Examining Protest as An Intervention to Reduce Online Prejudice: A Case Study of Prejudice Against Immigrants

Kai Wei
University of Pittsburgh
Pittsburgh, PA, USA
kaiwe@amazon.com

Yu-Ru Lin*
University of Pittsburgh
Pittsburgh, PA, USA
yurulin@pitt.edu

Muheng Yan
University of Pittsburgh
Pittsburgh, PA, USA
muheng.yan@pitt.edu

Abstract

There has been a growing concern about online users using social media to incite prejudice and hatred against other individuals or groups. While there has been research in developing automated techniques to identify online prejudice acts and hate speech, how to effectively counter online prejudice remains a societal challenge. Social protests, on the other hand, have been frequently used as an intervention for countering prejudice. However, research to date has not examined the relationship between protests and online prejudice. Using large-scale panel data collected from Twitter, we examine the changes in users' tweeting behaviors relating to prejudice against immigrants following recent protests in the U.S. on immigration related topics. This is the first empirical study examining the effect of protests on reducing online prejudice. Our results show that there were both negative and positive changes in the measured prejudice after a protest, suggesting protest might have a mixed effect on reducing prejudice. We further identify users who are likely to change (or resist change) after a protest. This work contributes to the understanding of online prejudice and its intervention effect. The findings of this research have implications for designing targeted intervention.

CCS Concepts

• **Applied computing → Law, social and behavioral sciences**;
• **Information systems → Web mining**.

Keywords

computational social science, social movement, online hate and prejudice, civic protest, immigration

ACM Reference Format:
Kai Wei, Yu-Ru Lin, and Muheng Yan. 2020. Examining Protest as An Intervention to Reduce Online Prejudice: A Case Study of Prejudice Against Immigrants. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380307

1 Introduction

There has been a growing concern about the surge of online hate groups and their influence ranging from shaping social values and perceptions [10, 16, 39] to spreading racist beliefs and to incite violence offline [7, 15]. Expressions of such prejudice and even hatred, as well as their prevalence and influence, have become a serious societal issue. This is of particular concern for youths and young adults because they are not only active social media adapters [26] but more likely to be affected and influenced by hatred and extremist ideas propagated through the Web [16]. Recent research also found that people from minority ethnic, gender, and sexual minority groups have continued to be the primary targets for various forms of prejudice in social media [47], with immigrants being one of the groups particularly susceptible to the real, adverse effects of online prejudice [11, 17].

Prejudice refers to individuals' antipathy towards a person or a group. People may convey their prejudice through private or public speech. This work focuses on prejudice conveyed in the public online space. We define prejudiced speech as antipathetic (a deep-seated dislike) remarks against a person, group, or community. It should be noted that prejudiced speech and hate are related but separate concepts. Hate speech expresses hatred; prejudiced speech expresses feelings of a strong dislike, opposition, or anger, which are not necessarily hate.

While there has been growing interest in developing automated techniques to identify online prejudice acts and hate speech, how to effectively counter online prejudice remains a societal challenge. Previous research on prejudice has proposed that protest can suppress prejudice against an out-group [18]. One notable example is that civil rights movement drastically reduced prejudice against blacks in the U.S. [37]. However, research to date has not specifically examined the effects of social protests in online prejudice. Therefore, this research takes the first step towards understanding the effect of protest on online prejudice, by focusing on the changes in users' tweeting behaviors relating to prejudice against immigrants following recent protests in the U.S. on immigration related topics. Our major contributions are as follows:

- We present the first empirical study on the effect of using protests as an intervention to reduce online prejudice. We empirically show that there were both negative and positive changes in the measured prejudice after a protest, suggesting protest might have a mixed effect on reducing prejudice.
- We propose a study design that includes building a prejudice classifier to measure the effect of a social protest on reducing online prejudice. We also built prediction model to identify users who are likely to have a change (or resist change) in the measured prejudice following protests.
- We further identify themes to contextualize the change of measured prejudice. Specifically, we seek to understand in what way the identified prejudiced users change their prejudiced expression

*Corresponding author.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380307

following protests. Our results have implications for designing targeted interventions for reducing online prejudice.

2 Related Work

We review studies on online prejudice and the theories and preliminary evidence for the effects of social protests in online prejudice.

2.1 Online Prejudice

Prejudice stems from feelings of dislike, false assumptions, and stereotypes about a personal or a group [1]. The close line of research on studying online prejudice is hate speech detection. There are three approaches to online hate speech detection: 1) dictionary-based and sentence-syntax based approaches [31, 34, 49], 2) traditional machine learning approach [21], and 3) deep learning approach [6].

Early research has used dictionary-based and sentence-syntax based approaches. For example, [31] used racial slurs in Twitter posts to detect racist users, and found that racist tweets appeared in both political leaders' Twitter followers. While this method offers a simple way to identify hate speech, using racial slurs to determine people's prejudices against a group is unable to capture prejudiced statements that do not have racial slurs. As a result, this approach often has a low recall for retrieving prejudiced speech and can introduce racial bias in detecting hate [48]. In addition, some research also adopted syntactic approach to detect hate speech [49]. However, this approach also suffered a low recall, as the way people express prejudice or hate is not as explicit as defined above.

Traditional machine learning techniques were also applied to detecting hate speech detection. For example, [13] applied a supervised machine learning approach and studied cyber hate across multiple groups (race, sex, and disability) on Twitter and defined cyberhate as the "othering" language such as using "them" to refer an out-group. However, this work did not distinguish offensive language from hate; for example, using terms such as "jokes" and "really druck" to mock disabled athletes seems to be offensive but not necessarily hateful. As suggested by a recent research [21], hate speech detection should consider the difference between offensive language and hate speech as hate speech is used to target a social group with an intention to exclude that group.

Recent works have applied deep learning techniques to online hate speech detection. For example, [6] focused on the detection of hate speech against immigrants and women in Spanish and English messages extracted from Twitter. This work showed that Support Vector Machines (SVM) outperformed sophisticated systems such as Convolutional Neural Networks (CNNs) and Long Short Term Memory networks (LSTMs) in hate speech detection task. Consistent with this finding, [21] reported the best performing classifier for hate speech detection is SVM, with a F1 score of 0.90 for classifying offensive and non-offensive language. These works suggested that traditional machine learning techniques remain effective in related tasks.

While these previous works have extensively studied hate speech, few have focused on online prejudice. Overall, prejudiced speech has a broader scope than hate speech. While prejudice can be manifested in hate speech, not all manifestations of prejudice are hate speech. For example, someone can express prejudice against immigrants by saying "*immigrants are lazy and they steal our jobs away!*" This statement express this person's prejudice because it shows antipathy towards immigrants. However, it is not hate speech. Hate speech against immigrants (e.g., "*these fucking illegals are not here to mow your lawn - they're here to blow up your buildings and kill your children, and you, and me*"), on the other hand, would express much more intense and explicit feeling of dislike about immigrants.

This work takes the first step to identify online prejudice. Moreover, we examine using protest as an intervention to reduce prejudice. This is because while monitoring prejudice or hatred content through automated techniques is an important step in regulating online space [9], solely relying on monitoring is not sufficient to reduce the development and influence of such behaviors. Yet, to date, there have been no studies focusing on the effect of intervention strategies such as protest.

2.2 Protest as Intervention

Protest is a form of sociopolitical collective action in which members of a group act together to express objection to particular actions or situations [2]. It can take many forms, such as letter writing, public denunciations, marches, sit-ins, and boycotts directed toward prejudiced, offensive or stigmatizing practices [19]. Participants in a protest often believe that their actions can make the public more aware of certain critical issues [25], pressure the government to take actions in policy change [2], and shift the social values and norms [5]. Recent research suggests that protests could serve as critical counter-political voices [51] to resist prejudice and discrimination [53]. However, there has also been growing concerns about the effect of protest on prejudice. This section reviews the literature on protest and prejudice in order to highlight the need to further study the effect of protest on online prejudice.

Social movement theory poses that protest can have an impact on society, leading to changes in political and cultural outcomes [24]. Previous research has linked nonviolent protest to attitude change, and found that the changes in attitudes can persist [37]. For example, protest was found to have a positive effect on reducing negative stereotypes, and that the positive effect remained at one-week follow-up [11]. It was also found that after the 2006 immigration protests, foreign-born Latinos were reported more positive attitudes towards immigrants and support for benign immigration policy (e.g., immediate legalization of current unauthorized immigrants) [12]. White Southerners living in counties where a sit-in occurred were observed to be more likely to support the protest, compared with those counties with no sit-in event [3]. These previous studies highlighted the relationship between protests and broader attitudinal change, and supported the premises that protest could be used as an intervention to reduce online prejudice.

However, there has been research showing that protest has no effect and even a negative effect on prejudice. In a meta-analysis, Corrigan [18] examined publications between 1972 to 2010 that focused on the effects of the anti-stigma approaches on public stigma related to mental illness. Among 72 examined studies, only one tested the effectiveness of protest, which yielded non-significant findings for the effect of fact sheets from Psychiatrists' Changing Minds campaign on reducing prejudiced attitudes against schizophrenia and alcoholism [35]. Several studies even suggested that protests that attempted to suppress prejudice can produce an unintended "rebound" in which prejudices about a group remain unchanged or

actually become worse [36, 54, 55]. A recent experimental studied also showed that the unintended "rebound" effect is conditioned on social norms and participants levels of prejudices. Specifically, participants suppressed their prejudice against homosexual group when they expected to share their responses with others. Participants with higher prejudice are more likely to exhibit prejudice rebound and rated homosexual group more stereotypically [23].

These previous studies suggested that protest can lead to changes in attitudes. However, there has been mixed findings on the effect of protest on prejudice. Moreover, little is known about protest as an intervention strategy to reduce online prejudice.

3 Research Questions and Study Design

This section introduces study design, study context, and research questions.

3.1 Study Design

In this work, we adapt "computational focus groups" method [32] to study users' online prejudiced speech and how immigrant protest events are related to its changes. Computational focus groups is a framework for tracking changes in social media users' emotions, attitudes, or opinions about a group or an issue following specific events [32]. Specifically, it tracks users' behavioral outcomes by analyzing the content of social media users' posts. This framework is similar to traditional intervention studies in that it requires an intervention (a focal event) and a measurable outcome (users' behavioral outcomes). However, the major difference between these two is the methods used to obtain outcomes. This is mainly because online users express their emotions, attitudes, and opinions in the form of unstructured texts, which requires researchers to leverage text mining techniques to turn unstructured texts into numbers. Our study design includes the following steps.

Identify a focal event A focal event is an event that has potential impact on people's behavioral outcomes. For example, previous research has used computational focus groups in studying events such as terrorist attacks and presidential debates [32, 33]. In this work, we select two most recent immigrant protests as focal events: "Day Without Immigrants" protest and the "No Ban, No Wall" protest. These two events are selected because they are the most recent nationwide immigrant protest in the U.S.

Construct focus groups Focus groups, traditionally, are a form of group interview that capitalizes on communication between research participants in order to generate text data [30]. Social media users generate data by communicating their emotions, attitudes, or opinions about a group or an issue by posting short text messages. This online platforms provides wealth of data that would otherwise require thousands of group interviews. In this work, we constructed a user panel who showed interest in discussing the topics relevant to immigrants and divided them into sub-groups based on their exposure level to protest cities.

Track user's behavioral outcomes Users' behavioral outcomes are tracked by leveraging text mining techniques to quantify users' social media posts. In this work, we leverage text mining techniques to identify whether a tweet is prejudiced speech against immigrants or not; and whether a tweet is about immigrants or not. Then, we aggregate all the tweets at the user level.

Compare users' behavioral outcome(s) before and after an event In this framework, users' behavioral outcome(s) before a focal event is considered a baseline measure. The differences in the user's behavioral outcome(s) before and after the event are regarded as the changes related to the focal event. In this work, we compared users' online prejudice against immigrants before and after protest events.

3.2 Study Context

To understand the impact of protest on online prejudice, we focused on two protest events: the "Day Without Immigrants" and the "No Ban, No Wall" protests in the U.S. These protest events were the most recent nationwide protests that aimed to show the important contributions of immigration and to resist punitive immigration policies.

"Day Without Immigrants" protest As a response to President Donald Trump's plans to build a border wall, strip sanctuary cities of federal funding, and deport potentially millions of undocumented immigrants [8], this protest took place on February 16, 2017 in multiple cities across the US. It aimed to show the importance of immigrants to the US economy and in the day to day lives of American citizens. Social media and other means were used to disseminate the information about this protest [46]. On the protest day, shops and restaurants were closed in several major US cities. For example, more than 50 restaurants were closed in Washington DC [20], and over 1000 businesses were closed in Dallas [45], and thousands of children did not attend school [14, 46].

"No Ban, No Wall" protest This protest took place on January 28, 2017 as a response to President Donald Trump's plan to ban citizens of certain Muslim countries from entering the US, and suspend admission of all refugees entering the country [22]. It was also planned and disseminated via social media, and simultaneously executed in multiple cities in the US, including New York, Los Angeles, and Philadelphia [4]. On the protest day, in Seattle-Tacoma Airport alone, about 3,000 protesters gathered to protest Trump's plan to ban citizens of certain Muslim countries from entering the US [43]. Thousands of protesters also gathered in major airports in cities such as Portland [44], Los Angeles [44], and Philadelphia [28]. Compared with the tactics employed by "Day Without Immigrants," the tactics used in this protest were both more traditional and less disruptive. Differences in social media responses to these tactics may provide important implications for achieving desirable protest outcomes.

3.3 Research Questions

This work aims to examine protest as an intervention to reduce prejudice, with a focus on prejudice against immigrants. Speciecially, we seek to answer the following research questions:

RQ1 To what extent protest increase awareness of immigrants?
RQ2 To what extent protest reduce online prejudice?
RQ3 What kind of users were more likely to change (or resist to change) after a relevant protest?
RQ4 In what way users changed the prejudiced expression?

4 Data

This section describes our data collection procedure. Panel data collection were carried out in March 2018. Users of interest were selected from multiple data sources and filtered based on exclusion criteria. For each selected user, all available tweets during the study period were collected from their timeline and profile.

Data sources Multiple data sources were used for selecting users of interest. These data were geo-based and hashtag-based datasets. Geo-based datasets were used as initial datasets because one of the study aims was to examine the relationship between levels of exposure to a local protest and user's online prejudice against immigrants. However, solely relying geo-tagged datasets posed risks to sampling bias , with only about 1% of tweets have geo-tagged information [29]. To mitigate the risks, we also collected additional data that contains protest event-related hashtags. The following section describes details about geo-based and hashtag-based datasets.

The geo-tagged dataset was provided by a research collaborator. These geo-based users were included in the initial dataset because we were interested in the role of geo-exposure in user's online prejudice against immigrants. To achieve our goal, one of the critical task is to exclude users who have never discussed immigrants in their tweets because these users are unlikely to be the ones who explicitly express prejudiced speech against immigrants. To this end, we included users who were located in the US and showed interest in discussing topics that were relevant to immigrants. These topics were identified based on a set of keywords or keyword patterns related to immigrants (e.g., "latino," "mexican," "muslim," or "immigra*"). In total, there were 138,759 users included in this study from the geo-tagged dataset.

Admittedly, solely relying on geo-based data introduced sampling bias because not all Twitter users choose to disclose their geographic locations. Thus, we collected additional users who showed interest in discussing topics related to the protest events. These users were identified based their mentions of #DayWithoutImmigrant," "#NoBanNoWall," and "BuildtheWall". Twitter API was used to users who mentioned these hashtags. In total, we identified 22,108 users. Among these users, 4,034 unique users mentioning "#DayWithoutImmigrant"; 8,949 unique users mentioning "#NoBanNoWall", and 9,125 unique users mentioning "#BuildtheWall".

Exclusion criteria Exclusion criteria were used to remove duplicated users, social bots, and organizational users. Since users were identified from multiple source, we first removed users who appeared more than once in the data. After removing these users, 159,702 users remained in the data.

Social bots are accounts controlled by software that automatically generates contents [52]. Given the interest of this study is human users, and thus bots were excluded prior to data analysis. To remove the social bots from these users. we used the Botometer API [52], a system that has been shown effective in bot detection (with an AUC of 94%), to detect and remove bots. After excluding social bots, a total of 112,142 users remained.

Organizational user accounts were also removed from this study. A user is considered to have an organizational account if the account represents an institution, corporation, agency, news media, or common interest group [40]. To identify such accounts, we used the machine learning tool developed by [38]. This tool has yielded an 88% overall classification accuracy. After excluding organizational accounts, a total of 102, 094 users remained.

Panel data collection Panel data were collected for each included user. In Twitter, panel data is embedded in user's timeline profile, which displays the latest tweets from the specified (public) Twitter account. For each of the user, Twitter REST API was used to collect all available tweets during the study period (two weeks before, two weeks after for each protest event). In total, we collected a total of 31,210,740 tweets posted during the study period from all included users (n_{users}= 102, 094).

5 Detecting Prejudice

One of the major challenges for studying the impact of a protest on online user's prejudice is to develop reliably measurement for quantifying the level of prejudice. To address this challenge, we develop supervised learning techniques, which includes (1) immigrant-related tweet classification: to identify tweets relevant to the immigration topics, and (2) prejudiced tweet classification: to identify tweets expressing prejudice against immigrants from the immigrant-related tweets. To build these classification techniques, we first (i) scope tweets that are potentially relevant to the immigration topics (section 5.1), (ii) establish ground-truth labels for training classifiers (section 5.2), and (iii) train and evaluate the supervised learning models (section 5.3). We report the classification performance in section 5.4.

5.1 Scoping Potentially Relevant Tweets

In order to correctly identify prejudice against immigrants, we need to first identify tweets that are relevant to immigrants or immigration topics. To make this step more efficient, we leverage keyword-based methods to filter the set of tweets that are potentially relevant, and later apply immigrant-related classification to this potentially relevant set. We combined the widely-used keyword query strategy with a keyword expansion method. Appropriate keywords can help filter irrelevant tweets and the strategy is relatively cost-effective. However, identifying appropriate keywords is challenging – narrowly defined keywords might result in the problem of missing relevant information, whereas broadly defined keywords might result in too much noise. To address this challenge, we adopted the idea of keyword expansion [41], where a set of initial keywords was used to bootstrap words similar to these words.

In our work, "immigrant" or "immigrants" were used as initial keywords. To expand the keywords, we first trained a Word2Vec model using the Gensim package to construct the semantic vectors of words. We then computed the cosine similarity between the word vectors and iteratively retrieved the words having the highest similarity with the set of relevant keywords and added them into the keyword set. The iterative process ended when no words were identified at the top by manual inspection. In total, there were 889,579 potentially relevant tweets from 71,919 Twitter users. From these potentially relevant tweets, we further classified them as tweets about immigrants, and tweets prejudiced against immigrants.

5.2 Establishing Ground Truth

As discussed in section 2, existing online hate speech detection techniques are not suitable for detecting prejudiced speech as the latter has a broader scope. To develop suitable prejudiced speech classifiers we need to establish proper ground truth, which is constructed based on the following human coding process.

5.2.1 Human coding process In this study, a codebook was developed for human coders to classify whether a tweet was about immigrants, and whether it expresses prejudice. The codebook included a definition of immigrants and prejudice, and related examples and rationales for coding a tweet as about immigrants or prejudice. The indicators and examples were derived from coding a random sample of tweets. Rationales were brief descriptions of reasons for coding the tweets.

Coding Two independent coders were recruited to assist the coding process, both being native English speakers of college-level education who are active social media users and check social media posts every day. Two training sessions were conducted before coders were asked to code the tweets independently. In the first training session, we provided a brief overview of the study, discussing coding tasks and overall work flow. Following the training session, coders were asked to code a random sample of 200 immigrant-related tweets that had already been coded by the author based on the codebook; this batch was used to facilitate training. In the second training session, we discussed the coding results with the coders, and we each explained our reasons for the answer codes in the batch. Through the discussion, we found that misclassifications were due primarily to the misinterpretation of the tweets. For example, one coder misclassified this tweet, "FBI's pre-election sweep of Muslim Americans raises surveillance fears," to be about immigrants. The tweet is about Muslim Americans, not Muslim immigrants. After the discussion, we reviewed our coding. The final codes for this batch were based on majority rule, and this batch was then used as the gold standard for future coding. After training sessions, coders proceeded to code four batches of 200 randomly sampled tweets. Each batch was coded independently by the coders. The results of these batches were used to test codebook reliability.

Reliability Evaluation Both coders had substantial agreement in terms of Cohen's kappa coefficient ($\kappa > 0.61$) with the author on the gold standard batch before they proceeded to test codebook reliability, which consisted of four batches. Coders had substantial agreement in coding each batch, which indicates that the codebook achieved high reliability.

Following the codebook reliability testing, coders and the author coded an additional 2000 tweets. The average pairwise Cohen's kappa coefficient was 0.87 for coding tweets about immigrants, and 0.63 for coding tweets that exhibited prejudice against immigrants. The majority rule was used to decide the final code for each tweet. In total, there were 3000 labeled tweets. Table 2 shows the distribution of labeled tweets. A total of 1717 tweets (about 50%) were labeled as about immigrants. A total of 471 (about 16%) were prejudiced against immigrants. This coding process generated a random sample of 3000 labeled tweets from all relevant tweets during the study period, which was used for training and evaluating the machine classifiers.

Table 1: Distribution of labeled tweets

	Yes	No
About immigrants	1717	1283
Prejudice against immigrants	471	2529

5.3 Experiment Setup

Experiments were carried out to select machine learning models for classifying tweets about immigrants, and tweets with prejudice against immigrants. Both were binary classification tasks with the objective of classifying whether a tweet belonged to one category or the other.

Pre-processing Prior to training classification models, text preprocessing was performed on both labeled and unlabeled tweets to remove noise and prepare the text for classification. In this process, we removed stop words, URLs, and mentions (@username). The labeled tweets were split into 60% as training, 20% as test, and 20% as development, a common practice in machine learning.

Feature The mean vectors of the Word2Vec model were used as features to train the classification models. Specifically, each tweet consists of words. After training the Word2Vec model, each word is represented in a 300-dimensional vector. The mean vectorization of the embedding model for a given tweet is defined as taking the average of all the word vectors in the tweet.

Classifier In the experiments, we tested the following supervised machine learning models: Naive Bayes, Adaptive Boosting, Support Vector Machines, Logistics Regression, and Extreme Gradient Boosting. These models were chosen because they have been shown to perform well in classifying tweets [21, 27].

Imbalanced data and over-sampling As shown in Table 1, there was a major issue with imbalanced data where only about 16% were labeled prejudiced against immigrants. Previous research has shown that classification of data with imbalanced class distribution can suffer significant drawbacks in model performance because most standard classifier learning algorithms assume a relatively balanced class distribution and equal misclassification costs. This would lead classifiers to be more sensitive to detecting the majority class and less sensitive to the minority class [50]. To address the issue of imbalanced tweets that contained prejudiced speech against immigrants, we used the naive random over-sampling technique to generate tweets that were labeled as prejudiced speech (the minority class). This over-sampling technique generates new samples by randomly sampling the replacements of the current available samples. The over-sampling was only applied to the training dataset. The random over-sampling was implemented using the imbalanced-learn Python package.

5.4 Classification Performance

Accuracy of the models was determined based on *precision*, *recall*, *F1-score*, and *AUC*. The models that performed the best were used to classify the remaining data.

Table 2 shows the accuracy of supervised learning models for (1) immigrant-related tweet classification and (2) prejudiced tweet classification. In the first task, both AdaBoost and SVM had good precision (above 80%), recall (above 80%), and F1-score (above 80%),

suggesting that the selected features in combination with the models were able to retrieve most of the tweets that were about immigrants and had few false positives. In addition, both models also reached an AUC above 0.8, showing that they were reliable prediction models for classifying whether a tweet was about immigrants. When comparing AdaBoost with SVM, the overall performance of SVM was slightly better, with 1.6% performance gain over AdaBoost for F1-score. Therefore, the performance of SVM was the best among all evaluated models. For the second task, the overall performance of AdaBoost and XGBoost was better than the other models. Both of these models reached good precision (above 80%), recall (above 80%), and F1-score (above 80%), and AUC (above 0.8), suggesting that these models were able to reliably classify whether a tweet was prejudiced speech against immigrants. When comparing XGBoost with AdaBoost, the overall performance of XGBoost was slightly better, with a 1.5% of performance gain over AdaBoost for F1-score. Therefore, the XGBoost performed the best among all evaluated models.

Table 2: Model performances for detecting *relevant* and *prejudiced* tweets

Classification	F1-score	Precision	Recall	AUC	Method
Immigrant (Relevant)	82.9	83.0	82.9	0.824	AdaBoost
	73.8	74.5	73.5	0.736	NB
	74.1	78.3	76.0	0.712	XGBoost
	84.5	**84.9**	**84.4**	**0.846**	**SVM**
	75.2	78.1	76.7	0.724	LR
Prejudiced	83.4	84.3	82.7	0.722	AdaBoost
	74.9	83.3	71.3	0.714	NB
	84.9	**85.4**	**84.5**	**0.737**	**XGBoost**
	79.8	85.1	77.3	0.754	SVM
	66.6	85.1	61.5	0.720	LR

5.4.1 Error analysis While classifier overall achieved good accuracy, there are still cases where machine mis-classified the tweets. For example, machine mis-classified relevant tweets as irrelevant, "*This National Guard ""rounding up immigrants"" story is a smear. The MSM is also smearing our military. This is disgusting.* " In this tweets, the use was discussing his or her opinions about immigrants, indicated by National Guard ""rounding up immigrants"" story. There were also cases where machine classified the following tweets about immigrants, but it is about Muslims, "*Jibaal (3 Muslims) got fired from Capital Hill for spying &; ""sending"" data to external server* ". In this case, machine seems to make similar mistakes as human would make. As discussed in Section 5.2.2, one of the common mistake human coders made during the training was to mis-classify minority group such as Muslims as immigrants. This error could be because when we discuss Muslims immigrants, often the case we would use Muslims to refer this group.

In addition to mis-classify relevant tweets, machine also misclassified prejudiced tweets. For example, machine falsely identified a tweet that was not prejudiced as prejudiced, "*Not Criminals, Not Illegals, We Are International Workers, Fuck Donald Trump and His Pinche Wall* ". This error might be because the classifier tends to classified tweets that contain certain hostile words as prejudiced such as " Illegals" and "Criminals". This issue could be addressed by adding a negation feature when classifying prejudiced tweets.

Last, we also saw false negative cases where machine classified the tweet as non-prejudice when the tweet is prejudiced. For example, "*Give me your poor, your tired, your disgusting masses..."* ". As we know, a similar phases were written on the Statue of Liberty in the US, "*Give me your tired, your poor, Your huddled masses yearning to breathe free, The wretched refuse of your teeming shore. Send these, the homeless, tempest-tossed to me, I lift my lamp beside the golden door!* ". This original poem was expressing welcome to immigrants and refugees to this country. However, this user changed its meaning to a negative way and linked immigration to "disgusting masses". While machine classify this tweet as about immigrant but not prejudiced against immigrants, the history context of this expression would tell us that this user still expressed prejudice but in a more implicit way.

5.4.2 Automatically label data To automatically label the unlabeled data, we used the trained SVM to label whether a tweet is about immigrants and the trained XGBoost to label whether a tweet is prejudiced speech against immigrants. Among 889,579 immigrant-related tweets, 490,622 tweets (about 55%) were labeled as about immigrants, and 157,014 (about 18%) were labeled as prejudiced speech against immigrants.

6 Analysis Results

We organize our analysis results to answer *RQ 1* and *RQ 2* in section 6.1, *RQ 3* in section 6.2, and *RQ 4* in section 6.3. The unit of analysis is on the user level.

For each user, we measure the user's interest in the immigration-related topics as "*%Relevance*" – the proportion of user's tweets that are classified as relevant tweets among all of the tweets posted by the user within the study window. This quantity captures the user's relative interest in tweeting about immigrants or immigration-related topics, which serves as an indicator for his/her being aware of the topics. We measure the level of prejudice per user as "*%Prejudice*" – the proportion of user's relevant tweets that are classified as prejudiced tweets within the study window[1]. The two quantities are measured for each user in the dataset in the pre-protest window (within 14 days prior to the protest) and in the post-protest window (within 14 days after the onset of the protest). We further group users into the *prejudiced group* if the user posted any prejudiced tweets during the pre-protest window, and in the *non-prejudiced group* if no prejudiced tweet was observed from the user.

Given our interest in protest exposure, we also identify each user's level of exposure based on their geo-locations – a user is categorized as "high exposure" if the user located in the cities where the protests happened, and "low exposure" otherwise. Table 3 shows a summary of users breaking down by the two events, across two groups (prejudiced and non-prejudiced), and by the level of exposure (high and low), respectively.

6.1 Changes After Protests

6.1.1 Change in Awareness Figure 1 shows changes in user's percentage of tweets that are about immigrants following the "Day Without Immigrants" protest (Figure 1 (a)) and "No Ban, No Wall" protest (Figure 1 (b)). For the "Day Without Immigrants" protest, a

[1]The %Prejudice is set to be 0 if the user didn't post any relevant tweets in the study period.

Table 3: User categorization per event

Protest	User group	High	Low
"Day Without Immigrants"	Prejudiced	698	22,416
	Non-prejudiced	7606	165,616
"No Ban, No wall"	Prejudiced	416	29,486
	Non-prejudiced	4042	168,364

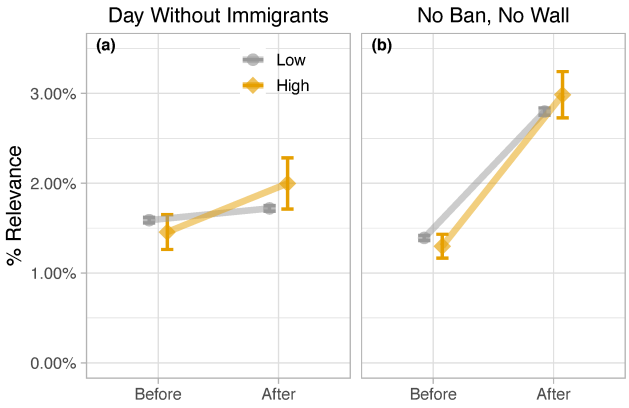

Figure 1: Change of the measured awareness on immigration issues after protests.

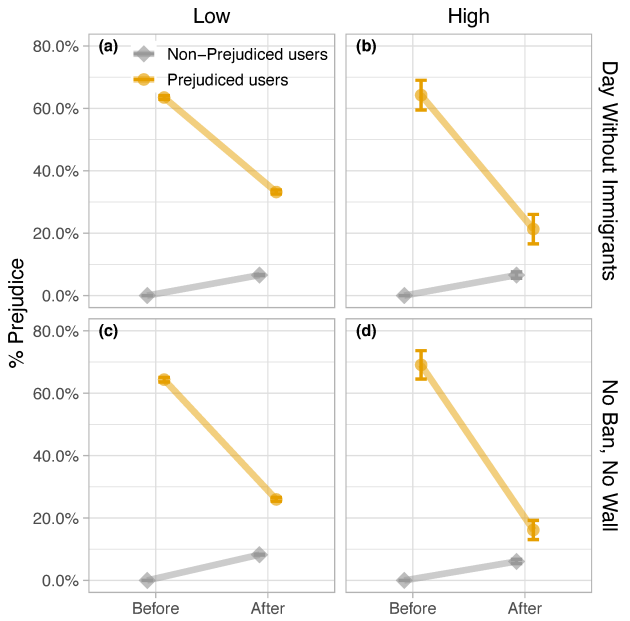

Figure 2: Change of measured prejudice after protests.

Wilcoxon Signed-ranks test indicated that among high-exposure user group, the percentage of tweets about immigrants increased significantly after the protest ($Z = -58.48$, $p = 0.001$, $r = 0.01$). Among low-exposure user group, there was also an increased in the percentage of tweets about immigrants after the protest ($Z = -87.47$, $p < 10^{-3}$, $r = 0.04$).

Similar pattern was observed for "No Ban, No Wall" with a slightly larger magnitude. Among high-exposure user group, the percentage of tweets about immigrants increased after the protest ($Z = -79.661$, $p < 10^{-3}$, $r = 0.06$). Among low-exposure user group, there was also an increased in the percentage of tweets about immigrants after the protest ($Z = -87.26$, $p < 10^{-3}$, $r = 0.03$).

Overall, these findings suggest that there was a significant increase in the percentage of tweets about immigrants among high-exposure and low-exposure users. Compared to the "Day Without Immigrants" protest, the magnitude in the increase among high exposure users was slightly larger for the "No Ban, No Wall" protest. This might be because the "No Ban, No Wall" protest happened on a larger scale. For example, the "No Ban, No Wall" protest happened in 22 cities within the US as well as cities outside of US such as London and Berlin, whereas the "Day Without Immigrants" protest took place in 17 cities within the US.

6.1.2 Change in Prejudice Figure 2 shows the changes in prejudice following protest. Figure 2 (a) and (b) shows high and low exposure groups for "Day Without Immigrants" protest". Figure 2 (c) and (d) shows high and low exposure groups.

After the "Day Without Immigrants" protest, there was a significant decrease in percentage of prejudice among prejudiced user across both high exposure ($Z = -16.96$, $p < 10^{-3}$, $r = 0.66$) and low exposure condition ($Z = -141.97$, $p < 10^{-3}$, $r = 0.07$) However, the measured prejudice significantly increased for non-prejudice users across both high exposure ($Z = -56.74$, $p < 10^{-3}$, $r = 0.27$) and low exposure condition ($Z = -89.19$, $p < 10^{-3}$, $r = 0.06$).

Similar pattern was also found for the "No Ban, No Wall" protest. After the protest, percentage of prejudice significantly decreased among prejudiced user across both high exposure ($Z = -21.29$, $p < 10^{-3}$, $r = 0.76$) and low exposure condition ($Z = -127.68$, $p < 10^{-3}$, $r = 0.07$). However, the measured prejudice significantly increased for non-prejudice users across both high exposure ($Z = -77.77$, $p < 10^{-3}$, $r = 0.28$) and low exposure condition ($Z = -89.04$, $p < 10^{-3}$, $r = 0.03$). It is worth noting that a large effect size was observed for the decrease in the measured prejudice among high exposure prejudiced users across the "Day Without Immigrants" protest ($r = 0.66$) and the "No Ban, No Wall" protest ($r = 0.76$). This finding show that about 80% of prejudiced users would have a prejudice score that was below the mean of their baseline before the protest. It is worth noting that for non-prejudiced users, there was a small effect size for the the increase of prejudice across the "Day Without Immigrants" protest ($r = 0.27$) and the "No Ban, No Wall" protest ($r = 0.28$). This shows that about 60% of non-prejudiced users would have a prejudice score that would be above the mean of their baseline before the protest.

Overall, these findings suggest that users' measured awareness of immigrants increased after the protest. Moreover, the change of users' measured prejudice varied by their exposure level and measured prejudice before the protest.

6.2 Predicting Prejudice Changes

To answer *RQ 2* (what kind of users were more/less likely to change after a protest), we conduct a prediction experiment. The prediction task seeks to predict whether a user's prejudice behavior change (e.g., an increase or a decrease in prejudiced expression) given the user's traits observed before the protest event. By representing

users' traits by various features, we can learn features that are informative for predicting users' behavioral change.

Table 4: Distribution of User Labels

		Prejudice Dropped	Unchanged
Prejudiced User Group	DayWithoutImmigrant	8,038	6913
	NoBanNoWall	7,803	4,754
		Unchanged	Prejudice Increased
Non-prejudiced User Group	DayWithoutImmigrant	77,928	8,275
	NoBanNoWall	73,950	12,661

Table 5: Model performances of prejudice change prediction.

DayWithoutImmigrants Protest

	Model	Accuracy	Precision	Recall	F1
Prejudice Group	Random Forest	0.7437	0.7558	0.7731	0.7643
	Lasso	0.7460	0.7431	0.8064	0.7735
	Ridge	0.7453	0.7406	0.8100	0.7737
	Linear SVM	0.7327	0.7070	**0.8684**	**0.7797**
	SVM w/ Gaussian Kernel	0.7438	**0.7608**	0.7635	0.7622
	XGBoost	**0.7529**	0.7595	0.7906	0.7748
Non-Prejudice Group	Random Forest	0.7122	0.7133	0.7097	0.7115
	Lasso	0.7227	0.7208	0.7270	0.7239
	Ridge	0.7219	0.7209	0.7240	0.7225
	Linear SVM	0.7216	0.7185	0.7287	0.7235
	SVM w/ Gaussian Kernel	0.7217	0.7211	0.7231	0.7221
	XGBoost	**0.7285**	**0.7280**	**0.7297**	**0.7288**

NoBanNoWall Protest

	Model	Accuracy	Precision	Recall	F1
Prejudice Group	Random Forest	0.7329	0.7330	0.7326	0.7328
	Lasso	0.7274	0.7035	0.7861	0.7425
	Ridge	0.7279	0.7038	0.7871	0.7431
	Linear SVM	0.7257	0.6947	**0.8052**	0.7459
	SVM w/ Gaussian Kernel	0.7381	0.7371	0.7402	0.7387
	XGBoost	**0.7495**	**0.7502**	0.7480	**0.7491**
Non-Prejudice Group	Random Forest	0.6885	0.6818	**0.7068**	0.6941
	Lasso	0.6706	0.6635	0.6920	0.6775
	Ridge	0.6705	0.6635	0.6920	0.6775
	Linear SVM	0.6646	0.6543	0.6977	0.6753
	SVM w/ Gaussian Kernel	0.6988	0.6962	0.7052	**0.7007**
	XGBoost	**0.7021**	**0.7039**	0.6975	**0.7007**

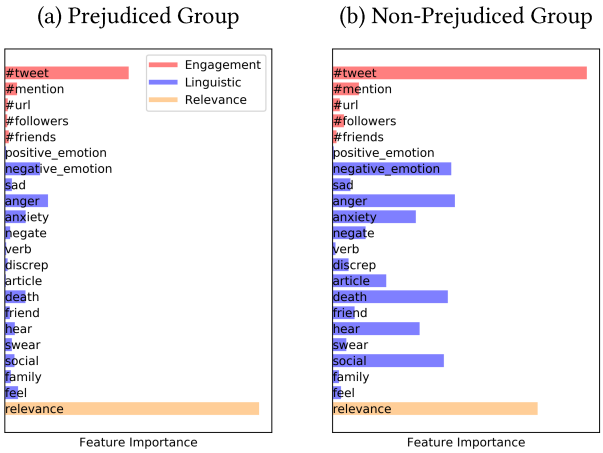

Figure 3: Feature importance in predicting prejudice change.

Experiment Setting. A user's behavioral change is determined by the user's change in prejudice score (i.e., %Prejudice) before and after a protest. We denote a user's prejudice scores as s_{pre} and s_{post} for the pre-and post-protest periods, respectively. As described earlier, each user is assigned to one of the two groups: prejudiced

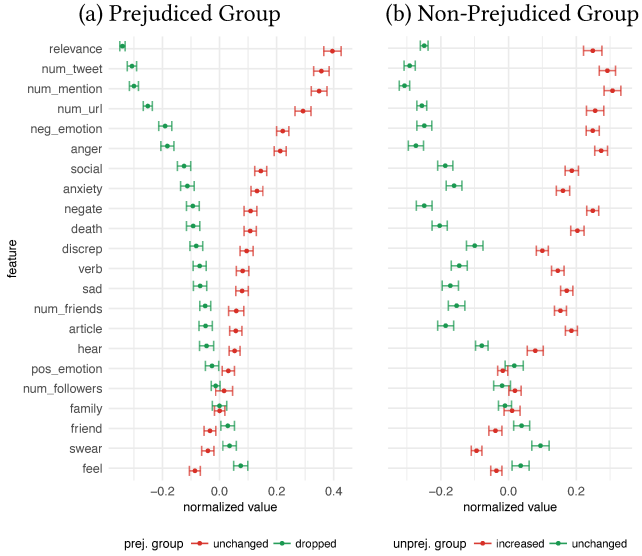

Figure 4: Feature differences between groups. (a) Users in the prejudiced group whose %Prejudice dropped or unchanged after the protest. (b) Users in the unprejudiced group whose %Prejudice unchanged or increased after the protest. Feature values are standardized with error bars indicating the 95% confidence interval.

group if the user's $s_{pre} > 0$, and non-prejudiced group otherwise. For the prejudiced group, we predict whether a user had "prejudice dropped" ($s_{post} < s_{pre}$) or "unchanged" (i.e., remaining the same level of prejudice); for the non-prejudiced group, we predict whether a user had "prejudice increased" ($s_{post} > s_{pre}$) or "unchanged" ($s_{post} = s_{pre} = 0$, i.e., remaining non-prejudiced) . We examine both of the "DayWithoutImmigrants" and "NoBanNoWall" events. The label distributions are given in Table 4. The labels of the prejudiced user in the "Day Without Immigrants" protest are roughly balanced. In the other three cases, we down-sample the instances to make the labels evenly distributed.

Feature Engineering. We extract features from users' data observed in the pre-protest periods (2 weeks before the protests). The feature *relevance* is derived from the user's %Relevance as described earlier. In addition, we include two arch-types of features: *Engagement Features* and *Linguistic Features*.

Engagement Features capture users' engagement in social media. It consists of five features: *number of tweet* reflects the user's online activity level; *number of mention* and *number of friends* together reflect user's interaction with others; *number of followers* captures the extent to which the user's (information) reach to others; *number of urls* (the relative frequency of user's posts containing URL links) reflects the level of the user's information consumption from other online sources. The first two features are normalized by *number of tweet*, while others are logarithmically transformed.

Linguistic Features are derived from users' tweet content using the LIWC lexicon [42]. For each lexicon category, we compute the user-level proportion of tweets containing the corresponding lexicon category within the study period. We include categoreis that are related to our problem context, including *Standard Linguistic*

Dimensions (article, negate, verb), *Social Processes* (social, family, friend), *Affective Processes* (positive/negative emotion, sad, anger, anxiety), *Perceptual processes* (feel, hear), *Informal language* (swear), *Personal concerns* (death) and *Cognitive processes* (discrepancy).

Model Evaluation We evaluate various standard machine learning models including Random Forest (number of trees = 100, max tree depth = 4), Logistic Regression with l1 norm (Lasso), Logistic Regression with l2 norm (Ridge Regression), Linear SVM (C = 0.5), SVM with Gaussion Kernel (C = 0.5), and XGBoost (number of trees = 100, max tree depth = 4).

The prediction performances are evaluated with the 10-fold cross-validation for both tasks. We report *accuracy*, *precision*, *recall*, and *F1 score* as the evaluation metrics in Table 5. The models result in similar performances in each prediction task and XGBoost slightly outperforms other models in all cases. The results show that we are able to make reasonable predictions on prejudice change from the features we engineered.

Feature Analysis To compare the informativeness of various features, we extract and plot the feature importance scores from the Random Forest model. Fig. 3 shows the results from the "Day Without Immigrants" event. The results from the "No Ban, No Wall" exhibit similar pattern and is omitted. As shown in Fig. 3 (a), when predicting the prejudice change among the prejudice group, *relevance* is the most informative feature, the engagement feature *number of tweet* is the second, and several linguistic features including *negative emotion*, *anger*, and *anxiety* are also informative to some extent.

Similar patterns can be observed in predicting the prejudice change for the non-prejudiced group, as shown in Fig. 3 (b), where overall *relevance* and *number of tweet* are the most informative features. Interestingly, the linguistic features – especially *anxiety*, *article*, *death*, *hear* and *social* – are generally more informative in this prediction task. Together, the informativeness of these features suggest that users' behavioral change in response to a protest event are predictable by his/her prior interest in the related topics (*relevance*), prior online activity level *number of tweet*, and through a particular set of linguistic signals.

To further examine "what types of users are more/less likely to change" after a protest, we show in Fig. 4 the differences in feature values aggregated by different change outcomes, for both prejudiced and non-prejudiced groups from the "Day Without Immigrants" event. The results from the "No Ban, No Wall" is omitted as the patterns are highly similar.

Features are ordered from top to bottom by the differences in the means of the two outcomes – the dropped vs. unchanged in Fig. 4 (a), and the increased vs. unchanged in Fig. 4 (b). The plots indicate that, the decrease (or non-increase) in prejudice after the event is often associated with (1) a lower level of relevance and engagement (*number of tweet*, *number of mention*, *number of url*), (2) a lower level of language use in negative affects (*negative emotion*, *anger*, *anxiety*), and (3) a lower level of language use signifying *social* and *death* concepts. In other words, users in the prejudiced group who exhibited higher values in these features were more likely to *resist to change* their prejudiced tendency, compared to users having lower values in these features. Similarly, in the non-prejudiced group, users with higher values in the aforementioned features were more likely to show prejudiced behavior after the event, compared with users having lower values in these features.

In sum, our analysis suggests that Twitter users who are more engaged in tweeting, who engaged more in discussing the immigration-related topics, and who expressed more negative affects, are less likely to have a change in measured prejudice against immigrants throughout the protest events.

6.3 Qualitative Analysis of Change

The previous analysis indicates there was a measurable effect of protests in reducing online prejudiced speech on Twitter. In this section, we study what kind of changes have happened. By qualitatively examining users' tweet content, we identified four major themes as follows.

6.3.1 Seeking prejudice justification One of the pronounced themes among prejudice users is that after the protest, some users started to seek justifications for their prejudices. For example, before the "No Ban, No Wall" protest, a user bluntly expressed prejudice against undocumented immigrants, "*I hope passing Kate's Law is on the 100 day agenda. Sanctuary cities are cesspools of crime and lawbreakers* " (classified as prejudiced) and "*I don't feel bad for illegals who broke laws. I feel bad for families who had loved ones that were killed by illegals* "(classified as prejudiced). This user insulted undocumented immigrants as "illegals" and assumed that these immigrants would commit crimes such as murder (e.g., "killed by illegals"). However, after the protest, this user started to provide justifications for expressing prejudice against undocumented immigrants, "*We must restore rule of law to our borders. #JointAddress #Jointsession #BuildTheWall* " (classified as about immigrants but not prejudiced), and "*Are you in favor of refugee ban, to keep America safe and secure?* " (classified as about immigrants but not prejudiced), and "*Saudi Arabia currently bans ALL refugees from entering their country and they are building a 600 mile wall on their border* " (classified as about immigrants but not prejudiced). These tweets suggested that this user was trying to justify his or her prejudice for example, the reason why US needs to build a wall is because Saudi Arabia is doing it. In addition, this user also associated his or her support for building this wall with "restore rule of law" and the safety of the country. It is important to note that in this case, machine classified this tweet, "*Saudi Arabia currently bans ALL refugees from entering their country and they are building a 600 mile wall on their border* " as classified as about immigrants but not prejudiced. However, it conveys hostility towards immigrants in subtler way because prejudice but the intention of this tweet was to promote punitive immigration policy to exclude immigrants. This also highlights the challenge of using machine to classify prejudice due to the complexity and subtlety of the way people express their prejudices.

6.3.2 Attacking prejudice suppressors In addition to provide justifications, some users also started to attack users who are supportive of immigrants and their protests. For example, before the "No Ban, No Wall" protest, a user explicitly express prejudice against undocumented immigrants, "*All illegals should be deported within six months and fined 250k* " (classified as prejudiced) and "*I rather my tax dollars go towards a wall on the border than supporting tens of millions of illegal aliens for decades to come.* "(classified as prejudiced) These are prejudiced expressions that intended to exclude

undocumented immigrants from this country and have the assumption that undocumented immigrants live on welfare programs (e.g., cost tax dollars). However, this user started to attack users who are supportive of immigrants and their protests, "*If you love refugees so much you should open your home to them, and sponsor them. Don't judge others by standards you don't live* " (classified as about but not prejudiced) and "*Who gives a crap! Every country has the right to allow in their country whom they want. I can't immigrate to any country I want* " (classified as not about immigrants). It should be noted that while this example convey user's prejudice against immigrants, the way that this user expressed was not explicit.

6.3.3 Gloating over protest consequences Protest comes with a social price, especially true for the disruptive ones such as the "Day Without Immigrants" protest. On the protest day, many people stayed home rather than go to work. A price associated with their behavior might be making their bosses unhappy and got fired. After the protest, rather than expressing sympathy to these immigrants, prejudiced users gloated over immigrants who suffered from these consequences. For example, users posted,"*#DayWithoutImmigrants means billionaires make less money because they can't hire illegals -they have to pay Americans FAIR WAGES* " (classified as prejudiced) and "*Companies Firing Illegals who Participated in #ADayWithoutImmigrants* ". Some even said, "5 immigrants were just fired for not showing up for work. Apparently they forgot Louisiana is a hire/fire #ADayWithoutImmigrants" (classified as about immigrants but not prejudiced). Note that in the last tweet, while it did not explicitly express the user's prejudice, it did convey this user's hostility about immigrants in an implicit way.

6.3.4 Confirming prejudiced belief Before the "DayWithoutImmigrants", users who cried for excluding immigrants started to use the protest to confirm their prejudiced belief. For example, before the protest, "*a user posted "Our legal system is broken! "77% of refugees allowed into U.S. since travel reprieve hail from seven suspect countries* "(classified as about immigrants but not prejudiced) and "*the illegal Muslim who lived4 yrs in Indonesia w NO immigration papers to return to US.over threw US gov. =not in Jail* ". These tweets show that this user blamed refugees (illegal Muslim, refugees) (classified as prejudiced) for the broken legal system in the US. However, after the protest, this user used this protest to confirm his or her prejudiced belief. For example, this user posted, "*What US jobs did immigrants steal from Americans? Find out on #Daywithoutimmigrants.* ", "*Immigrant day just proved they aren't needed. Might not have been their goal!! Went shopping, ate out. #immigrantday #Daywithoutimmigrants* " (classified as about immigrants but not prejudiced), and "*Welfare/food stamps/victims of crimes could use a #DayWithoutImmigrants ILLEGAL IMMIGRANTS* " (classified as prejudiced). These tweets show that the user used the protest as a way to confirm his or her prejudices such as immigrants take away jobs, live on welfare, commit crimes, and should be excluded from the country.

Overall, the qualitative analysis show that while prejudiced users might be less likely to express their prejudice in a explicitly way, they still remain prejudiced. However, their ways of expressing prejudice became sutler after the protest. This is especially true when user became subtler in the ways of expressing their prejudice.

7 Discussion

To understand protests as an intervention to counter online prejudice, we proposed a study design that leverages a prejudice classifier to measure the effect of a social protest on reducing online prejudice. Specifically, we focused on online prejudice against immigrants and explored the changes in awareness of immigrant and prejudice following protests. Our findings indicate there were both negative and positive changes in the way online prejudice is being expressed following protests. We also found that users who were less engaged in social media (e.g., having fewer tweets, friends, and followers) might be more likely to have a drop in the measured prejudice following a protest, compared with the highly engaged online users. To contextualize prejudice change, we analyzed in what way prejudiced users changed their prejudice after protests. One notable finding is that after a protest, ways of expressing prejudice against immigrants became subtler. It should also be noted that determining whether an expression involves prejudice may be subjective to varying degrees and thus is non-trivial even for humans. This reveals new challenges and needs to train more adaptable machine classifiers to detect online prejudice.

While the results of machine learning models showed high reliability for the classification tasks, it is important to note that the trained classifiers were not immune to the pre-existing biases in the training data, e.g., the data may have more tweets from users from particular demographic or geographic groups. In addition, the classifiers for user prejudice change prediction may inherit biases from the imperfect classifiers for identifying prejudiced tweets. Future research might consider further improving the user-level change prediction with better annotated data.

In this work, we selected users based on a particular set of hashtags and geo-tagged tweets that contained specific related keywords. This convenient sampling approach may introduce potential selection bias where the sample included in this study might not reflect the general Twitter population. The focus on the tweets and protest events in the U.S. also limits the generalizability of the findings drawn from this study. As for study duration, we examined the short-term effect of change in the measured prejudice following protests; the results may not be generalizable to longer-term changes. Future work may consider studying a broader population across different countries and studying the long-term effect of protests on online prejudice.

Last, this research had no control over a number of confounding variables because protest events occurred in a natural environment. The decrease in online prejudice might be related to user's political ideology, immigrant population size in a given city or state.

Acknowledgments

The authors would like to acknowledge the support from NSF #1637067, #1739413, the DARPA UGB, and AFOSR awards. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

References

[1] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice. (1954).

[2] Edwin Amenta and Michael P Young. 1999. Democratic states and social movements: Theoretical arguments and hypotheses. *Social problems* 46, 2 (1999), 153–168.

[3] Kenneth T Andrews, Kraig Beyerlein, and Tuneka Tucker Farnum. 2015. The legitimacy of protest: explaining White Southerners' attitudes toward the civil rights movement. *Social Forces* 94, 3 (2015), 1021–1044.

[4] J Bacon and A Gomez. 2017. Protests against Trump's immigration plan rolling in more than 30 cities. https://www.usatoday.com/story/news/nation/2017/01/29/homeland-security-judges-stay-has-little-impact-travel-ban/97211720/

[5] Lee Ann Banaszak and Heather L Ondercin. 2016. Public opinion as a movement outcome: The case of the US women's movement. *Mobilization: An International Quarterly* 21, 3 (2016), 361–378.

[6] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 54–63.

[7] Anat Ben-David and Ariadna Matamoros-Fernández. 2016. Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication* 10 (2016), 1167–1193.

[8] E Bermudez, R Vives, S Kohli, and M Etehad. 2017. Day Without Immigrants' resonates across Los Angeles, even if many still go to work. http://www.latimes.com/local/california/la-me-ln-day-without-immigrants-20170216-story.html

[9] Catherine Blaya. 2019. Cyberhate: A review and content analysis of intervention strategies. *Aggression and violent behavior* 45 (2019), 163–172.

[10] Denise M Bostdorff. 2004. The internet rhetoric of the Ku Klux Klan: A case study in web site community building run amok. *Communication Studies* 55, 2 (2004), 340–361.

[11] Michael P Boyle, Lauren Dioguardi, and Julie E Pate. 2016. A comparison of three strategies for reducing the public stigma associated with stuttering. *Journal of fluency disorders* 50 (2016), 44–58.

[12] Regina Branton, Valerie Martinez-Ebers, Tony E Carey Jr, and Tetsuya Matsubayashi. 2015. Social protest and policy attitudes: The case of the 2006 immigrant rallies. *American Journal of Political Science* 59, 2 (2015), 390–402.

[13] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, 1 (2016), 11.

[14] CBS. 2017. Day Without Immigrants Boycott Having An Impact In North Texas. http://dfw.cbslocal.com/2017/02/16/day-without-immigrants-boycott-having-an-impact-in-north-texas/

[15] Jason Chan, Anindya Ghose, and Robert Seamans. 2016. The internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly* 40, 2 (2016), 381–403.

[16] Michael Chau and Jennifer Xu. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies* 65, 1 (2007), 57–70.

[17] Rubén Comas-Forgas, Jaume Sureda-Negre, and Aina Calvo-Sastre. 2017. Characteristics of Cyberbullying Among Native and Immigrant Secondary Education Students. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)* 7, 1 (2017), 1–17.

[18] Patrick W Corrigan, Scott B Morris, Patrick J Michaels, Jennifer D Rafacz, and Nicolas Rüsch. 2012. Challenging the public stigma of mental illness: a meta-analysis of outcome studies. *Psychiatric services* 63, 10 (2012), 963–973.

[19] Patrick W Corrigan, David Roe, and Hector WH Tsang. 2011. *Challenging the stigma of mental illness: Lessons for therapists and advocates*. John Wiley & Sons.

[20] S Daniels and M Graham. 2017. Immigrants across the U.S. skip work, school in anti-Trump protest. https://www.reuters.com/article/us-usa-immigration-protest/immigrants-across-the-u-s-skip-work-school-in-anti-trump-protest-idUSKBN15V19M

[21] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[22] B Demick. 2017. Thousands at JFK airport in New York protest new visa and refugee suspensions. http://www.latimes.com/politics/washington/la-na-trailguide-updates-thousands-at-jfk-airport-in-new-york-1485642642-htmlstory.html

[23] Thomas E Ford, Sabrina R Teeter, Kyle Richardson, and Julie A Woodzicka. 2017. Putting the brakes on prejudice rebound effects: An ironic effect of disparagement humor. *The Journal of social psychology* 157, 4 (2017), 458–473.

[24] William A Gamson. 1989. Reflections onThe Strategy of Social Protest. In *Sociological Forum*, Vol. 4. Springer, 455–467.

[25] Marco Giugni. 2004. *Social protest and policy change: Ecology, antinuclear, and peace movements in comparative perspective*. Rowman & Littlefield.

[26] Shannon Greenwood, Andrew Perrin, and Maeve Duggan. 2016. Social media update 2016: Facebook usage and engagement is on the rise, while adoption of other platforms holds steady. Pew Research Center.

[27] Xingsheng He, Di Lu, Drew Margolin, Mengdi Wang, Salma El Idrissi, and Yu-Ru Lin. 2017. The signals and noise: actionable information in improvised social media channels during a disaster. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 33–42.

[28] B Horn. 2017. Thousands at Philly airport call Trump's order 'un-American'. Delaware Online. https://www.delawareonline.com/story/news/local/2017/01/29/delaware-immigration-trump-protest/97215226/

[29] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. 2013. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd international conference on World Wide Web*. 667–678.

[30] Jenny Kitzinger. 1995. Qualitative research: introducing focus groups. *Bmj* 311, 7000 (1995), 299–302.

[31] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

[32] Yu-Ru Lin, Brian Keegan, Drew Margolin, and David Lazer. 2014. Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. *PloS one* 9, 5 (2014), e94093.

[33] Yu-Ru Lin, Drew Margolin, and Xidao Wen. 2017. Tracking and analyzing individual distress following terrorist attacks using social media streams. *Risk analysis* 37, 8 (2017), 1580–1605.

[34] Estefanía Lozano, Jorge Cedeño, Galo Castillo, Fabricio Layedra, Henry Lasso, and Carmen Vaca. 2017. Requiem for online harassers: Identifying racism from political tweets. In *2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, 154–160.

[35] Jason Luty, Okon Umoh, Mohammed Sessay, and Arghya Sarkhel. 2007. Effectiveness of Changing Minds campaign factsheets in reducing stigmatised attitudes towards mental illness. *Psychiatric Bulletin* 31, 10 (2007), 377–381.

[36] C Neil Macrae, Galen V Bodenhausen, Alan B Milne, and Jolanda Jetten. 1994. Out of mind but back in sight: Stereotypes on the rebound. *Journal of personality and social psychology* 67, 5 (1994), 808.

[37] Soumyajit Mazumder. 2018. The persistent effect of US civil rights protests on political attitudes. *American Journal of Political Science* 62, 4 (2018), 922–935.

[38] James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations are users too: Characterizing and detecting the presence of organizations on Twitter. In *Ninth International AAAI Conference on Web and Social Media*.

[39] Lacy G McNamee, Brittany L Peterson, and Jorge Peña. 2010. A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs* 77, 2 (2010), 257–280.

[40] Richard Jayadi Oentaryo, Jia-Wei Low, and Ee-Peng Lim. 2015. Chalk and cheese in twitter: Discriminating personal and organization accounts. In *European Conference on Information Retrieval*. Springer, 465–476.

[41] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The effect of extremist violence on hatexffful speech online. In *Twelfth International AAAI Conference on Web and Social Media*.

[42] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC 2015*. Technical Report.

[43] Associated Press. 2017. More Than 3,000 Protest Trump Executive Order At Sea-Tac Airport. https://www.kiro7.com/news/local/anti-trump-protesters-gather-again-in-seattle/489087028

[44] B. Quimby. 2017. Thousands protest Trump policies at 2 rallies in Portland. https://www.pressherald.com/2017/01/29/anti-trump-rallies-are-popping-up-around-maine-today/

[45] K Reece. 2017. Dallas workers join national 'Day Without Immigrants' protest. http://www.wfaa.com/news/local/dallas-workers-to-take-part-in-day-without-immigrants/408856343

[46] L Robbins and A Correal. 2017. On a 'Day Without Immigrants,' Workers Show Their Presence by Staying Home. https://www.nytimes.com/2017/02/16/nyregion/day-without-immigrants-boycott-trump-policy.html

[47] Jennifer D Rubin and Sara I McClelland. 2015. 'Even though it's a small checkbox, it's a big deal': stresses and strains of managing sexual identity (s) on Facebook. *Culture, health & sexuality* 17, 4 (2015), 512–526.

[48] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[49] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

[50] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. 2009. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence* 23, 04 (2009), 687–719.

[51] Imogen Tyler and Katarzyna Marciniak. 2013. Immigrant protest: an introduction. *Citizenship studies* 17, 2 (2013), 143–156.

[52] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*.

[53] Maykel Verkuyten and Borja Martinovic. 2015. Behind the ethnic–civic distinction: public attitudes towards immigrants' political rights in the Netherlands.

Social science research 53 (2015), 34–44.
[54] Daniel M Wegner, Ralph Erber, and Sophia Zanakos. 1993. Ironic processes in the mental control of mood and mood-related thought. *Journal of personality and social psychology* 65, 6 (1993), 1093.
[55] Daniel M Wegner and David J Schneider. 1989. Mental control: The war of the ghosts in the machine. *Unintended thought* (1989), 287–305.