Diversity from Emojis and Keywords in Social Media

Melanie Swartz
George Mason University, Department of Computational and Data Sciences, Fairfax, VA
mswartz2@gmu.edu

Andrew Crooks
George Mason University, Department of Computational and Data Sciences, Fairfax, VA
acrooks2@gmu.edu

William G. Kennedy
George Mason University, Department of Computational and Data Sciences, Fairfax, VA
wkennedy@gmu.edu

ABSTRACT

Social media is a popular source for political communication and user engagement around social and political issues. While the diversity of the population participating in social and political events in person are often considered for social science research, measuring the diversity representation within online communities is not a common part of social media analysis. This paper attempts to fill that gap and presents a methodology for labeling and analyzing diversity in a social media sample based on emojis and keywords associated with gender, skin tone, sexual orientation, religion, and political ideology. We analyze the trends of diversity related themes and the diversity of users engaging in the online political community during the leadup to the 2018 U.S. midterm elections. Our results reveal patterns along diversity themes that otherwise would have been lost in the volume of content. Further, the diversity composition of our sample of online users rallying around political campaigns was similar to those measured in exit polls on election day. The diversity language model and methodology for diversity analysis presented in this paper can be adapted to other languages and applied to other research domains to provide social media researchers a valuable lens to identify the diversity of voices and topics of interest for the less-represented populations participating in an online social community.

CCS CONCEPTS

• **Human-centered computing** → Collaborative and social computing; • **Social and professional topics** → User characteristics; • **Applied computing** → Law, social and behavioral sciences.

KEYWORDS

Social media, emoji, diversity, elections, political campaigns

ACM Reference Format:
Melanie Swartz, Andrew Crooks, and William G. Kennedy. 2020. Diversity from Emojis and Keywords in Social Media. In *International Conference on Social Media and Society (SMSociety '20), July 22–24, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3400806.3400818

SMSociety '20, July 22–24, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7688-4/20/07…$15.00
https://doi.org/10.1145/3400806.3400818

1 INTRODUCTION

Social media studies provide insights on themes contained within social media content and user interactions across a variety of topics including, for example, natural disasters [10], vaccinations [45], and politics [36]. While social media analysis has been used to study a variety of social and political issues, there has been less attention given to measuring the diversity represented by the users and content within the social media sample for these various studies. Applying a diversity lens to social media analysis enables researchers to better understand the diversity representation of the population being studied as well as to identify diversity related themes within the social media content. This is particularly important with respect to social media and politics. In an era when news and political leaders are using social media to deliver their messages [3, 34] and political groups use social media to rally support or engagement [39, 44], it has never been more important to ensure that the diverse population of a nation is being reached and the voices of less represented populations in online social-political communities are not lost in the noise [20, 22].

To understand the political landscape of a country, including the concerns of the population and composition of political parties, traditional research methods are popular because they are designed to be rigorous, targeted, statistically valid, and typically representative of the diverse populations interviewed or surveyed [19, 31]. With social media now comprising a large part of political activity and campaigning [7, 37, 44], these formal survey methods may not adequately capture or account for the topics and concerns expressed in less formal styles and behaviors of communication (e.g. slang, emotion, sarcasm, gestures) in social media [16, 17]. Studying social media presents its own set of opportunities and challenges [18, 38]. Many approaches that study the diversity and demographics of social media users rely on location-based information associated with where content is posted [15, 35]. Often researchers will infer demographics and diversity attributes of users based on the location of the user's profile or content and compare it with other locational datasets for the same geographic area such as census or voter statistics aggregated at varying scales of geography [2, 18, 31]. Using location from social media content relies on the provider of the platform as well as the individual user settings. Accuracy and precision of this location information varies greatly and currently ranges from precise coordinates to broad geographic areas such as a city or country [29]. However, as the availability of precise geolocation information varies substantially across platforms and is becoming less available due to privacy concerns [11], alternative approaches are needed to explore diversity within social media communities and datasets.

To fill this gap, this paper presents a novel method using a diversity language model to associate diversity related attributes to social media user accounts and content by analyzing the emojis and keywords used. We apply this model to publicly available tweets that contain keywords related to American politics, specifically the 2018 U.S. midterm. Our methodology for diversity analysis is then applied to identify the groups of social media users and trends in content with similar diversity attributes. Our results reveal patterns of social media engagement across political lines among the diverse populations that otherwise would not have been apparent if we only analyzed the content for key political terms without taking diversity of the users into account.

The three main contributions of this paper are: (1) the development of a diversity language model based on the use of emojis and keywords, (2) the development of a methodology for diversity analysis to label and analyze diversity attributes within a social media sample, and (3) applying the methodology to analyze the diversity of the online community using political party campaign slogans associated with American politics during the lead-up to the 2018 U.S. midterm elections. The remainder of this paper presents a review of related research with respect to diversity and the use of emojis in social media. This is followed by a description of the datasets collected for this research and the methodology to develop a diversity language model and conduct diversity analysis. Then we present and discuss the results of our analysis of the 2018 U.S. midterm elections and conclude the paper with areas for further research.

2 BACKGROUND

The diversity of political party membership and engagement are often measured and analyzed by methods such as interviews, surveys, voter registration, exit polling, and attendance of (e.g. [13, 33, 42]). However, while much political activity takes place in an online setting, such as social media [7, 44], there has been limited research or methods developed to measure and compare the diversity of online political userbase and engagement. While there are challenges with working with social media data with regards to studying politics [12, 22], it is important to try to observe the diversity of a social media sample [18, 31]. As bots and trolls try to influence social and political outcomes with the creation of accounts and the use of language and characteristics similar to the groups they are trying to influence [1, 25], it can be challenging to accurately measure the true identity and diversity of social media user presence [12]. Identifying diversity attributes associated with these accounts and content may reveal which groups are being targeted and in what way.

In addition, viral content, retweets, influencers, and organized campaigns may drown out unique and relevant content of less represented populations [20, 23]. Analyzing the diversity of a social media sample of users can potentially provide cues about the groups engaging with content and enable these voices to be heard even within the massive volumes of social media content. This analysis can also serve as a baseline measurement of diversity associated with topics and users in social media which can then be compared over time to identify behaviors and accounts associated with bots, trolls, or proliferators of fake or viral content [25].

For this research we compare the diversity of the user base sharing content containing political campaign slogans and election-related themes pertaining to the U.S. Democratic and Republican party activities. The goal of this research is to characterize the diversity of the population of users that utilized election related or political slogans or phrases in their social media content, specifically related to the U.S. 2018 midterm elections. While the diversity and demographics of the political base at rallies and events is measured directly via more formal approaches (e.g. [13]), here we attempt to measure the diversity of the user base engaging in online political communities solely with digital online social media data.

There are many concerns with studying social and political issues using social media data. A common one is bias arising from the differences in social media users compared to the real-life population [6, 20]. While identifying the demographic makeup of the user base of a social media platform is already an area of research [24, 28, 35], it should also be noted that the users represented in a social media sample may not even reflect the composition of the user base of that same social media platform [8]. Further, there are differences in the styles, language, ways that people engage, and even how individuals represent themselves in real-life compared to social media [37]. Diversity analysis can help to identify these differences and also quantify the bias represented in the sample along diversity related characteristics and themes.

Computer mediated communication styles on social media are a unique set of linguist patterns that users have evolved to adapt to the social trends and norms of an online community using a social media application. Often these adaptations arise based on the availability of the platform's features as well as to work within limitations such as the amount or type of content that can be posted (e.g. [7, 21]). For example, stickers, website url shorteners, and emojis are popular within social media and text messaging applications [17]. Thus, the digital linguistic styles and patterns of users should also be considered during social media analysis. In this research, we focus on the use of emojis and keywords as part of our social media analysis.

There are three common approaches to conducting social media analysis with content containing emojis. Many studies will remove or sometimes replace emojis with words, (e.g. [45]). Another popular approach is content analysis that assigns a score based on the presence of specific emojis as indicators of emotion, sentiment, or sarcasm [16]. However, these approaches typically only analyze the few emojis which are anthropomorphic, such as face- and body-gestures and do not consider the thousands of other emojis that may exist, such as emojis depicting symbols, animals, food, and objects. The other approach is semantic analysis to assign meaning to emojis [5, 41]. Although the Unicode Consortium provides a standard for emoji codepoints and names, the definition and digital character representation of emojis are only suggestions [43]. As a result, emoji presentation, specifically the color, shape, or details of an emoji, may vary across devices and even social media platforms. These differences can impact how emojis are used and perceived [17, 21, 41]. The interpretation of emojis within social media content will also be impacted by the socio-cultural context of the users [5], nearby text [14, 27], or even the type of document such as a username, tweet, or user profile [40].

We chose to examine the use of emojis as cues for diversity, inspired by the handful of studies which show differences in the most common emojis used per various user populations. For example, trends in emoji use were described based on culture and geography at the country level (e.g. [26]). Others have described emoji use based on age [30] or differences based on gender [9, 21]. And since 2016, with the availability of skin-toned emojis, some researchers have shown that people typically use emojis with a similar appearance as their own, such as skin-tone [4, 32]. However, these studies all describe the aggregated emoji use across populations when the diversity attributes are already known. Nonetheless, the current state of emoji related research demonstrates the utility of including emojis in social media analysis and provides a useful starting point for examining diversity of users based on keywords and emoji use.

3 DATA COLLECTION

For our research, we conducted social media analysis on publicly available tweets and user profiles collected from Twitter. We describe the datasets in this section.

3.1 Tweets

We collected over 44 million publicly available tweets during the timeframe of October 1, 2018 to November 7, 2018, to coincide with the timing of the one month prior to and including the day of the 2018 U.S. midterm elections. The tweets were collected using Twitter's free streaming application programming interface (API), which provides only a sampling of all available tweets. The keywords and account names we used to collect tweets related to the 2018 U.S. midterm elections, campaign slogans, and specific user accounts associated with the Democratic and Republican political parties and candidates, Table 1.

Table 1: Subset of keywords for 2018 U.S. midterm election

Election related	Campaign slogans	Accounts
election2018	BlueWave	@TedCruz
midterms2018	FlipTheSenate	@BetoORourke
democrat	FlipTheHouse	@FLGovScott
DNC	VoteThemOut	@SenateDems
republican	MAGA	@HouseDemocrats
RNC	KAG	@SenateGOP
GOP	TakeItBack	@HouseGOP

Twitter returns the tweet content and metadata about the tweet. One attribute in the metadata indicates whether the tweet is a retweet. For our analysis we divided the tweet collection into retweets (RT) and non-retweets (Non-RT). Retweets accounted for 84% of all the tweets we collected. Of the over 3 million unique authors in our collection, 61% only sent retweets, 19% never sent retweets, and 20% sent both retweets and non-retweets.

3.2 User profiles

For the 3 million authors of tweets in our collection, we used the free Twitter search API to collect the user profiles. The user profile information returned from Twitter includes an attribute for the user profile description, a free-form narrative that users can fill in if they wish. Most of the user profiles descriptions we collected contained information and only a handful were blank. While the amount of detail and length varied greatly, the type of information in the user profiles included hobbies, interests, political beliefs, religious beliefs, race, language, national or cultural identity, relationship status, educational status, gender, sexual orientation, employment status, history, and more. Although it is difficult to validate the authenticity of the information contained in user profile descriptions, this volunteered narrative provides a wealth of user information. Next we describe how we assign diversity categories and subcategories based on diversity-related keywords and emojis contained in user profiles and tweets.

4 METHODOLOGY

In this section we present our methodology for conducting diversity analysis of social media. We discuss the creation of a diversity language model to associate emojis and keywords with diversity-related characteristics. The model is used to label social media content with diversity categories and subcategories. The diversity-labeled content is then analyzed to identify themes and patterns of behavior across users with similar diversity characteristics. Figure 1 provides a graphical summary of our diversity analysis workflow.

4.1 Diversity language model

We developed a diversity language model based on specific keywords and emojis that may be associated with the following five diversity categories: gender, religion, race/skin tone, sexual orientation, and political ideology. Categories were divided into subcategories, e.g. the gender category contains the following subcategories: female, male, and transgender. Figure 2 lists the diversity categories and the main subcategories of our diversity language model. It also includes the emojis and a few of the Spanish and English keywords associated with a particular category and subcategory.

While the diversity language model displayed is not the exhaustive list of keywords we used, it is representative of how the final model was constructed. We acknowledge these are not fully representative lists and may not be endorsed by everyone; however, the keywords and emojis were used by more than 100 users in our collection in a way as to indicate diversity characteristics. We developed the model through an iterative process and refined the keywords and emojis included with the goal of having a robust model that describes the diversity characteristics of many users while also minimizing the number of terms that could lead to misclassifications. We also acknowledge that the presence of any of these terms or emojis from the diversity model that appear within a tweet or user profile does not always mean the user is self-identifying, and that the use of the term could be about someone else, discussion of a diversity topic, or in some cases may not be related to diversity at all.

The subcategories we used in our model in some cases were limited based on the emojis we selected as part of the diversity language model. For the skin tone category, we chose five subcategories based on the five Fitzpatrick skin tone emoji modifiers. We chose these emoji modifiers for use as diversity cues because a user

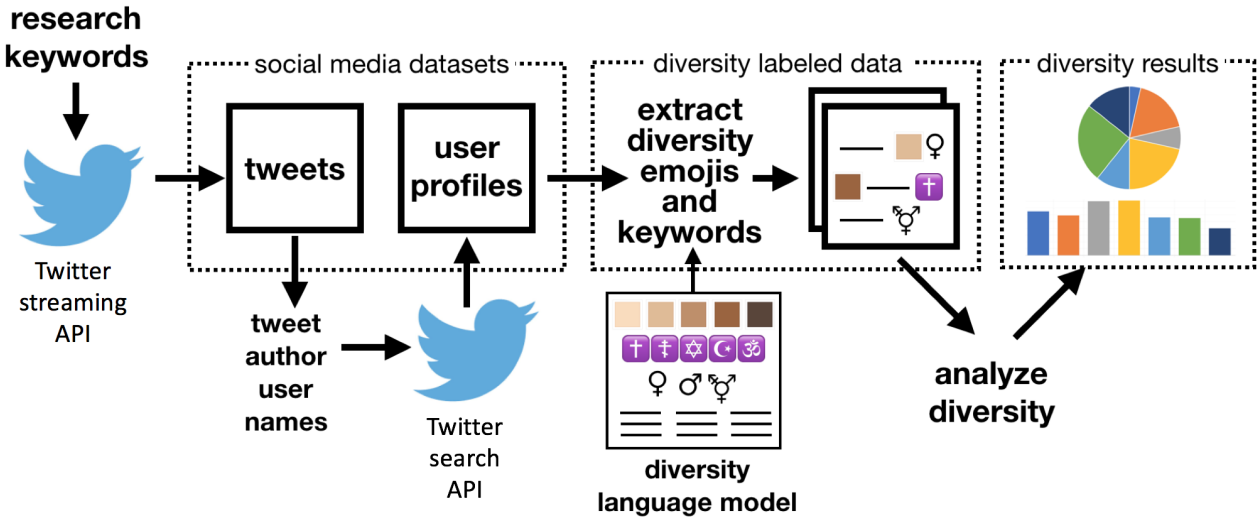

Figure 1: Workflow for diversity analysis of social media content.

must specifically select an emoji of a person or body-part that contains a skin tone because most default emoji presentations do not have skin tone modifiers. Users typically prefer to use a skin tone emoji with a similar appearance as their own [4, 32]. For gender, we chose the subcategories of female, male, and transgender based on the availability of corresponding emoji symbols and modifiers. For religion, we created subcategories for Christian, Jewish, Orthodox, Islam, and Hindu, based on emoji symbol availability. However, we also added the category for Atheist due to the large number of profiles using this term despite the lack of an emoji. For sexual orientation, we decided to keep this as a general category because of the lack of emoji differentiation for subcategories. And finally, we added political ideology as the fifth diversity category and chose only keywords based on their presence and context of use within user profiles in our collection.

As we tuned the diversity language model based on analysis of keywords and emojis in user profiles more heavily than tweets, there were a few cases where diversity associated terms used in the profile to indicate a diversity attribute occasionally had different meanings in tweets. For example, the keyword "white" for the skin tone subcategory for light worked well for labeling user profiles, but for tweets resulted in content containing a phrase such as "white house" to be wrongly labeled as a diversity attribute for skin tone. We mitigated this by only using the skin tone emojis during our analysis of tweets for the skin tone diversity category.

4.2 Assign diversity label

Once the diversity language model, Figure 2, is created and verified, we then use the model to assign diversity category and subcategory labels to tweets and user profiles that contained one of the terms or emojis in the model. In some cases, a single user profile or tweet may be assigned multiple subcategories within the same category. When this happened, which was rare, we would associate the content with the diversity subcategory as "mixed".

Category	Subcategories	Emojis	Keywords examples (English/Spanish)
Skintone	Light Light-medium Medium Medium-dark Dark		white, blanco moreno black, negro, negra
Gender	Female Male Transgender	♀ ♂ ⚧	woman, she/her, man, he/him transgender
Sexual Orientation	Indicated		gay, lesbian, bisexual, straight, hetero, polyamor
Religion	Christian Jewish Islamic Hindu Atheist		Catholic, Catolic, Jewish, Judaism Muslim, Islam Hindu, Sikh atheist, ateo
Political	Conservative Liberal		Republican Democrat

Figure 2: Diversity Language Model.

4.3 Diversity analysis

Using the social media content labeled with diversity categories and subcategories, we then review the composition of the dataset collected and identify trends based on the diversity characteristics. As part of the initial exploratory analysis we measure the prevalence of diversity characteristics as the percent of user's profiles and tweets containing specific diversity emojis and keywords. The composition of the datasets is then summarized as the proportion or percent of user accounts, retweets, and non-retweets for each of the diversity categories and subcategories. To understand differences in the way diversity emojis and terms are used, we also compared if a user includes the same diversity emojis and keywords in both their profile and tweets.

After reviewing the overall summary metrics of diversity for the collections, we analyze specific trends per set of users or tweets with the same diversity characteristics. We also examine temporal trends in tweet and author volume for each diversity characteristic. With this baseline understanding of the datasets along the diversity measures, we then focus on the patterns from combining multiple categories, e.g. political ideology and religion.

5 RESULTS AND DISCUSSION

As the aim of this paper is to explore diversity characteristics in a social media sample based on use of keywords and emojis in tweets and user profiles, building upon our methodology, we now turn to the results of our diversity analysis of social media content related to the 2018 U.S. midterm elections and discuss our findings. In what follows, we first gauge how prevalent the use of diversity emojis and keywords are within tweets and user profiles. We then establish a baseline by analyzing the patterns and trends within the dataset collections associated with diversity characteristics. Finally, we summarize the diversity of the user base in relation to election-related themes and campaign slogans.

5.1 Diversity in tweets and profiles

To understand how prevalent diversity characteristics are in our datasets, we measured what percent of tweets and user profiles contained diversity emojis or keywords. We also compared if these proportions were different between retweets and non-retweets. We present a summary of the results and key findings below.

Presence of diversity-related emojis and keywords in tweets. Across the corpus of 44 million tweets, approximately 15% (6.6 million tweets) contained either a diversity keyword or emoji, which we refer to as diversity tweets. Diversity tweets were sent by 36% of the authors in our collection. Of diversity tweets, 95% contained at least one of the diversity keywords and 5% contained at least one diversity-associated emoji, regardless of retweet or non-retweet status. We then examined the value of including diversity emojis as part of the language model by comparing the overlap of tweets with diversity emojis and diversity keywords. Of diversity tweets containing diversity emojis, only 15% also had diversity keywords. In terms of emoji use in general, only 9% of all tweets containing emojis had a diversity-related emoji for skin-tone, gender, or religion.

Presence of diversity-related emojis or keywords in user profiles. Next we examined the prevalence of diversity characteristics within user profiles. In the full set of 3.3 million user profiles in our collection, 15% of them (approximately 500,000) contained either a diversity keyword or diversity emoji, which we refer to as the diversity profiles. For authors sending both retweets and non-retweets, 22% had diversity characteristics in their profile. Authors sending only retweets had 18% and authors of only non-retweets had 12% of profiles containing diversity keywords or diversity emojis. For comparison, 19% of authors of diversity tweets had a diversity profile. In regards to the use of both diversity keywords and diversity emojis in the same profile, this was much more likely with authors that sent both retweets and non-retweets (44%) compared to authors of only retweets (27%) and authors of only non-retweets (25%).

Key findings for presence of diversity characteristics in tweets and profiles. While only 15% of tweets and user profiles in our collection contained diversity characteristics, over a third of the authors sent at least one tweet containing a diversity emoji or keyword. Users that included diversity characteristics in their profile were often as a way to self-identify diversity characteristics. Authors with diversity profiles were not any more likely to use diversity characteristics in their tweets, and vice versa. This indicates that both user profiles and tweets should be considered when taking diversity into account as they each may reveal different insights about the diversity of the social media sample. The use of diversity emojis and keywords within the same document occurred for a small percentage of tweets and user profiles. This indicates that emojis provide diversity cues not fully covered by keywords alone, thus, both are valuable to include in the diversity language model.

5.2 Analysis of diversity

Next we established a baseline of patterns and trends of diversity characteristics in our collection by analyzing tweets and profiles along each diversity category and subcategory. We summarize noteworthy results below.

Proportion of profiles and tweets by diversity category. We compared the proportion of diversity keywords and diversity emojis used in profiles and tweets across the diversity categories, Figure 3 and Figure 4. This analysis reveals that user profiles were slightly more likely than tweets to contain diversity cues associated with skin tone, gender, and sexual orientation. However, tweets were more likely to contain cues associated with religion. The use of skin tone emojis was more prevalent in user profiles than tweets.

Diversity composition of user profiles per diversity category. We then analyzed the diversity composition of the collection of user profiles by measuring the percent of profiles labeled with

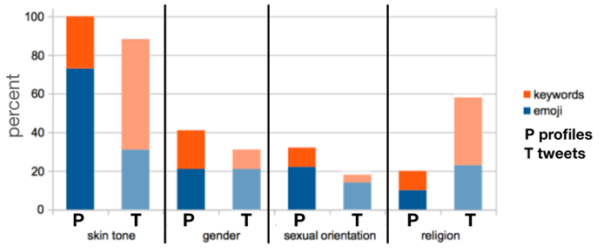

Figure 3: Proportion of diversity profiles and tweets with diversity emojis and keywords.

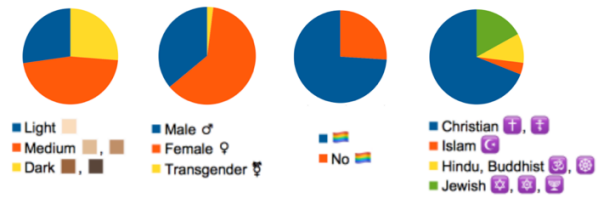

Figure 4: Proportion of user profiles by diversity subcategory.

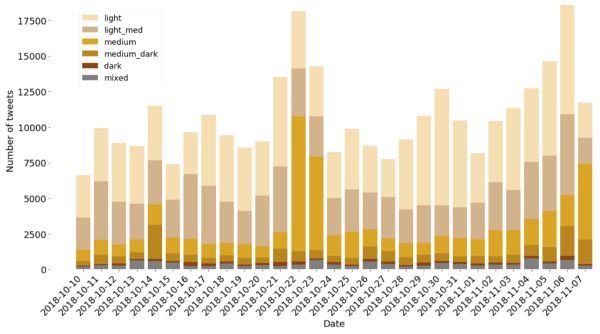

Figure 5: Volume of tweets with skin tone emoji by date.

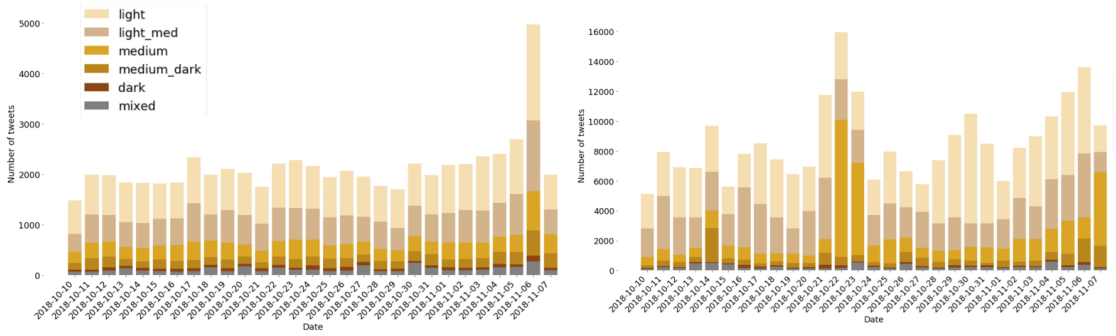

Figure 6: Volume of Tweets with skin tone emoji by date for non-retweets (a) and retweets (b).

a subcategory for each diversity category, Figure 4. Although this analysis only focuses on users with diversity profiles, it does provide new insight about the diversity characteristics represented within our collection of user profile descriptions. Of these diversity profiles in our collection, there was a greater percent with characteristics associated with: medium or medium-light skin tone emojis, female gender, and Christian religion. In addition, profiles that contained the rainbow flag also predominantly included keywords for sexual orientation,

Temporal analysis of diversity tweets. Analyzing the volume of tweets by diversity category per day provides a baseline to observe changes over time and identify peaks and valleys of activity. In Figure 5 we show the number of tweets in our collection per day that contained diversity emojis for the skin tone category. This plot reveals two spikes of activity one in mid-October and the other the day before the elections. Further analysis plotting the volume of non-retweets and retweets separately reveals more insight into the activity, Figure 6. The spike the day before the elections came from several non-retweets encouraging voting. The spike mid-October was retweet activity of a single tweet by a celebrity that included a political message, included an emoji with medium skin tone, and asked to be retweeted.

Content analysis of tweets based on diversity in author profiles. We analyzed tweet content to identify themes and topics expressed by authors whose profiles were labeled with a diversity subcategory. We processed tweet text using natural language processing methods to remove common words (e.g. the, and, as) and sentence punctuation. The remaining words were stemmed to identify topics. For example, stemming of voting, voter, votes, and vote would all reduce to the term vote. For each tweet, we kept the unique terms and hashtags. Next we identified the terms and hashtags used in retweets and non-retweets per diversity subcategory. Our results did not yield distinct differences based on diversity due to the large amount of terms and keywords we used to collect the data. However, we identified that two hashtags in particular "#maga" and "#bluewave" were among the top used hashtags in tweets across users of all diversity subcategories. These hashtags were associated with political campaigns for the Republican and Democratic political parties and we analyze the diversity representation of these users next.

Key findings from analysis of diversity characteristics. We used diversity analysis to establish a baseline of the diversity compositions of the collection of tweets and user profiles. With this baseline, we identified content that was trending or going viral associated with a specific diversity subcategory that otherwise would not have been as easily found in the volume of tweets in our collection. We also identified hashtags within our collection common across diversity categories. This insight into the composition and behavior of engagement is especially useful to understand the approaches and dynamics of online political campaigns.

5.3 Diversity and political ideology

Most tweets in our collection contained terms related to political ideology, however their use often did not reflect political party

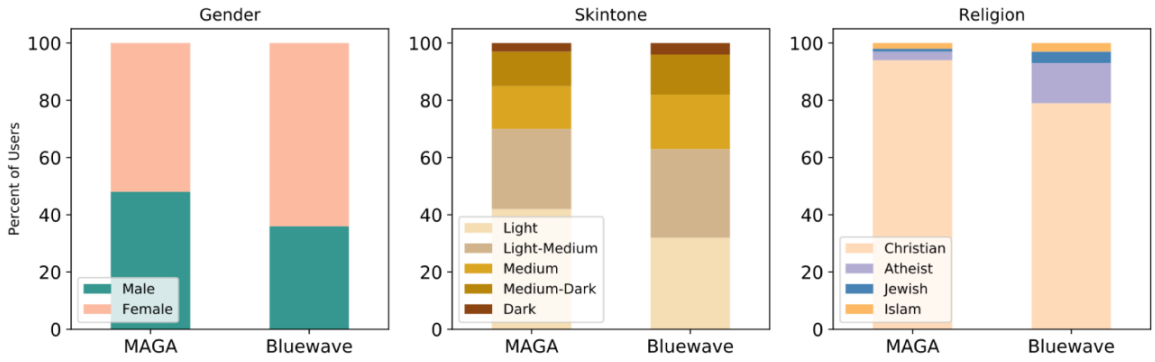

Figure 7: Composition of users in our collection for two political campaigns.

affiliation of the user and it was common to see keywords for both political parties included in the same tweet. For user profiles, only 5 percent contained political ideology keywords. From earlier content analysis we did find that tweet authors predominantly used campaign slogans across their tweets for only one of the political parties and similarly within user profiles. So rather than using political ideology, we analyzed the diversity represented in user profiles for authors of tweets that contained election related phrases and political party campaign slogans associated with the U.S. 2018 midterm elections. One of the top political campaigns for the Democratic party was the "Bluewave", which included terms such as "Blue Wave", "#bluewave", and the wave emoji. There were 22,051 users in our collection that had authored tweets including terms associated with Bluewave and also had diversity keywords or emoji in their user profile description. For the Republican party, the "MAGA" slogan, which stands for "Make America Great Again", and the hashtag "#maga" were used by 65,695 users who had diversity profiles. We compared the diversity composition of users based on their use of these political campaign slogans. We found users of Bluewave campaign phrases had a greater proportion of user profiles attributed with female gender; skin tones for medium-light, medium, and medium-dark; and atheist and Jewish for religion. In contrast, MAGA users had a higher percent of diversity profiles containing keywords and emojis representing the diversity subcategories of male, light skin tone, and Christian. Figure 7 shows the proportion of users in our collection per diversity category associated with MAGA and Blue wave.

Key findings from diversity analysis of political ideology and campaigns. While we were not able to easily divide users based on political ideology, we were able to differentiate users and content based on use of political campaign slogans in tweets and profiles. We found users did not typically self-identify political party affiliation in their user profile or in tweets. The use of political terms such as party names in tweets were more often used in banter or occasionally in discussion of social issues and policies associated with a political party. Alternatively, we were able to conduct diversity analysis along use of political campaign slogans. Campaign slogans were used across the diversity categories and were polarizing in that users typically only used campaign slogans for one political party. In addition, the diversity analysis of political campaigns yielded proportions across diversity categories similar to those reported in exit polls on election day for the 2018 U.S. midterm elections [42]. This indicates that the diversity analysis of social media users expressing support for a political party online through the use of campaign slogans may indicate diversity composition associated with voting outcomes. Further, diversity analysis provided additional insight about the content and user base represented in our social media sample and is a valuable addition to social media analysis.

6 CONCLUSIONS

Diversity analysis of social media data, as presented in this paper, provides an additional lens for studying diversity related themes in political discussion and the diversity of users engaging in online social communities. This paper presents a novel methodology for labeling and analyzing diversity represented in a social media sample based on keywords and emojis associated with gender, skin tone, sexual orientation, religion, and political ideology. In applying this methodology on a social media dataset collected during the leadup to the 2018 U.S. midterm elections, we established a baseline and identified trends of diversity related content that would not have been found in the volume of tweets otherwise. Furthermore, the diversity composition of users of political party campaign slogans yielded proportions along political party lines similar as those measured in exit polls on election day.

Our results indicate that both social media content and user profiles reveal different insights related to diversity and both should be considered when conducting a diversity analysis. Specifically, we observed that users were more likely to self-identify using diversity keywords and emojis in the user profile description rather than in their tweets. This means that for deriving a diversity composition of users, analysis of user profiles is preferred over aggregating social media posts by user. We also found semantic differences in the way diversity related emojis and keywords were used, which indicates that emojis are a useful addition to the analysis of diversity in social media.

This research is not without its challenges. The diversity keywords and emojis do not represent all diversity attributes and at

times may also take on additional meanings not related to diversity. Further, the number of social media posts and user profiles containing diversity related keywords and emojis will vary based on how the data are collected and social norms of the social media platform. In addition, when conducting social media analysis, there is always inherent bias when comparing the number and composition of users represented in a social media sample to a real-life population. This is further impacted by the difficulty in validating the authenticity of user accounts and veracity of content, especially with the prevalence of fake accounts, bots, trolls, and misinformation on social media platforms.

While this research focused on baseline and trends of diversity representation in social media, there are several opportunities for future work. Our approach for labeling diversity attributes in social media data can be compared to other methods using machine learning or manual tagging. Detailed content analysis, such as topic modeling or sentiment analysis, could be used to connect diversity with issues such as hate speech. Diversity analysis can be used to identify the representation of accounts associated with bots, misinformation efforts, and political influence campaigns which may reveal insights about their intent and targeted audience. Social network analysis can be used to examine diversity networks of users with similar diversity characteristics to measure how connected diverse groups are on social media related to political issues. In addition, the diversity language model presented in this paper can be further adapted for other languages or topics. It can be applied beyond just tweets to assess the extent to which diversity has been discussed as well as identifying diversity represented in online community engagement associated with other social and political topics.

REFERENCES

[1] Adam Badaway, Kristina Lerman, and Emilio Ferrara. 2019. Who falls for online political manipulation? In Proceedings of the 2019 World Wide Web Conference, May 13-17, 2019, San Francisco, CA, 162-168.

[2] Pablo Barberá. 2016. Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. Working Paper, New York University.

[3] Pablo Barberá and Thomas Zeitzoff. 2018. The new public address system: Why do world leaders adopt social media? International Studies Quarterly, 62, 1, 121-130.

[4] Francesco Barbieri and Jose Camacho-Collados. 2018. How gender and skin tone modifiers affect emoji semantics in Twitter. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (SEM18), June 5-6, 2018, New Orleans, Louisiana, 101-106.

[5] Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In Proceedings of the 2016 ACM on Multimedia Conference (MM'16), October 15-19, 2016, Amsterdam, The Netherlands, 531-535.

[6] Grant Blank and Christoph Lutz. 2017. Representativeness of social media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. American Behavioral Scientist, 61, 741-756.

[7] Leticia Bode and Kajsa E. Dalrymple. 2016. Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter. Journal of Political Marketing, 15, 4, 311-332.

[8] Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P. Gummadi. 2017. Who makes trends? Understanding demographic biases in crowdsourced recommendations. In Eleventh International Conference on Web and Social Media (ICWSM-17), May 16-18, 2017, Montréal, Québec, Canada, 22-31.

[9] Zhenpeng Chen, Xuan Lu, Wei Ai, Huoran Li, Qiaozhu Mei, and Xuanhe Liu. 2018. Through a gender lens: Learning usage patterns of emojis from large-scale Android users. In Proceedings of the 2018 World Wide Web Conference (WWW'18), April 23-27, 2018, Lyon, France, 763-772.

[10] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. 2013. #Earthquake: Twitter as a distributed sensor system. Transactions in GIS, 17, 1, 124-147.

[11] Elizabeth Culliford. 2019. Twitter makes global changes to comply with privacy laws. Reuters, December 2, 2019. https://www.reuters.com/article/us-twitter-privacy/twitter-makes-global-changes-to-comply-with-privacy-laws-idUSKBN1Y622J.

[12] Ashok Deb, Luca Luceri, Adam Badaway, and Emilio Ferrara. 2019. Perils and challenges of social media and election manipulation analysis: The 2018 US midterms. In Companion Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13-17, 2019, San Francisco, CA, 237-247.

[13] Carroll Doherty, Jocelyn Kiley, and Olivia O'Hea. 2018. Wide gender gap, growing educational divide in voters' party identification. Pew Research Center.

[14] Giulia Donato and Patrizia Paggio. 2017. Investigating redundancy in emoji use: Study on a Twitter based corpus. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2017), September 8, 2017, Copenhagen, Denmark, 118-126.

[15] Maeve Duggan and Joanna Brenner. 2013. The demographics of social media users, 2012. Pew Research Center.

[16] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion, and sarcasm. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), September 7-11, 2017, Copenhagen, Denmark, 1615-1625.

[17] Lauren Gawne, and Gretchen McCulloch. 2019. Emoji as digital gestures. Language@ Internet, 17, 2. http://nbn-resolving.de/urn:nbn:de:0009-7-48882

[18] Salvatore Giorgi, Veronica Lynn, Sandra Matz, Lyle Ungar, Hansen Andrew Schwartz. 2019. Correcting sociodemographic selection biases for accurate population prediction from social media. rXiv preprint arXiv:1911.03855.

[19] Sandra Halperin and Oliver Heath. 2020. Political research: Methods and practical skills. Oxford University Press, USA, 2020.

[20] Eszter Hargittai. 2018. Potential biases in big data: Omitted voices on social media. Social Science Computer Review, 38, 1, 10-24.

[21] Susan C. Herring and Ashley R. Dainas. 2018. Receiver interpretations of emoji functions: A gender perspective. In1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji 2018), June 25, 2018, Stanford, CA, USA.

[22] Jaigris Hodson and Brigitte Petersen. 2019. Diversity in Canadian election-related Twitter discourses: Influential voices and the media logic of #elxn42 and #cdnpoli hashtags. Journal of Information Technology & Politics, 16, 3, 307-323.

[23] Adam Hughes and Nida Asheer. 2019. National politics on Twitter: Small share of U.S. adults produce majority of tweets. Pew Research Center.

[24] Kyriaki Kalimeri, Mariano G. Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019. Predicting demographics, moral foundations, and human values from digital behaviours. Computers in Human Behavior, 92, 428-445.

[25] Dijana Kosmajac and Vlado Keselj. 2019. Twitter bot detection using diversity measures. In Proceedings of the 3rd International Conference on Natural Language and Speech Processing, September 12-13, 2019, Trento, Italy, 1-8.

[26] Nikola Ljubešić and Darja Fišer. 2016. A global analysis of emoji usage. In Proceedings of the 10th Web as Corpus Workshop (WAC-X 2016), August 12, 2016, Berlin, Germany, 82-89.

[27] Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In Eleventh International Conference on Web and Social Media (ICWSM-17), May 16-18, 2017, Montréal, Québec, Canada, 152-161.

[28] Alan Miller, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of Twitter users. In Fifth International Conference on Web and Social Media (ICWSM-11), July 17-21, 2011, Barcelona, Spain, 554-557.

[29] Ahmed Mourad, Falk Scholer, Walid Magdy, and Mark Sanderson. 2019. A practical guide for the effective evaluation of Twitter user geolocation. ACM Transactions on Social Computing, 2, 3, 1-23.

[30] Nao Na'aman, Hannah Provenza, and Orion Montoya. 2017. Varying linguistic purposes of emoji in (Twitter) context. In ACL 2017, Student Research Workshop, July 30-August 4, Vancouver, Canada, 136-141.

[31] Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In Proceedings of the 27th International Conference on Computational Linguistics (COLING-18), August 20-26, 2018, Santa Fe, New Mexico, USA, 1534-1545.

[32] Alexander Robertson, Walid Magdy, and Sharon Goldwater. 2018. Self-representation on Twitter using emoji skin color modifiers. In 12th International Conference on Web and Social Media (ICWSM-18), June 25-28, 2018, Stanford, CA, USA, 680-683.

[33] Jonathon Schuldt, and Adam Pearson. 2016. The role of race and ethnicity in climate change polarization: evidence from a U.S. national survey experiment. Climatic Change, 136, 495–505.

[34] Elisa Shearer and Jeffrey Gottfried. 2017. News use across social media platforms 2017. Pew Research Center.

[35] Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. 2013. Knowing the tweeters: Deriving sociologically

relevant demographics from Twitter. Sociological Research Online, 18, 3, 1-11.
[36] Felipe Bonow Soares, Raquel Recuero, and Gabriela Zago. 2019. Asymmetric polarization on Twitter and the 2018 Brazilian presidential elections. In Proceedings of the 10th International Conference on Social Media and Society, July 19-21, 2019, Toronto, Ontario, Canada, 67-76.
[37] Joanna Sterling, John T. Jost, and Richard Bonneau. 2020. Political psycholinguistics: A comprehensive analysis of the language habits of liberal and conservative social media users. Journal of Personality and Social Psychology, 18, 4, 805-834.
[38] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. Social media analytics–Challenges in topic discovery, data collection, and data preparation. International Journal of Information Management, 39, 156-168.
[39] Sebastien Stier, Arnim Bleier, Haiko Lietz, and Markus Strohmaier. 2018. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. Political Communication, 35, 1, 50-74.
[40] Melanie Swartz and Andrew Crooks. 2020. Comparison of emoji use in names, profiles, and tweets. In IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 2020, 375-380.
[41] Garreth W. Tigwell and David R. Flatla. 2016. Oh that's what you meant!: Reducing emoji misunderstanding. In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI'16), September 6-9, 2016, Florence, Italy, 859-866.
[42] Alec Tyson. 2018. The 2018 midterm vote: Divisions by race, gender, and education. Pew Research Center.
[43] Unicode Emoji. 2018. In Unicode Technical Standard #51. Unicode.org.
[44] Maurice Vergeer, Liesbeth Hermans, and Steven Sams. 2013. Online social networks and micro-blogging in political campaigning: The exploration of a new campaign tool and a new campaign style. Party Politics, 19, 3, 477-501.
[45] Xiaoyi Yuan and Andrew T. Crooks. 2018. Examining online vaccination discussion and communities in Twitter. In Proceedings of the 9th International Conference on Social Media and Society, July 18-20, 2018, Copenhagen, Denmark, 197-206.