Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Identifying Hurricane Evacuation Intent on Twitter

Xintian Li[1] and **Samiul Hasan**[2] and **Aron Culotta**[1]
[1]Department of Computer Science, Tulane University, New Orleans, LA
[2]Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL
xli71@tulane.edu, samiul.hasan@ucf.edu, aculotta@tulane.edu

Abstract

Evacuations have a significant impact on saving human lives during hurricanes. However, as a complex dynamic process, it is typically difficult to know individual evacuation decisions in real-time. Since a large amount of information is continuously posted through social media platforms, we can use them to understand individual evacuation behavior. In this paper, we collect tweets during Hurricane Irma in 2017 and train a text classifier in an active learning way to distinguish tweets expressing positive evacuation decisions from both negative and irrelevant ones. Additionally, we perform a demographic analysis and content clustering to investigate the potential causes and correlates of evacuation decisions. The results can be used to help inform planning strategies of emergency response agencies.

1 Introduction

Extreme weather events like hurricanes often lead to significant physical and economic losses. Evacuation has always been an important measure to save human lives and reduce losses during such events. However, it is usually difficult to know about the evacuation behavior of a population especially as a hurricane develops in real time. Post-hurricane surveys can help us learn about the behaviors, but they are costly, have low response rates, and are of limited use during unfolding crisis events. Since a considerable number of people nowadays share their activities and thoughts through social media platforms, a massive volume of real-time information is provided with valuable understandings of individual behaviors.

Recently, a number of methods have been proposed to study evacuation behaviors during hurricanes using geolocated Twitter data (Martín, Li, and Cutter 2017; Kumar and Ukkusuri 2018; Stowe et al. 2018; Martín, Cutter, and Li 2020; Hong and Frias-Martinez 2020; Roy and Hasan 2021). While these approaches have provided fine-grained insights into mobility patterns, they have several limitations: (1) only about 1% of tweets carry geolocation information (Tasse et al. 2017), (2) they provide little information about the reasons people give for evacuating or not, (3) they do not provide enough lead time to allow emergency managers and transportation engineers to respond to evacuation compliance that is greater or less than expected.

To address these limitations, in this paper we propose a text classification approach to identify tweets expressing a user's intent to evacuate or stay in place in the days before a hurricane's landfall. Using text rather than geocoordinates greatly expands the sample size, enabling a more fine-grained analysis of demographic correlates of behavior, as well as a content analysis to identify reasons for user decisions. Additionally, because such tweets are posted hours or days before the user actually evacuates, they can potentially serve as a leading indicator of behavior, providing greater lead time to practitioners. We present a case study of the approach applied to Hurricane Irma, which made landfall in Florida in 2017. The three primary research questions of this paper are as follows:

RQ1: Can we identify tweets expressing evacuation intent? We create a new dataset of 5,000 evacuation related tweets from Hurricane Irma annotated into one of three classes (will evacuate, will not evacuate, neutral). We train a text classifier that achieves 0.94 AUC on identifying tweets expressing an intent to evacuate.

RQ2: How do temporal trends of evacuation tweets align with other measures of evacuation behavior? Using the above classifier to annotate a larger sample of tweets, we compare the volume of evacuation tweets with survey data and traffic data, finding a moderate agreement ($\sim$.6 correlation) between data sources, suggesting that evacuation tweets serve as a leading indicator of evacuation behavior.

RQ3: What factors contribute to evacuation decisions? A regression analysis over Twitter users finds that non-evacuees tend to be older, less wealthy, less educated, and less white. A cluster analysis of tweets from non-evacuees finds that typical reasons given for not evacuating include taking care of family and pets, work requirements, and lack of fuel for driving.

2 Related Work

Understanding evacuation behavior is an active area of research in emergency management and transportation engineering (Huang, Lindell, and Prater 2016). Typically, data are collected through surveys of residents after the disaster, to understand who evacuated, when, and why (Yang et al.

regular expression	matched	labeled	positive	negative	neutral
`([.,!"\-:]\s+|^(@\w+ )*)(evacuat|leav|escap|head)ing`	1,162	500	253	11	236
`n[\'o]?t (evacuat|leav|escap)`	1,346	1,000	22	326	652
`\bi(\'| a)?m( \w+)*? (evacuat|leav|escap|head)ing`	503	500	240	19	241
`\bwe((\'| a)re| r)( \w+)*? (evacuat|leav|escap|head)ing`	3,762	500	345	9	146
remaining	29,409	500	59	12	429
total	36,182	3,000	919	377	1,704
after active learning annotation		5,000	1,727	1,267	2,006

Table 1: Number of tweets matched by each regular expression and labeled in each class.

2016; Huang, Lindell, and Prater 2017; Wong, Shaheen, and Walker 2018). In addition to factors like perceived risk and storm conditions, other variables such as socio-economics have also been found to play a role. Traffic data have also been used to estimate the number of evacuees over time, though these can result in undercounts since the number of occupants in each car is unknown (Dow and Cutter 2002).

Several factors have motivated the use of social media to understand evacuation behavior, including dwindling survey response rates (Johnson and Wislar 2012), limited coverage of vulnerable populations, and a need for real-time situational awareness during an unfolding disaster. For example, being able to forecast evacuation demand in advance allows traffic congestion mitigation strategies. Likewise, understanding the demographics of those who will/will not evacuate allows communication managers to tailor their messaging to the appropriate groups. The challenges of this domain have spurred numerous avenues of research in computational and informatics fields, such as the Workshop on Disaster Management at CIKM'16 (Castillo et al. 2016). Early work in this area focused on extracting information from social media and building information management tools to triage multiple streams of data (Yin et al. 2012; Rudra et al. 2015; Li et al. 2016; Wen, Lin, and Pelechrinis 2016).

Martín, Li, and Cutter (2017) were among the first to use social media data to study evacuation behavior, using geolocated tweets to identify 747 users who evacuated South Carolina during Hurricane Matthew in 2016. They find similar evacuation compliance rates as survey data. Kumar and Ukkusuri (2018) use a similar methodology to study Hurricane Sandy in New York City, identifying 98 evacuees based on geolocated tweets. Hong and Frias-Martinez (2020) use geolocated tweets to identify evacuees during Hurricane Irma to estimate evacuation patterns – i.e., to where do residents of each county evacuate? Stowe et al. (2018) annotated 200 Twitter users as evacuated or not during Hurricane Sandy, using geolocation and textual features to predict the class label, though with limited accuracy (.64 F1). Roy and Hasan (2021) identified 252 geolocated Twitter users who evacuated during Hurricane Irma and trained a hidden Markov model to predict movements (e.g., when a user will choose to evacuate). Recently, Martín, Cutter, and Li (2020) also analyzed geolocated Twitter users during Hurricane Irma, and manually inferred gender, age, and race/ethnicity attributes for each user. They found that evacuees tended to be younger and white, while gender was not a significant factor. While Twitter is the most common social media used in this area due to its public nature, Metaxa-Kakavouli, Maas, and Aldrich (2018) is a notable exception, using Facebook data to conclude that users with more geographically diverse friend networks are more likely to evacuate, perhaps due to the existence of social connections outside of their home location.

While the above approaches demonstrate the potential of Twitter to provide insights into evacuation behavior, the use of geolocated tweets severely limits the sample size, as such tweets are estimated to be about 1% of all tweets (Tasse et al. 2017). Furthermore, while geolocated tweets are useful for observing real-time location, they rarely contain textual information to reveal the rationales people give for evacuating or not. Finally, by identifying tweets indicating an intent to evacuate in the future, we can identify much earlier leading indicators of evacuation behavior, potentially providing practitioners with the lead time to intervene as appropriate. Our primary contributions are to provide a methodology to identify tweets expressing an intent to evacuate or stay and to conduct an in-depth demographic and temporal analysis to understand factors related to such decisions.

3 Methods and Results

In this section we describe the data, methods, and results to address the research questions from §1.[1]

3.1 Data

Our primary data are tweets related to Hurricane Irma, which made landfall in Florida, USA on September 10th, 2017. We are primarily interested in tweets expressing evacuation intent, so we can analyze the content expressing rationales for decision making. To collect an initial sample of interest, we used the full-archive search endpoint provided by Twitter API (Twitter 2021) to search for all public tweets from September 4th to September 17th containing one of the following keywords: *evacuate, evacuating, leave, leaving, escape, escaping*. To focus on tweets describing the user's own situation, and exclude discussions of news stories, we

[1]Replication materials are at https://github.com/tapilab/icwsm-2022-hurricane

	All labeled tweets						Tweets before Florida landfall					
class	training	precision	recall	f1-score	auc	support	training	precision	recall	f1-score	auc	support
positive	3,000	0.77	0.80	0.78	0.926	924	2,399	0.79	0.83	0.81	**0.929**	877
	3,500	0.79	0.83	0.81	0.921	1,323	2,840	**0.81**	0.86	0.83	0.915	1,228
	4,000	0.78	0.86	0.82	0.908	1,678	3,246	0.80	**0.88**	**0.84**	0.906	1,532
	4,500	0.78	0.85	0.81	0.917	1,697	3,696	0.80	0.87	0.84	0.918	1,549
	5,000	0.78	0.83	0.81	0.922	1,727	4,121	0.80	0.85	0.83	0.923	1,578
negative	3,000	0.73	0.64	0.68	0.917	410	2,399	0.74	0.66	0.69	0.923	363
	3,500	0.70	0.64	0.67	0.915	428	2,840	0.73	0.63	0.68	0.919	380
	4,000	0.71	0.58	0.64	0.907	464	3,246	0.72	0.60	0.65	0.907	405
	4,500	0.80	0.80	0.80	0.946	895	3,696	**0.82**	0.81	0.82	0.948	800
	5,000	0.79	0.85	0.82	0.951	1,267	4,121	0.80	**0.86**	**0.83**	**0.950**	1,122

Table 2: Classification using logistic regression: metric scores of 10-fold cross validations in active learning. The training column shows the size of the entire training set for each corresponding classifier. The support is the actual number of examples from that class in the training set. All other columns indicate different metric scores of evaluations.

method	precision	recall	f1	accuracy	auc
GloVe + LSTM	0.74	0.75	0.75	0.75	0.89
LSTM	0.76	0.76	0.76	0.76	0.89
Tf-idf + LogReg	0.79	0.79	0.79	0.79	0.92
LogReg	0.80	0.80	0.80	0.80	0.91
BERT	**0.83**	**0.83**	**0.83**	**0.83**	**0.94**

Table 3: 10-fold classification accuracy for the three-class evacuation intent task.

removed retweets, tweets that have URLs, and tweets that contain the terms *people* or *residents*. This filtering resulted in a total of 969,796 tweets posted by 676,655 users.

To further focus the tweets on Florida residents, we infer the state location of each user using Carmen (Que and Dredze 2017), which considers both geocoordinates (when present) as well as the location field from the user's profile. We also searched the location field for Florida city names (Miami, Orlando, Tampa, etc.) to identify additional relevant users. After this filtering, the data contained 36,280 tweets from 23,814 users. Of these, only 4,571 tweets have geocoordinates, indicating the increased sample size made possible by considering the user location field instead of just geocoordinates.

Data Annotation After initial data exploration, we identified three classes of interest to categorize tweets in terms of evacuation intent: 1) *positive*: the user was evacuating, had evacuated, or showed a clear intention of evacuation; 2) *negative*: the user indicated that they would not or could not evacuate; 3) *neutral*: the tweet is irrelevant, or it is hard to infer the user's intent.

We performed several rounds of annotation, mixing random sampling with keyword search to ensure an adequate number of examples from each class. Two co-authors of the paper annotated data initially to converge on annotation guidelines. In the first round, we sampled 500 tweets at random, then augmented that with 2,500 tweets matching one of four regular expressions (Table 1), aimed at identifying more positive and negative examples. The first pattern matches sentences containing one of the four words: evacuating, leaving, escaping, heading, and where subject is omitted. Matched sentences are likely to describe the user's own behavior. The second pattern matches one of the three keywords with a preceding "not", intending to find negative tweets. The following two expressions are used to match sentences in which the subject of the evacuation is "I" or "we". Among tweets obtained by applying above regular expressions, we annotated 500 of them in each group except the second one in which 1,000 tweets were labeled. We also labeled 500 tweets among the remaining ones not matched by any mentioned expressions. In total, we annotated 3,000 tweets to form our initial dataset, of which 919 are positive, 377 are negative, and the remainder neutral.

We used this initial labeled dataset to train a logistic regression classifier (more details below), which we applied in an active learning process to label an additional 2,000 tweets. To do so, we sampled tweets from each class as predicted by the classifier, annotated them, added them to the training set, and repeated the process. The final annotated dataset contains 5,000 tweets: 1,727 positive, 1,267 negative, and 2,006 neutral. Based on a sample of 100 tweets labeled by two co-authors of the paper, we find 84% agreement, and Cohen's kappa of 0.76.

In the following sections, we first describe text classification experiments with this labeled data, then perform a number of subsequent analyses to address the research questions

negative		neutral		positive	
term	coefficient	term	coefficient	term	coefficient
NOT_evacuating	2.447	alone	1.527	evacuating	2.537
NOT_leaving	2.403	if	1.508	i'm evacuating	1.732
NOT_evacuate	2.256	about evacuating	1.332	we are	1.731
NOT_fu*king leaving	2.123	is evacuating	1.305	we're leaving	1.699
supplies	2.026	they	1.304	driving	1.634
NOT_fu*king	1.990	or	1.168	tomorrow	1.624
i'm NOT_evacuating	1.387	girl	1.157	escaping	1.420
i'm	1.386	leave	1.126	had to	1.381
i'm staying	1.363	county	1.109	planning to	1.313
staying	1.352	if we	1.102	leaving tomorrow	1.296

Table 4: Classification using logistic regression: top 10 features for each class

prediction	truth	probability	tweet
negative	neutral	0.51	Everyone I'm with wants to evacuate [CITY]. But my mom, brother and sister are here, dad is still south. I'm so torn.
negative	neutral	0.51	305 till I die translates to don't evacuate huh
neutral	positive	0.50	Evacuating to Palm Beach and the bugs are evacuating to [CITY], one of us is a bird brain ... #HurricaineIrma2017
neutral	positive	0.50	Ma get your things we're leaving, [NAME] and [NAME] coming for us too
positive	neutral	0.51	@*** Leaving town?
positive	neutral	0.51	supposed to be leaving to [CITY] on Thursday now cus of this hurricane they might cancel my flight.

Table 5: Sample of false positive predictions for each class (with minor edits for anonymity).

outlined in §1.

3.2 Classification

Our goal is to fit a classifier on the annotated tweets, then apply it to the remaining tweets to analyze temporal and demographic trends in evacuation behavior. We compared three classification approaches: *Logistic Regression*, a *Long short-term memory* (LSTM) neural network and *Bidirectional Encoder Representations from Transformers* (BERT). For logistic regression, we considered both binary and tf-idf features, using unigrams and bigrams. Our tokenizer separates mentions, hashtags, and emojis into distinct tokens. We additionally include a simple negation feature to track polarity (e.g., "not going to leave" becomes "NOT_going, NOT_to, NOT_leave").

For the LSTM model, we also have two settings: 1) use pre-trained word vectors for the embedding layer (**GloVe + LSTM**); 2) learn the embedding together with the whole model (**LSTM**). For the pre-trained embedding, we use 200-dimension vector representations trained using Twitter data from Global Vectors for Word Representation (GloVe) (Pennington, Socher, and Manning 2014). In the latter setting, the embedding dimension is 32. The following parts of the model are the same for both settings: a 1D convolution layer containing 64 filters with the kernel size of 5, a max pooling layer with the window size of 3; an LSTM layer with 32 units, and the output layer with softmax activation. This architecture follows the C-LSTM model for text classification, which was found to be more effective than using a basic LSTM (Zhou et al. 2015). To address class imbalance, for all models instances are weighted by the inverse of class frequency.

For the BERT model, we use a pretrained model on English language with a sequence classification head and fine-tune it with our data. The model is uncased, with 12 hidden layers, hidden size of 768, 12 attention heads, containing about 110M parameters. The original vocabulary consists of 30,522 tokens and we add 181 new ones to include emojis. We fine-tune the model by training it for 3 epochs with a learning rate of 5e-5.

Table 3 summarizes the accuracy of the five models using 10-fold cross validation.[2] Precision, recall, and F1 are the

[2]These results are for the 4,121 tweets before landfall, since this

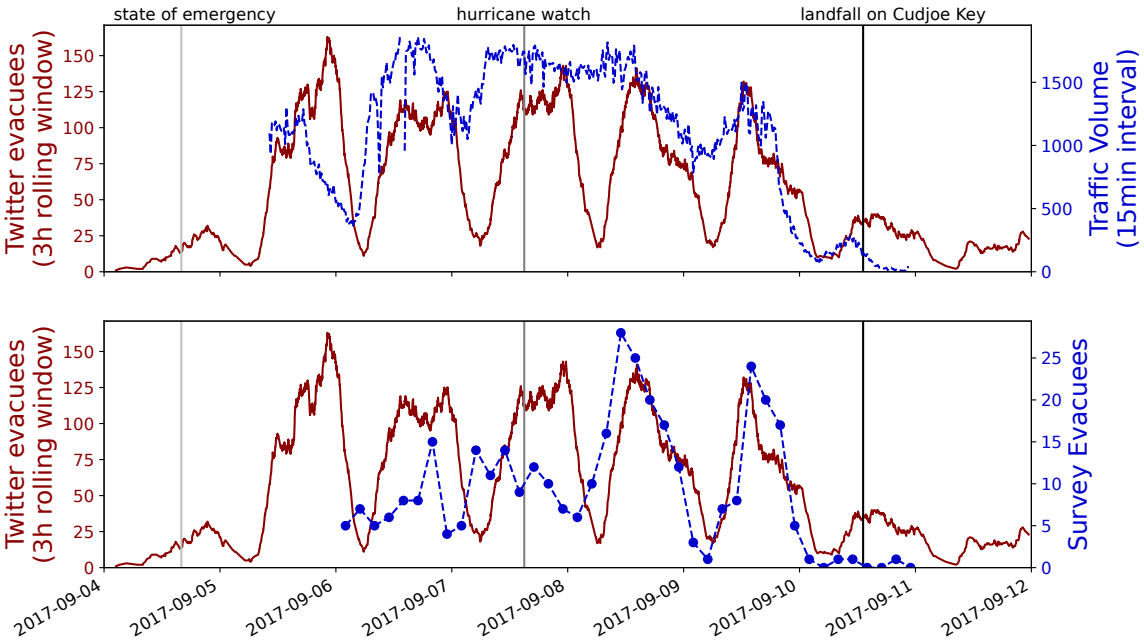


Figure 1: Comparison of the volume of tweets expressing an intent to evacuate with traffic volume (top) and results from a retrospective survey of Florida residents (Wong, Shaheen, and Walker 2018) (bottom).

weighted average across the three classes. The BERT model outperforms the second-best approach by about 3%. Despite its simplicity, the logistic regression model with binary input features (**LogReg**) is the second-best approach. We speculate that the short tweet length, combined with a small training set, limits the effectiveness of the more complex LSTM approach. Due to its simplicity and effectiveness, we use logistic regression for subsequent analyses. Additionally, logistic regression is easier to interpret, allowing us to view the top coefficients (Table 4).

Table 2 additionally shows how the accuracy of the logistic regression classifier improves with more labeled data. For tweets posted before landfall, the positive F1 increases from .81 to .84 when we increase the training set from 3,000 to 4,000 examples; the negative recall improves further from .60 to .86 with the addition of 1,000 more examples.

Table 4 lists the top 10 features for each class in terms of importance. For the negative class, we can see the influence of the negation feature to capture polarity. For the neutral class, we see many messages discussing someone else's plans to evacuate, as well as equivocation ("thinking about evacuating"). For the positive class, we see messages describing plans for a later time (e.g. "tomorrow"). Table 5 shows a sample of typical errors made by the classifier (altered slightly for anonymity). As is often the case with social media, additional context may help disambiguate difficult examples.

For subsequent analysis, we select the best performing logistic regression classifier, which has a positive class precision of .80 and recall of .85 when applied to tweets posted before landfall. We apply this classifier to all remaining tweets to further understand evacuation decisions during Hurricane Irma.

3.3 Alignment with Other Measures of Evacuation Behavior

Our second research question asks: *How do temporal trends of evacuation tweets align with other measures of evacuation behavior?* This serves both as an additional validation measure of the classifier, as well as an indicator of whether Twitter trends are reflective of real-world behaviors. Applying the classifier above to all of the tweets, we identify an additional 2,446 positive tweets from 2,228 unique users. Combined with the annotated data, our final data contain 4,173 positive tweets from 3,732 users. We aggregate positive examples over time to create a time series of evacuation intent tweets.

We compare the Twitter-derived time series with two alternative measures of evacuation behavior. The first is an online survey of 645 Florida residents conducted during the three months after Hurricane Irma (Wong, Shaheen, and Walker 2018; Wong et al. 2020). The survey asked respondents whether they evacuated or not, and if so, when they did so, motivations for doing so, and the nature of their evacuation. Evacuation time responses are reported at three hour intervals. Such post facto surveys are common in emergency management research (Huang, Lindell, and Prater 2016), as it is often infeasible to collect data in the chaotic days leading up to a hurricane. Even such traditional methods have their limitations – e.g., this survey has a relatively small number of respondents, over-samples wealthy and white individuals, and is susceptible to recall bias, since it relies on

is the focus of our work.

respondents remembering the time of evacuation.

The second measure is derived from traffic sensor data from the Florida Department of Transportation. Using the Regional Integrated Transportation Information System (www.ritis.org), we collect data from two highway sensors: I-75 north bound (detector id-9828), and I-95 north bound (detector id-10077). These detectors, just north of Orlando, cover the two primary interstate routes for evacuees leaving southern Florida, reporting estimates of the number of vehicles that pass the detector in 15-minute intervals. Similar data has been used to reconstruct and forecast evacuation patterns during Hurricane Irma (Feng and Lin 2021; Ghorbanzadeh et al. 2021; Roy et al. 2021).

Figure 1 plots the time series for each of these measures, along with the Twitter-derived measures. Overall, there is moderate agreement between the time series – Pearson's correlation is .609 between traffic and Twitter data, and .604 between survey and Twitter data.[3] By comparison, the corresponding correlations using all tweets that match the regular expressions, without using the classifier for filtering, is .345 with traffic and .412 with surveys. One noticeable deviation is that the highest peak of tweets occurs late on September 5th, even though the survey indicates that September 8th was the most common day for evacuation. This is expected, since the Twitter series is a leading indicator of behavior – i.e., a user posts "we're evacuating tomorrow" the day before they actually evacuate. Hurricane Irma, in particular, had a long lead time, with a state of emergency being declared over five days before landfall (in contrast with Hurricane Michael in 2018, in which a state of emergency was only declared two days before landfall). Evacuation orders were issued on September 6th for low lying areas, then expanded to additional areas on the 7th and 8th. More work is needed to extract time expressions from the posts (e.g., "leaving tomorrow" vs. "leaving today") to examine these temporal trends more closely. Furthermore, these correlations should be seen only as offering some face validity to the Twitter trends — more rigorous analysis would be needed to account for endogenous factors that may influence these results.

Nevertheless, the presence of so many "intent to evacuate" tweets so long before landfall suggests that many people make evacuation decisions at an early stage. During an unfolding hurricane, being able to identify such tweets and analyze user characteristics may help emergency managers gauge levels of evacuation compliance in real-time, providing enough lead time to intervene as necessary to better target public messaging. This motivates the following section, in which we analyze the demographics of users expressing pro- or anti-evacuation intent.

[3]Each time series was aligned to 3 hour intervals to match the survey data frequency. Correlation was also computed after removing daily trends using the `seasonal_decompose` method of the `statsmodels` Python library; correlations were lower but comparable (.609 → .577 for traffic and .604 → .592 for survey.)

Characteristic	Category	n (%)
Age	18-34	2,586 (52%)
	35+	2,404 (48%)
Children	Has Kids	2,250 (45%)
	No Kids	2,740 (55%)
Education	No College	2,445 (49%)
	Some College	1,223 (25%)
	Graduate	1,322 (26%)
Race/Ethnicity	Asian	375 (8%)
	Black	1,609 (32%)
	Hispanic	1,419 (28%)
	White	1,587 (32%)
Gender	Female	3,118 (62%)
	Male	1,872 (38%)
Income	$0-100k	3,171 (64%)
	$100k+	1,819 (36%)

Table 6: Inferred Twitter user demographics

3.4 Demographic Correlates of Evacuation Decisions

Our third research question asks *What factors contribute to evacuation decisions?* Inequities in who chooses to or is able to evacuate is a major area of interest in emergency management (Elder et al. 2007; Deng et al. 2021). Understanding how these decisions vary by group allow practitioners to target messaging to priority groups, to identify evacuation barriers to be removed, and to track disparities over subsequent hurricanes. In the following sections, we analyze both user demographics as well as tweet contents to better understand factors that correlate with evacuating versus not.

To investigate how user demographics correlate with evacuation decisions, we first identify the 3,514 users who tweeted at least one message classified as positive, and the 1,476 users who tweeted at least one message classified as negative (a small number of users were removed who tweeted one of each). To estimate user demographics, we use a classifier from prior work (Culotta, Ravi, and Cutler 2016), which was trained to predict user demographics based on a sample of their tweets. Table 6 lists the inferred demographic distribution of users in this sample. While undoubtedly such a classifier is imperfect (and oversimplifies gender and race/ethnicity categories), prior results indicate that it exhibits high concordance both with labeled data (.83-.86 F1) and with population level characteristics (e.g., .73 average correlation with panels of matched web traffic demographics). Compared to population statistics, the sample appears to over-sample users who are female, young, and high income. Compared to the retrospective survey (Wong, Shaheen, and Walker 2018), this sample has greater representation of people of color (e.g., 32% Black vs. 1.6%) and male users (38% vs. 18.1%).

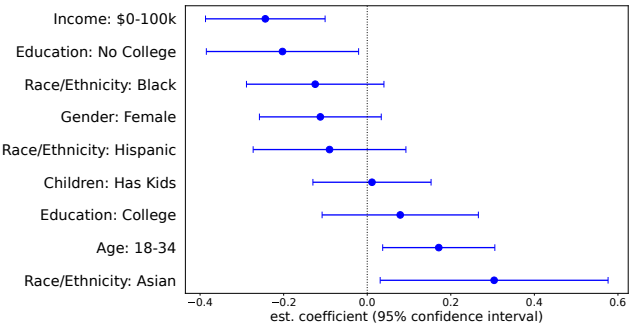


Figure 2: Coefficients (and 95% confidence intervals) of a binary logit model predicting whether a user posts a tweet expressing an intent to evacuate (1) or not evacuate (0) based on inferred user demographics.

Given the predicted demographic categories as binary independent variables, we next fit a binary logit model to predict whether a user authored a tweet to evacuate (1) or not to evacuate (0) based on both the human-annotated data and the classifier predictions. Figure 2 plots the estimated coefficients (and 95% confidence intervals). For each demographic characteristic, we omit one dummy variable to avoid collinearity (34+, No Kids, Grad School, White, Male, $100k+). Overall, the data suggest that non-evacuees tend to be older, less wealthy, less educated, and less white. These results mostly agree with the survey described above (Wong, Shaheen, and Walker 2018) with respect to age, ethnicity, and parental status. However, results based on income and education levels were not available from the survey.

While other surveys have also investigated the demographic correlates of evacuation decisions, a potential advantage of using social media is again the real-time nature, which allows managers to identify specific groups who are not evacuating for the present hurricane. Furthermore, a content analysis (§3.6) can provide rationales for evacuation decisions, stratified by demographics.

Since demographic classification from tweets is a noisy process, to provide additional validation we randomly sampled 100 users and manually annotated them with respect to age, gender, and race/ethnicity based on an inspection of their user profile. As this manual annotation is itself an admittedly difficult task, we selected 100 users for which we had high confidence in the annotations based on the available information (and adjusting age estimates based on the time frame of Hurricane Irma). From this sample, we find the classifier to have 96% accuracy for age (18-34 vs. 35+), 86% accuracy for gender (male vs. female), and 76% accuracy for race/ethnicity (Asian vs. Black vs. Hispanic vs. White). The gender results appear in line with the results reported in the original paper (Culotta, Ravi, and Cutler 2016) (84% F1 when using only text features, as we use here), though the race/ethnicity results are a bit lower (86% F1). It is possible that focusing on a geographically homogeneous set of users will have a different error profile than applying to the classifier to a wider sample of users. We recommend as future work domain adaptation methods to fine-tune demographic classifiers on a specific sample of data.

3.5 Evacuation Time by Demographics

Since we know the time of the first positive tweet of each evacuating user, we can additionally plot the cumulative number of users expressing an intent to evacuate over time for each demographic category (Figure 3). Such plots provide insights into who the "early adopters" are with respect to expressing an intent to evacuate. Two of the most notable differences occur in Age and Race/Ethnicity, indicating that younger users and non-white users tend to post evacuation messages sooner than others.

3.6 Clustering

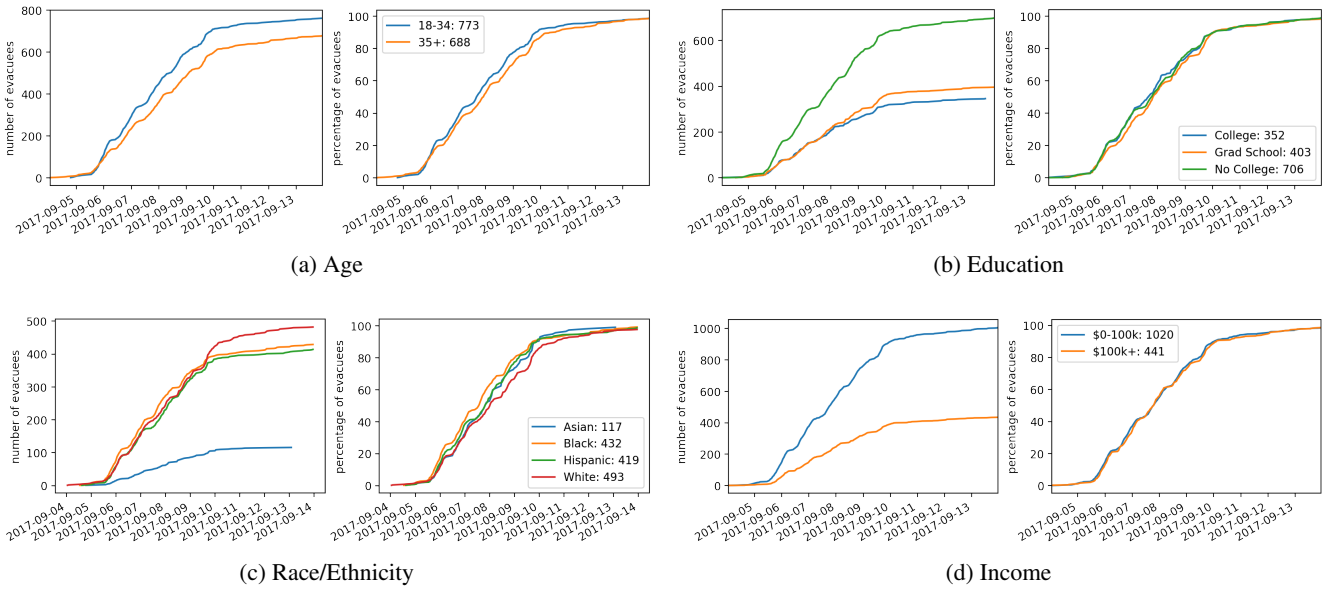
To further investigate the rationales of people's evacuation decisions, we performed a clustering analysis on tweets from the negative and positive classes. Understanding these rationales can be helpful both as a retrospective analysis (why did people behave as they did?), as well as for real-time analysis. For example, if people are using incorrect information to rationalize their decision making (e.g., thinking a road or shelter is closed when it is in fact open; being unaware of options for evacuating with pets), emergency managers can address these issues directly.

We use a simple k-means approach to cluster the tweets. Each tweet is represented by a vector of term frequencies and normalized with L2 norm. We perform the tokenizing process similar to that in the classification, additionally removing stop words and very common terms.

We first perform the clustering on the 1,267 negative tweets with the number of clusters set to 8[4]. Table 7 shows the clustering result including the size, top terms and a centered instance of each cluster (reworded slightly for anonymity). We derive the top terms according to the cluster centers, where a large value in coordinates indicates a high frequency of the corresponding term in that cluster. The most centered sample in a cluster is the one closest to the center point. From the clusters, we can notice some factors causing people's non-evacuation decisions, including families, pets, shortage of gasoline, and that they are not in an evacuation or flood zone. The survey of Wong, Shaheen, and Walker (2018) also asks respondents reasons for not evacuating. The most popular answers include "Didn't want to sit in traffic," "Didn't want to leave," and "Wanted to protect my property." Work requirements and pets account for 22% and 18% of respondents, respectively. Interestingly, family members did not come up in the survey results. This may be in part because the survey skewed toward older and wealthier individuals and heads of households, whereas in the Twitter data such tweets often come from users who are staying with their parents.

We also cluster the 1,727 positive tweets using the same approach to separate them into 8 groups. Table 8 shows the 8 clusters produced by the model. Families are also frequently

[4]We did not tune this parameter extensively, but given the many short and similar messages, we tried several small values of k (5-20), and selected 8 based on a subjective analysis of the clusters.



(a) Age

(b) Education

(c) Race/Ethnicity

(d) Income

Figure 3: Cumulative number of users expressing positive evacuation intent over time by demographic characteristics: Age, Ethnicity, Education, Income. For each characteristic, the left figure shows the raw number of users while the right one shows the cumulative percentages over time.

cluster	size	top terms	centered sample
1	711	work, dad, late	I'm a nurse in Fl. Can't evacuate. Told 2 bring supplies 2 work and be ready 2 be at work till Tue at least. Prayers welcome #IrmaHurricane
2	134	family, friends, prepared	But i cant leave my family and my puppies :(
3	71	pets, family, family pets	yep us too! I would never leave my dogs behind
4	63	mom, dad, work	No smoke no smoke I won't go won't go can't leave my momma alone
5	61	gas, family, stations	Not evacuating. Not enough gas to get out of state. Stressing out
6	35	evacuation, zone, evacuation zone	we did not evacuate and will not because we are not in an evacuation zone.
7	33	parents, coming, right	I can't leave. My parents won't evacuate, so I'm staying with them.
8	17	zone, flood, flood zone	Hunkering down is not easy with 5 huge dogs! We are not in a flood zone so we did not evacuate. Everyone be safe.

Table 7: Clusters of tweets showing negative evacuation intent (with minor edits for anonymity)

mentioned in positive tweets and such tweets form a cluster as a result. The remaining clusters are mostly related to the approximate time of people's evacuation, e.g., now, tomorrow, tonight. The place which people evacuate to or from also serves as a significant feature in some of the clusters, e.g., Miami, Georgia. These results suggest avenues for future research to extract more granular time and location information from such tweets to perform a more fine-grained analysis of evacuation behavior. Additionally, more advanced clustering methods and word representations may improve the clustering results.

4 Conclusions, Limitations, Future Work

In this paper, we have presented a method to identify tweets expressing evacuation behavior and have analyzed temporal, demographic, and linguistic patterns to understand factors that contribute to such decisions. In many cases, our findings agree with those from traditional surveys, but we also observe novel factors such as the influence of family members on evacuation decisions and the correlations with education and income levels. We also find that tweets expressing an intent to evacuate peak well before other measures, suggesting that they may serve as a leading indicator of evacuation behavior.

As with most studies based on social media, there are several important limitations of this work. First, the Twitter users are not a representative sample of the general population. By inferring the demographics of our sample, we can stratify our results and better understand the nature of this bias. Our initial keyword selection was limited to English, so we are likely undersampling non-English speakers, although the demographic analysis indicates a sizable

cluster	size	top terms	centered sample
1	934	morning, friday, heading	Yeah, we're leaving on Friday morning. We just got a hotel!
2	235	tomorrow, morning, tomorrow morning	Thank you. Sincerely. Evacuating tomorrow.
3	132	family, friends, north	I'm evacuating with my family regardless.
4	99	now, fine, live	Evacuating now.
5	78	tonight, tomorrow, miami	I'm leaving tonight
6	73	today, tomorrow, live	Thanks. Evacuating today.
7	70	miami, now, tomorrow	im evacuating out of Miami
8	56	georgia, now, family	No, we're evacuating up to Georgia. But yes, we normally do that. 30 yrs in Florida, you learn.

Table 8: Clusters of tweets showing positive evacuation intent (with minor edits for anonymity)

sample of Hispanic users. Furthermore, as the classifier was trained on data sampled in part via select keywords, we have not assessed how the classifier would perform on all Twitter messages. We instead recommend using the keyword filter as a first step, after which the classifier is applied. Additionally, our work relies on multiple imperfect classifiers to infer evacuation intent and demographics, which can introduce additional bias. Finally, our work is based on self-reported plans to evacuate; it is possible that users who express an intent to evacuate end up not doing so, and visa versa. To partly address these concerns, we have attempted to validate our results both with traditional survey data and traffic volume measures.

In future work, it will be useful to test the ability of our approach to generalize to different hurricanes. Hurricane Irma struck a heavily populated area, and thus had a large evacuation. Events with smaller evacuations, and in less densely populated areas, may not have sufficient Twitter data to perform a similar analysis. Drawing upon additional data sources (e.g., Instagram, Reddit, Google Search Trends, etc.) may help in such cases. Additionally, domain adaptation approaches should be investigated to ensure the classification models can generalize to new locations, which may have different linguistic and demographic characteristics.

An additional area of future work is to determine how predictive the initial spike of evacuation-related tweets is of overall evacuation behavior. For example, the initial online reaction to a declared state of emergency may give planners guidance as to what level of evacuation compliance to expect and to adjust accordingly.

Ethical Statement

A risk of data and analysis of this paper is the possible harm from inferring demographic information of online users. Such inferences could be used for unintended purposes, and any biases in the classifiers could result in disparate impacts. Any application of the ideas presented here must carefully weigh these potential risks with any potential benefit to emergency planning. To allow replication while complying with terms of service and privacy concerns, only the Twitter IDs of tweets included in this study will be released.

Acknowledgements

This research is supported in part by the National Science Foundation under grant #2133960.

References

Castillo, C.; Diaz, F.; Lin, Y.-R.; and Yin, J. 2016. The Fourth International Workshop on Social Web for Disaster Management (SWDM 2016). In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2503–2504.

Culotta, A.; Ravi, N. K.; and Cutler, J. 2016. Predicting Twitter user demographics using distant supervision from website traffic data. *Journal of Artificial Intelligence Research* 55: 389–408.

Deng, H.; Aldrich, D. P.; Danziger, M. M.; Gao, J.; Phillips, N. E.; Cornelius, S. P.; and Wang, Q. R. 2021. High-resolution human mobility data reveal race and wealth disparities in disaster evacuation patterns. *Humanities and Social Sciences Communications* 8(1): 1–8.

Dow, K.; and Cutter, S. L. 2002. Emerging hurricane evacuation issues: hurricane Floyd and South Carolina. *Natural hazards review* 3(1): 12–18.

Elder, K.; Xirasagar, S.; Miller, N.; Bowen, S. A.; Glover, S.; and Piper, C. 2007. African Americans' decisions not to evacuate New Orleans before Hurricane Katrina: A qualitative study. *American Journal of Public Health* 97(Supplement_1): S124–S129.

Feng, K.; and Lin, N. 2021. Reconstructing and analyzing the traffic flow during evacuation in Hurricane Irma (2017). *Transportation Research Part D: Transport and Environment* 94: 102788.

Ghorbanzadeh, M.; Burns, S.; Rugminiamma, L. V. N.; Erman Ozguven, E.; and Huang, W. 2021. Spatiotemporal Analysis of Highway Traffic Patterns in Hurricane Irma Evacuation. *Transportation Research Record* 03611981211001870.

Hong, L.; and Frias-Martinez, V. 2020. Modeling and predicting evacuation flows during hurricane Irma. *EPJ Data Science* 9(1): 29.

Huang, S.-K.; Lindell, M. K.; and Prater, C. S. 2016. Who leaves and who stays? A review and statistical meta-analysis of hurricane evacuation studies. *Environment and Behavior* 48(8): 991–1029.

Huang, S.-K.; Lindell, M. K.; and Prater, C. S. 2017. Multistage model of hurricane evacuation decision: Empirical study of Hurricanes Katrina and Rita. *Natural Hazards Review* 18(3): 05016008.

Johnson, T. P.; and Wislar, J. S. 2012. Response rates and nonresponse errors in surveys. *Jama* 307(17): 1805–1806.

Kumar, D.; and Ukkusuri, S. V. 2018. Utilizing geo-tagged tweets to understand evacuation dynamics during emergencies: A case study of Hurricane Sandy. In *Companion Proceedings of the The Web Conference 2018*, 1613–1620.

Li, T.; Zhou, W.; Zeng, C.; Wang, Q.; Zhou, Q.; Wang, D.; Xu, J.; Huang, Y.; Wang, W.; Zhang, M.; et al. 2016. DIDAP: an efficient disaster information delivery and analysis platform in disaster management. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, 1593–1602.

Martín, Y.; Cutter, S. L.; and Li, Z. 2020. Bridging Twitter and survey data for evacuation assessment of Hurricane Matthew and Hurricane Irma. *Natural hazards review* 21(2): 04020003.

Martín, Y.; Li, Z.; and Cutter, S. L. 2017. Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew. *PLoS one* 12(7): e0181701.

Metaxa-Kakavouli, D.; Maas, P.; and Aldrich, D. P. 2018. How social ties influence hurricane evacuation behavior. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 1–16.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. URL http://www.aclweb.org/anthology/D14-1162.

Que, R.; and Dredze, M. 2017. Carmen: a library for geolocating tweets. https://github.com/mdredze/carmen-python. Accessed: 2021-05-01.

Roy, K. C.; and Hasan, S. 2021. Modeling the dynamics of hurricane evacuation decisions from twitter data: an input output hidden markov modeling approach. *Transportation research part C: emerging technologies* 123: 102976.

Roy, K. C.; Hasan, S.; Culotta, A.; and Eluru, N. 2021. Predicting traffic demand during hurricane evacuation using Real-time data from transportation systems and social media. *Transportation Research Part C: Emerging Technologies* 131: 103339. ISSN 0968-090X. doi:https://doi.org/10.1016/j.trc.2021.103339.

Rudra, K.; Ghosh, S.; Ganguly, N.; Goyal, P.; and Ghosh, S. 2015. Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 583–592.

Stowe, K.; Anderson, J.; Palmer, M.; Palen, L.; and Anderson, K. M. 2018. Improving classification of twitter behavior during hurricane events. In *Proceedings of the sixth international workshop on natural language processing for social media*, 67–75.

Tasse, D.; Liu, Z.; Sciuto, A.; and Hong, J. 2017. State of the geotags: Motivations and recent changes. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Twitter. 2021. Search the full archive of Tweets. https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all. Accessed: 2021-05-25.

Wen, X.; Lin, Y.-R.; and Pelechrinis, K. 2016. Pairfac: Event analytics through discriminant tensor factorization. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 519–528.

Wong, S.; Shaheen, S.; and Walker, J. 2018. Understanding evacuee behavior: a case study of Hurricane Irma. UC Berkeley: Transportation Sustainability Research Center. URL https://escholarship.org/uc/item/9370z127.

Wong, S. D.; Pel, A. J.; Shaheen, S. A.; and Chorus, C. G. 2020. Fleeing from hurricane Irma: Empirical analysis of evacuation behavior using discrete choice theory. *Transportation Research Part D: Transport and Environment* 79: 102227.

Yang, H.; Morgul, E. F.; Ozbay, K.; and Xie, K. 2016. Modeling evacuation behavior under hurricane conditions. *Transportation research record* 2599(1): 63–69.

Yin, J.; Karimi, S.; Robinson, B.; and Cameron, M. 2012. ESA: emergency situation awareness via microbloggers. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2701–2703.

Zhou, C.; Sun, C.; Liu, Z.; and Lau, F. 2015. A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630* .