 128 tokens. Additionally, we have the 6 features derived in Section 2.3 that we will be using.

For the cluster representation, we take the 1000 most informative words from each cluster's recent tweets and average their respective 200-dimensional GloVe Twitter word embeddings, as described in Section 3.2.4.

So, our model has three inputs: the tweet tokens, the selected tweet features, and the cluster representation.

4.4.2 Model Architecture

The model architecture is shown in Figure 4-1. As mentioned, we use BERTweet to produce a representation for the KW tweet. The first input, the tweet tokens, is fed into BERTweet, and we extract the final hidden state of the CLS token to use as the

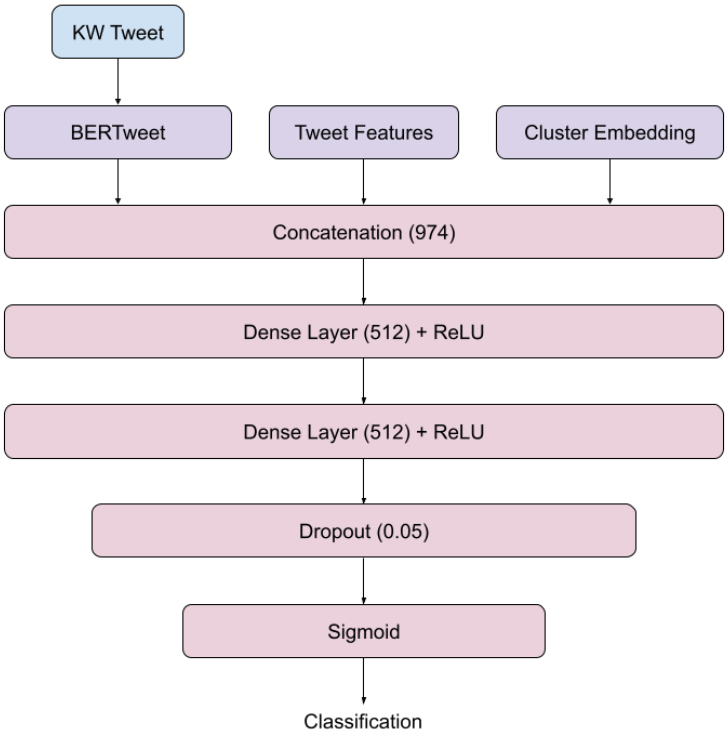


Figure 4-1: Cluster model architecture diagram

tweet embedding. We use a concatenation layer to concatenate the tweet embedding with the other two inputs, the selected tweet features and the cluster embedding. Then, we add on top a simple classifier consisting of two dense layers with ReLU activation (both with 512 units), a dropout layer with a dropout rate of 5%, and a sigmoid output layer.

4.4.3 Model Training

We first freeze all parameters of BERTweet and train only the classifier on top for 10 epochs. During this first training step, we use the Adam optimizer with a learning rate of 1e-4. Then, we unfreeze the last layer of BERTweet and train the entire model for 4 epochs, which simultaneously trains the classifier on top and fine-tunes BERTweet's outputted tweet embeddings. During this second fine-tuning step, we use the Adam optimizer with a linearly decaying learning rate and an initial learning rate of 2e-5. For both training steps, we use binary cross-entropy loss.

4.5 Results

Table 4.2 compares the performance metrics of the described models.

Model	Precision	Recall	F1	AUC
Random	0.4315	0.4957	0.4614	0.4975
Frequency-Based	0.4871	0.435	0.4596	0.5246
Simple	0.5839	0.5569	0.5701	0.6503
Fine-Tuned BERTweet	0.6973	0.6725	0.6847	0.7809

Table 4.2: Comparison of cluster model results

Our fine-tuned BERTweet model significantly outperforms both baseline models and the simple model.

With respect to our prediction task, a false positive means that a cluster we predict will respond does not respond, and a false negative means that a cluster we predict will not respond does respond. In general, it seems much better to have unexpected clusters respond than to have clusters predicted to respond not respond. That is, we want the clusters we predict will respond to a tweet to actually respond to that tweet, and we put less emphasis on the issue of missing clusters who would respond. For example, if our goal is to target a specific part of an audience, having false negatives would not hurt, while having false positives could mean that we ultimately fail to achieve our overall goal of engaging with a target group. Similarly, if our goal is to maximize audience interaction for a tweet, we would much prefer having false negatives over having false positives.

We now break down our model performance across KW tweets in Tables 4.3 and 4.4. Table 4.3 contains examples of KW tweets that our fine-tuned BERTweet model performs relatively well on. All these examples are related to social issues. They are also popular tweets, and generate responses across almost all clusters. Since these KW tweets are popular across clusters, it is possible that the model has learned to predict that every cluster responds to a social-themed tweet.

However, note that the text of the second, third, and fifth KW tweets in Table 4.3 do not seem to explicitly have any relation to social matters. Instead, all three of these tweets incorporate their connection to social issues in a hashtag. This might

KW Tweet	Precision	Recall	F1
We cannot be silenced. We cannot let them take away our voice. And our vote.	0.9400	1.0000	0.9691
Please remember to breathe today. To take care of yourself. Your bodies. And your minds. And please continue to say his name. #DaunteWright	0.9783	0.9375	0.9574
156 years later and #Juneteenth is becoming a Federal Holiday	0.9000	1.0000	0.9474
Getting vaccinated is not a political issue. It is safe. Republican AND Democratic officials have been vaccinated to keep themselves safe. Now its your turn. Please go get vaccinated if you can!!!! ❤️🙏 https://t.co/wmEasIFhmX	0.9535	0.9318	0.9425
#RemoveTrumpNow	0.9773	0.8776	0.9247
Today we honor and celebrate Dr. Martin Luther King. Whose prolific words ring as true now as they did then. 🙌🏽❤️🙏🏽	0.8600	1.0000	0.9247

Table 4.3: Tweet-specific performance of the cluster level fine-tuned BERTweet model on several test KW tweets the model does well on.

suggest that our model has learned to associate hashtags with popular tweets. In general, this seems to be a reasonable correlation, as hashtags are often used to reference trending topics. These three tweets, for example, all use hashtags to reference famous politically- and racially-themed events, which usually provoke a great deal of conversation.

KW Tweet	Precision	Recall	F1
#AURATExKERRY. New collection drops tomorrow. STAY TUNED.	0.0851	1.0000	0.1569
🗣️ THREAD https://t.co/2C8XCMbnKx	0.0938	1.0000	0.1714
Happy Birthday @SelenaGomez!!!! ❤️ https://t.co/3xMLMWJgY4	0.2941	0.2778	0.2857
Happy Birthday to my BX sister @iamcardib https://t.co/BqlT3qQmFJ	0.2449	1.0000	0.3934

Table 4.4: Tweet-specific performance of the cluster level fine-tuned BERTweet model on several test KW tweets the model does well on.

Table 4.4 provides examples of KW tweets that the model does not perform well

on. The first KW tweet in Table 4.4 contains a hashtag. If the model does learn to associate hashtags with popular tweets as we speculate, then the model would also predict this KW tweet evokes responses across many clusters. This would be consistent with the model's for this tweet.

The second tweet in Table 4.4 happens to be a quote tweet. The original tweet that Kerry Washington quotes is presented in Table 4.5. This original tweet is heavily political in nature. Recall that we concatenated the original tweet to the KW quote tweet to provide context as discussed in Section 2.3.1. Consequently, the majority of the tweet text we feed to our model actually belongs to the original tweet rather than the KW tweet. Since we truncate our input tweet text to 128 tokens, it is even possible that none of the input into our model comes from Kerry Washington's quote tweet. The model then essentially ends up predicting who would respond to the original tweet. As social issue-related tweets tend to generate responses across many clusters, it is likely that the model predicts that most clusters will respond to the KW tweet, resulting in the low precision for this tweet.

KW Response	**Original Tweet**
🗣 THREAD https://t.co/2C8XCMbnKx	Four years ago tonight, Mitch McConnell silenced me for reading Coretta Scott Kings letter. But remember this: every Republican in the chamber that night voted to shut me up. And every Republican voted for Trumps AG who was too racist to become a judge in the 1980s.

Table 4.5: Kerry Washington's quote tweet that the cluster level fine-tuned BERTweet model performs poorly on, along with the original quoted tweet.

The other two KW tweets in Table 4.4 wish someone happy birthday, which Kerry Washington does relatively frequently; almost 8% of the 709 KW tweets consist of birthday wishes. The responses to these birthday tweets would depend mostly on the recipient of the birthday wish, and we do not provide our model with any information about this person. As a result, we expect it to be extremely challenging for our model to predict who will respond to a tweet wishing someone a happy birthday. Our model cannot use any person-specific knowledge to make predictions, and it cannot differ-

entiate any information acquired during training about the various “happy birthday” tweets directed to various people. During evaluation, the model can then only apply the combined information it learned from all the birthday tweets in the training set, which would be inaccurate for the birthday tweets in the test set.

Chapter 5

Individual Level Prediction

In our individual level prediction task, our goal is to predict which users will respond to a given tweet. An advantage of the individual prediction task is that there is no ambiguity over what a response should be classified as: we know that a user responds to a KW tweet if they retweet, reply, or quote the tweet. Thus, our label generation is extremely straightforward.

5.1 Data Generation

As in the cluster level prediction task, our input data consists of a representation of the tweet combined with a representation of the audience, in this case an individual user. We use the exact same tweet representations as in Section 4.4.1, and we use the user representations from Section 3.3.2.

An issue with our data is that the negative examples greatly outweigh the positive examples. In addition, as indicated in Chapter 4, the signal from these negative examples is noisy. To help mitigate this, we will strategically select examples from the negative class in order to try extracting a stronger signal from the negative examples.

First, we narrow down the audience to the 187,609 unique users who have responded to any of the 709 KW tweets. The positive examples already only involve users from this set. So, our prediction task essentially becomes to predict which users will respond to future KW tweets out of the users who have responded to her tweets

in the past.

Then, we identify for each KW tweet a subset of the audience that are active on Twitter in the first 24 hours after the KW tweet has been posted, and intersect this subset with the 187,609 responders. To get each subset of users who were active for each KW tweet, we pulled the tweet histories for every user in the audience. For each KW tweet, we find the users who have exhibited any tangible evidence of activity on Twitter, e.g. tweeted, retweeted, replied, etc., in the first 24 hours after the tweet was posted. We treat this set of users who have tweeted within 24 hours of the KW tweet as the group of users who were exposed to the KW tweet. So, we select our negative examples for each KW tweet from the users in the set exposed to the tweet who did not respond to the tweet. This narrows down our negative examples from every user in the audience who did not respond to the set of users who likely saw the tweet and did not respond. This tries to address the issue where we do not know if a user does not respond because they did not see the tweet and chose not to respond, or if they simply did not see the tweet. By selecting negative examples from a set of users who likely saw the KW tweet, we are eliminating not seeing the tweet as a reason for not responding.

Finally, for each KW tweet we sample at most two negative examples for each positive example. This prevents the negative examples from overwhelming the positive examples.

We partition our KW tweets into training, validation, and test sets with an 80-10-10 split. Now, since we have tweet-specific examples, the number of data points per KW tweet will vary.

5.2 Baseline

We use the same two baselines from Section 4.2 to provide a benchmark for evaluation.

5.2.1 Random Guessing

Our first baseline simply guesses randomly whether each user responds to a KW tweet. In other words, this approach assumes each user has a 50% probability of responding to a KW tweet.

5.2.2 Naive Frequency-Based

Our second baseline naively uses each user's base rate of response as the probability that the user responds to a tweet. We define a user's base rate of response as the fraction of the training KW tweets that the user responds to. This approach assumes that each user responds to future KW tweets at the same rate that it has responded to past KW tweets, i.e. each user maintains its rate of response to KW tweets over time.

5.3 Fine-Tuning BERTweet

We essentially use the same model as in Section 4.4.

5.3.1 Data Generation

We generate inputs to BERTweet from the KW tweets and then extract tweet embeddings from BERTweet. Using the BERTweet tokenizer, we pad, truncate, and encode each KW tweet, yielding valid input sequences to BERTweet. BERTweet takes in sequences of length no greater than 128, so we pad and truncate our input sequences to 128 tokens. We also include the 6 features derived in Section 2.3.

For the user representation, we use the 128-dimensional binary vectors representing the following relationships of the individual users as obtained in Section 3.3.2.

So, our model has three inputs: the tweet tokens, the selected tweet features, and the user representation.

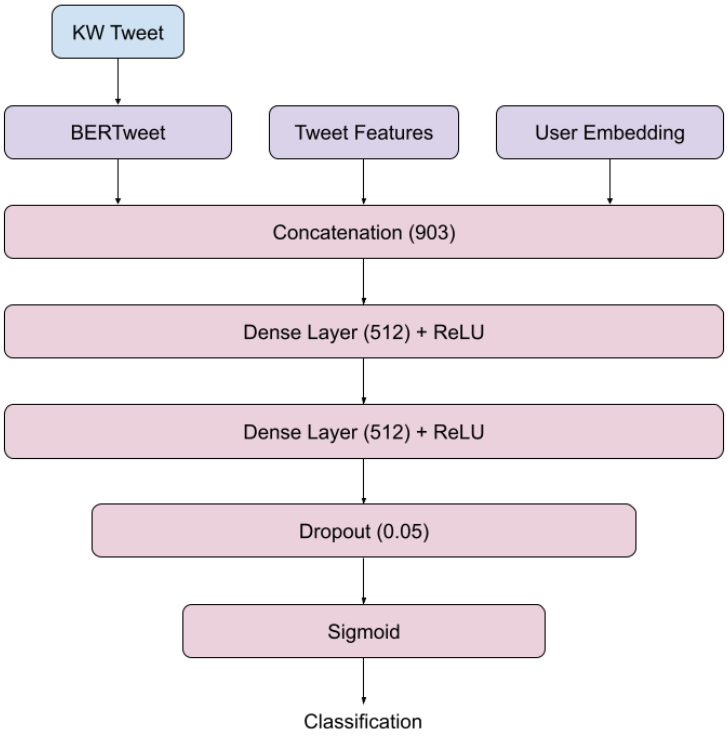


Figure 5-1: Individual model architecture diagram

5.3.2 Model Architecture

The model architecture is shown in Figure 5-1. We feed the tweet tokens into BERTweet and extract the final hidden state of the CLS token to use as the tweet embedding. Then, we concatenate the resulting tweet embedding with the other two inputs, the selected tweet features and the cluster embedding. Finally, we add on top a simple classifier consisting of two dense layers with ReLU activation (both with 512 units), a dropout layer with a dropout rate of 5%, and a sigmoid output layer.

5.3.3 Model Training

First, we freeze all parameters of BERTweet and train only the classifier on top for 10 epochs. During this first training step, we use the Adam optimizer with a learning rate of 1e-4. Next, we unfreeze the last layer of BERTweet and train the entire model for 4 epochs, which simultaneously trains the classifier on top and fine-tunes BERTweet's outputted tweet embeddings. During this second fine-tuning step, we

use the Adam optimizer with a linearly decaying learning rate and an initial learning rate of 2e-5. For both training steps, we use binary cross-entropy loss.

5.4 Results

Table 5.1 displays the metrics for each model.

Model	**Precision**	**Recall**	**F1**	**AUC**
Random	0.3591	0.5006	0.4182	0.4992
Frequency-Based	0.9854	0.0042	0.0083	0.5895
Fine-Tuned BERTweet	0.9334	0.5740	0.7109	0.8170

Table 5.1: Comparison of individual model results

Our fine-tuned BERTweet model significantly outperforms both baseline models. As in the cluster level prediction task, we again prioritize minimizing our false positives over false negatives. We break down our model's performance on several KW tweets in Tables 5.2 and 5.3.

KW Tweet	**Precision**	**Recall**	**F1**
No unity without accountability. Period.	0.9824	0.8016	0.8829
There are so many talented humans who make the movies and TV shows we love! You may not see them on screen but they are the magic makers and the glue that holds any set together. I urge the #IATSE to hear them. And I stand with my brothers & sisters in this strike #IASolidarity	0.9379	0.6189	0.7457
Floridians - this is happening TODAY! Call your senators and vote no now!!! https://t.co/lLZhLu30s9	0.9080	0.6320	0.7453
We cannot be silenced. We cannot let them take away our voice. And our vote.	0.9224	0.5983	0.7258
#RemoveTrumpNow	0.822830	0.573877	0.676166
Please use your voice. Call 202-499-6085 and demand senators pass the George Floyd Policing Act	0.955801	0.516418	0.670543

Table 5.2: Tweet-specific performance of the individual level fine-tuned BERTweet model on several test KW tweets the model does well on.

Table 5.2 lists examples of KW tweets that the fine-tuned BERTweet model performs relatively well on. These KW tweets seem to be primarily about social issues. This may be because the tweets related to social issues tend to have a narrower range of topics, or more well-defined and well-contained topics such as political or racial issues. It is highly likely that KW tweets in the training set that are about social issues would have similar fine-tuned BERTweet embeddings to KW tweets in the test set that are about social issues. So, it may be easier for our model to learn a relationship between these KW tweets and the types of users who respond to these tweets. Kerry Washington's social impact-related tweets also tend to garner more responses, which could contribute to our model's performance on these tweets.

On the other hand, the KW tweets that the model does not perform well on (as shown in Table 5.3) vary much more in topic and language, so the fine-tuned BERTweet embeddings of these KW tweets are likely all very different. As a result, the model may not have learned anything useful during training that it can apply to the KW test tweets.

Additionally, a limitation of our user representation is that users who follow the same accounts out of the 128 selected features appear the same, as noted in Section 3.3.2. Since almost all of the 128 selected accounts are related to politics, it may be that the model can become good at predicting responders to political tweets, as demonstrated by Table 5.2, and bad at predicting responders to other types of tweets, as demonstrated by Table 5.3.

Two of the KW tweets in Table 5.3 are birthday tweets. Just as we noted in Section 4.5, we expect model performance on these tweets to be poor.

The other two KW tweets listed in Table 5.3 contain very little text information. Most of the context of the tweet is provided via the link, which can be an attached photo, video, or URL. In these cases, the responders depend almost entirely on the material contained in the link. We only provide information on the quantity of URLs and whether the tweet contains an attached photo or video, so our model cannot use any of the relevant information in the link to make predictions. This would explain the poor performance on these tweets.

KW Tweet	Precision	Recall	F1
🎤⬇️ https://t.co/n3ni6OAGEL	1.0000	0.1159	0.2078
Here comes the 🌟👉🏽 https://t.co/fBKx0PtJXQ	0.7719	0.1477	0.2479
We got a #ScandalFam birthday in the house!!!!!!! Gladiators help me send @darbysofficial soooooooo much birthday love today. I miss you so much and can't wait to be reunited with your kind energy and positive spirit. Love you Darbs!!!!!!! HAPPY BIRTHDAY!!!!! 💞😍🎁🎉🪂🎈 https://t.co/X0kN3ZpYVx	0.7705	0.1643	0.2709
Wishing this badass titan, producer, mom, friend, & sister of mine @shondarhimes, a very HAPPY BIRTHDAY! 🎈🥳 Without her dreams & creativity, some of the greatest fake love stories of all time wouldnt exist (looking @ you @tonygoldwyn 😜). I ❤️ you & I am SO happy you were born! https://t.co/piHOtmq7HX	0.950000	0.177570	0.299213

Table 5.3: Tweet-specific performance of the individual level fine-tuned BERTweet model on several test KW tweets the model does poorly on.

Even in the examples where the model does not perform well, the precision is relatively high; the recall is where the model performs poorly. Since we seek to minimize false positives more than false negatives, this is the scenario we prefer.

Chapter 6

Conclusion

6.1 Contributions

In this thesis, we have explored multiple views of Kerry Washington's tweets and audience, as well as developed models to predict which audience members would interact with her tweets. We were able to show that our models could successfully predict who responds to tweets, particularly to tweets regarding social issues.

With these results, we can help influencers draft messages that appeal to a wider audience or spark discussion in a specific group of people. By knowing in advance who might engage with a tweet, we can also select an optimal messenger to deliver a tweet, whether we want to spread a message as broadly as possible or expose a particular part of an audience to some information. These applications all seek to modify the flow of information on Twitter.

Our analysis of the spread of tweets throughout the interest clusters suggested that we could more efficiently direct the propagation of information by identifying and targeting users who play influential roles within their respective interest clusters. Inducing responses from a few influential users within an interest cluster may be easier to achieve than aiming for a cluster response. The models we have built for predicting audience responses can then be employed to construct tweets that will evoke responses from these selected users.

We can apply these findings beyond Kerry Washington's tweets and audience.

The methods in this thesis can be utilized to similarly examine other influencers and thus affect how a group of influencers behave on Twitter.

6.2 Discussion

It is important to note the limitations of our approaches. We had a relatively limited amount of data. Since we only utilized one influencer's data and were subject to Twitter's Historic Powertrack API quotas, the set of tweets that we worked with was relatively small. Moreover, only looking at a single influencer's tweets may not yield generalizable observations.

On the cluster level side, there is uncertainty in the interest clusters we derived, as separating users into well-defined groups by interests is an inherently difficult problem. Since following relationships are an approximation for user interest, we may not obtain clusters that are cleanly distinguished by user interests. This directly translates into uncertainty in our data labels. Establishing a threshold for a cluster response also adds uncertainty to our label generation process. A cluster's median number of responses may not be a good threshold to define a cluster response with. These uncertainties introduce a level of instability to the overall cluster level prediction task and thus the results we achieved.

With the individual level prediction task, we have attempted to resolve the issue of imbalanced classes by imposing restrictions on the class of negative examples. Specifically, we limit the set of users to those who have responded to a KW tweet, and we also sample at most 2 negative examples for each positive example. We are evaluating our individual prediction model on a test set that is quite restricted, and so our results must be interpreted accordingly.

Our method of identifying users who have been exposed to a KW tweet is also noisy. By interpreting a user's tweet within 24 hours of a KW tweet as an indication that the user saw the KW tweet, it misses users who were active on Twitter but did not post any tweets, and it also assumes that users who tweeted did see the KW tweet. While this method captures some intuition about a user's likelihood of having seen

the KW tweet, it does not provide any exact information. Additionally, this method is computationally expensive. To get the list of users who have tweeted within 24 hours of each of the 709 KW tweets, we must fetch the tweet history for the entire audience of 2 million users from January 1 to November 20, 2021. Twitter also only allows access to a user's most recent 3200 tweets. so if any users have posted more than 3200 tweets over these 11 months, we have no knowledge of their activity levels prior to these 3200 tweets. Then, we could have inaccurate lists of active users for some KW tweets.

6.3 Future Work

This thesis has only begun to scratch the surface of the potential work that can be done in this area. In the future, there are multiple directions in which this project could be improved and developed further.

An area for improvement in the cluster level prediction task is the audience clustering step. Instead of exclusively using accounts followed by users in the audience as a basis for clustering, we can incorporate other factors to cluster the audience. For example, we can use some combination of user tweet, hashtag, URL, retweeting, and following similarities [17] as clustering metrics. As these would provide a more thorough representation of user interest, they would likely also provide a more accurate clustering and thus improve performance in the cluster level prediction task.

To help add interpretability into the predictions of the individual level prediction task, we can perform some post-processing on the predictions. A simple way to gain some insight into the users who we predict will respond to a KW tweet is to look at which cluster each user belongs to. For a more detailed perspective on these users, we can cluster them and look at each resulting cluster's most informative Twitter bio words or recent tweet words. We would then be able to understand what types of users respond to each tweet based off our individual user predictions.

A natural next step is to look beyond the tweet and audience, which we have analyzed in this thesis, and to consider the tweeter as the third factor that affects

tweet engagement. Furthermore, we can extend this work by predicting the sentiment along with the existence of a response from a cluster or user. We can also utilize these predictions to make recommendations about messenger or language choices in order to target different parts of an audience. A practical application of these results would be to build a system that influencers can use to see how altering the language or content of their tweets could help them reach different people in their audience, and to receive suggestions about the language or content choice that will maximize audience engagement. These areas of subsequent work could help shape and guide the social impact of a network of influencers on Twitter.

Appendix A

Tables

Cluster	Most Informative Twitter Bio Words
1	mom, news, wife, sport, fan, dog, anchor, lover, retir, husband, report, opinion, conserv, proud, famili
2	writer, 🌈, film, actor, black, queer, blm, #blacklivesmatt, matter, tv, nerd, filmmak, screenwrit, comic, write
4	game, gamer, video, youtub, twitch, wrestl, anim, comic, movi, streamer, fan, play, like, i'm, nerd
5	movi, tv, love, i'm, music, lover, show, girl, like, anim, thing, fan, obsess, life, addict
6	god, ghana, footbal, unit, soccer, manchest, music, chelsea, fc, sport, simpl, ghanaian, liverpool, tanzania, arsenal
7	market, founder, design, travel, pr, digit, communic, social, media, women, tech, tweet, communiti, strategist, opinion
8	polit, mom, democrat, liber, former, proud, opinion, dog, progress, wife, #resist, retir, vote, justic, junki
9	view, london, mum, uk, actor, film, theatr, director, writer, manag, research, feminist, british, rep, creativ
10	temporarili, unavail, violat, polici, twitter, account, i'm, guy, learn, man, soy, like, media

11	man, i'm, laid, money, fun, get, guy, cool, back, father, real, like, im, ig
12	mom, fashion, blogger, beauti, wife, makeup, design, lifestyl, blog, style, lover, jewelri, product, shop, travel
13	love, i'm, music, sport, follow, art, fashion, instagram, televis, news, funni, de, entertain, like
15	ig, black, sc, i'm, rap, love, beauti, snapchat, hip-hop, vibe, ✊, live, life, instagram, youtub
16	coach, basketbal, sport, athlet, footbal, father, alum, univers, school, head, high, husband, state, mom, colleg
17	god, nigeria, lover, nigerian, entrepreneur, engin, simpl, arsenal, easi, cool, analyst, music, chelsea, fc, consult
18	ig, sc, alumna, snapchat, follow, black, amosc, instagram, rip, 🙏, snap, ✊, futur, #blacklivesmatt, insta
19	indian, india, student, engin, cricket, bangladesh, bollywood, simpl, studi, hai
20	mom, lover, dog, wife, cat, enthusiast, nerd, human, feminist, anim, coffe, polit, liber, thing, tv
21	🌈, gay, drag, guy, #blacklivesmatt, queer, live, gaymer, actor, dad, pop, tran, matter, black
22	mother, love, wife, god, life, educ, sister, friend, mom, daughter, live, famili, retir, black, news
23	kenyan, kenya, god, nairobi, africa, enthusiast, fear, life, simpl, believ, entrepreneur, passion, develop, engin, humbl
24	love, mom, wife, life, lover, famili, girl, friend, mother, marri, live, dog, fan, countri, i'm
25	author, help, coach, market, speaker, busi, writer, entrepreneur, design, consult, inspir, media, brand, onlin, founder

26	god, author, mother, speaker, pastor, hair, life, wife, coach, entrepreneur, jesus, motiv, owner, woman, love
28	actor, canadian, film, canada, toronto, actress, produc, director, mom, lover, writer, communic, travel, hockey, view
30	artist, ig, follow, music, produc, im, songwrit, record, rapper, dj, book, ceo, contact, singer, email
31	educ, teacher, school, black, princip, author, learner, writer, wife, advoc, grade, elementari, mom, student, leader
32	zimbabwean, zimbabw, soccer, polit, god, entrepreneur, news, christian, technolog, develop, african, engin, music, sport, govern
33	grey, anatomi, account, justin, fan, follow, demi, ela, belieb, stan, scandal, bieber, @justinbieb, dela, lovat
34	pakistan, pakistani, muslim, khan, engin, student, allah, lahor, cricket, pak
37	mum, mummi, view, london, irish, fan, love, uk, ireland, xx, liverpool, rugbi, fc, live, dublin
38	sport, fan, footbal, father, basketbal, nfl, husband, nba, cowboy, basebal, famili, go, laker, man, love
39	love, follow, instagram, snapchat, i'm, like, music, youtub, insta, hi, life, sc, justin, bts, sing
40	#resist #blm #fbr democrat trump resist #theresist dms liber vote proud dm biden #voteblu #bidenharris2020
41	god, love, african, fear, south, life, i'm, daughter, child, music, africa, lover, humbl, radio, entertain
42	uganda, ugandan, god, kampala, simpl, engin, ug, passion, africa, develop, african, lawyer, self, busi, human
43	ig, gay, black, freak, guy, dude, i'm, model, kik, 🌈, sc, cool, fun
44	love, life, i'm, live, mother, fun, beauti, famili, god, music, im, laugh, fashion, bless, enjoy

45	simpl, aldub, sa, love, yang, ako, kathniel, ang, ng, happi, forev, lang, fan, na, bts
47	follow, #maga, music, trump, conserv, bir, youtub, back, i'm, like, und, name, countri, maga, channel
48	view, justic, tweet, phd, opinion, health, black, feminist, professor, advoc, alum, former, social, research, polici

Table A.1: The most informative Twitter bio words for all 50 clusters. Note that we have omitted all clusters with exclusively non-English most informative bio words.

Bibliography

[1] David Alvarez-Melis and Martin Saveski. Topic Modeling in Twitter: Aggregating Tweets by Conversations. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*. AAAI Press, 2016.

[2] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? *Proceedings of the 23rd international conference on World wide web - WWW 14*, 2014.

[3] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning, 2001.

[4] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. Prediction of Retweet Cascade Size over Time. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12. Association for Computing Machinery, 2012.

[5] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. *CoRR*, abs/2005.10200, 2020.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[7] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. RT to Win! Predicting Message Propagation in Twitter. In *ICWSM*, 2011.

[8] Eliana Sanandres, Raimundo Llanos, and Camilo Madariaga Orozco. Topic Modeling of Twitter Conversations. 06 2018.

[9] Sarah Shugars and Nicholas Beauchamp. Why Keep Arguing? Predicting Engagement in Political Conversations Online. *SAGE Open*, 9(1), 2019.

[10] Konstantinos Sotiropoulos, John W. Byers, Polyvios Pratikakis, and Charalampos E. Tsourakakis. TwitterMancer: Predicting Interactions on Twitter Accurately. *CoRR*, abs/1904.11119, 2019.

[11] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184, 2010.

[12] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583, 2019.

[13] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[14] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank: Finding Topic-Sensitive Influential Twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10. Association for Computing Machinery, 2010.

[15] T. Zaman, R. Herbrich, Jurgen Van Gael, and David Stern. Predicting Information Spreading in Twitter. 2010.

[16] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. A Bayesian Approach for Predicting the Popularity of Tweets. *CoRR*, abs/1304.6777, 2013.

[17] Yang Zhang, Yao Wu, and Qing Yang. Community Discovery in Twitter Based on User Interests. *Journal of Computational Information Systems*, 8, 03 2012.