Jonathan Rusert, Osama Khalid, Dat Hong, Zubair Shafiq, and Padmini Srinivasan

No Place to Hide: Inadvertent Location Privacy Leaks on Twitter

Abstract: There is a natural tension between the desire to share information and keep sensitive information private on online social media. Privacy seeking social media users may seek to keep their location private by avoiding the mentions of location revealing words such as points of interest (POIs), believing this to be enough. In this paper, we show that it is possible to uncover the location of a social media user's post even when it is not geotagged and does not contain any POI information. Our proposed approach JASOOS achieves this by exploiting the shared vocabulary between users who reveal their location and those who do not. To this end, JASOOS uses a variant of the Naive Bayes algorithm to identify location revealing words or hashtags based on both temporal and atemporal perspectives. Our evaluation using tweets collected from four different states in the United States shows that JASOOS can accurately infer the locations of close to half a million tweets corresponding to more than 20,000 distinct users (i.e., more than 50% of the test users) from the four states. Our work demonstrates that location privacy leaks do occur despite due precautions by a privacy conscious user. We design and evaluate countermeasures based JASOOS to mitigate location privacy leaks.

DOI 10.2478/popets-2019-0064
Received 2019-02-28; revised 2019-06-15; accepted 2019-06-16.

1 Introduction

Background. About seven-in-ten Americans report using social media platforms such as Facebook and Twitter [3]; however, they are increasingly wary about the privacy risks that come with their use. According to a recent survey by the Pew Research Center, about 80% and 71% American social media users are concerned about third-party companies and governments accessing their data on social media platforms, respectively [44]. For example, health care providers such as United-Health Group reportedly mine social media data along with other clinical information to assess health care risks and insurance premiums [5]. Businesses are also increasingly using social media to screen candidates before hiring [35]. Government agencies in the U.S. such as the FBI, DHS, and ICE are now surveilling content on social media platforms [2, 17, 36].

Privacy losses. A user's privacy is obviously at risk when their social media posts explicitly reveal some private information. For example, Mao et al. showed that social media users often include phrases such as "flying to" and "have cancer" [33]. Such posts can reveal users' location and medical conditions with high precision. Hecht et al. reported that approximately two-thirds of Twitter users provide valid location information (albeit at varying granularity) in their profile descriptions [22]. Such explicit revelations may either be because of a user's "carelessness" or "ignorance" in not knowing how to keep the information private.

Beyond explicit revelations, user privacy on social media platforms is also threatened by recent advances in statistical and machine learning techniques capable of inferring sensitive user attributes even when they are not explicitly revealed. On Facebook, a user's likes can be used to infer gender, ethnicity, relationship status, sexual orientation, religious views, location, political affiliation, use of addictive substances, and other private attributes [9, 28]. Similarly, on Twitter, a user's posts can be used to infer gender, ethnicity, age, political affiliation, and location [32, 50]. The recent Cambridge Analytica scandal [13] has further exacerbated privacy concerns about large-scale user profiling efforts by prying third-parties [7].

Privacy losses despite being cautious! Beyond the privacy losses described above, there is the uncharted problem of *inadvertent* leakage of information even by privacy seeking users on social media platforms. We suggest that such leakages can happen. Specifically, we examine whether a Twitter user's location can be revealed by a tweet *even when they take reasonable precautions*

Jonathan Rusert: University of Iowa, Email: jonathan-rusert@uiowa.edu
Osama Khalid: University of Iowa, Email: osama-khalid@uiowa.edu
Dat Hong: University of Iowa, Email: dat-hong@uiowa.edu
Zubair Shafiq: University of Iowa, Email: zubair-shafiq@uiowa.edu
Padmini Srinivasan: University of Iowa, Email: padmini-srinivasan@uiowa.edu

to keep location private. We assume that the users turn off location services on their devices, thus the tweets are not geotagged. We also assume that they are aware of tools that use geolocation databases (i.e., gazetteers) to infer location and so avoid explicit mentions of city names or points of interest (POIs) in the tweets and profile information. Moreover, since a user's social graph information (followers/followings) also can be revealing of location [46], we assume they protect this information as well. The question we ask is: can the location of a privacy seeking user taking due precautions still be leaked by just their tweet text? The answer we present in this paper is that it is indeed possible to identify locations for many tweets, even when they do not contain any obvious gazetteer[1] words.

Key Insight. Our key insight is that certain words, which may seem location neutral, can become location revealing due to their usage patterns. We discover such words by analyzing usage patterns amongst users who openly reveal their location (say through geotagging their tweets). We find that certain words (e.g. #isf2017) may be location revealing only during a short time interval while others (e.g. badgers) may continue to be location revealing over a long time interval. #isf2017 becomes location revealing in the month of August when the Iowa State Fair is held in the city of Des Moines. badgers is persistently location revealing as it is the name of a college sports team in Madison, WI. Neither term is a gazetteer term. Using such location revealing words, which are discovered from usage patterns in tweets of users who openly disclose their location, we can infer the locations of tweets by privacy seeking users who do not reveal their location. In essence, when vocabulary is shared between users who reveal location and users who do not then it raises the potential of location privacy leaks for the latter.

Proposed approach. We operationalize this insight by developing a Naive Bayes based location inference approach, named Jasoos[2] and testing it under non-gazetteer conditions. We intentionally pick a well established probabilistic framework (in contrast to selecting a leading location detection algorithm) in order to show how even a standard approach can threaten privacy. Jasoos adopts an integrated temporal and atemporal perspective. It also utilizes a *maxwordNB* variant of the standard Naive Bayes algorithm which essentially considers a single best feature (nouns and hashtags in the tweet text) to identify location. Our evaluation shows that Jasoos can accurately infer the locations of close to half of a million tweets collected from four different states in the United States.

Potential Countermeasures. The unpredictable nature and dynamics of vocabulary sharing between location revealing and location private users makes it challenging for privacy seeking social media users to anticipate words responsible for leaking their location. Therefore, to counter this location inference attack, Jasoos can be leveraged to develop a warning system that would ingest geotagged tweets to inform users about the usage of potential location revealing words in their tweets.

Key contributions. We summarize our key contributions as follows.

1. **Novel threat model:** The problem of inferring locations of tweets under the strict condition of explicitly excluding gazetteer words and any profile and social graph information, has not been studied in prior literature. This threat model is applicable for privacy seeking users who take due precautions to keep their location private. This core aspect of our research is novel.
2. **Extensive evaluation**: We present evaluations of Jasoos with tweets from four different states in the USA. We explore different feature sets and find that the combination of hashtags and nouns extracted from the tweets are the best. We also show that our maxword approach is more effective than a standard Naive Bayes algorithm.
3. **Strong performance in rank error:** We show that Jasoos is able to accurately infer locations of close to half a million tweets even when they do not include any obvious location revealing gazetteer words. As a highlight of our results, when using a strict version of Jasoos (explained later) we identify the correct city perfectly (rank error of 0) for 187,457 tweets from 277 different cities in Colorado, 36,276 tweets from 416 different cities in Iowa, 16,794 tweets from 548 cities in Wisconsin, and 205,454 tweets from 274 cities in Oregon. The median rank error is 3 for Colorado and Oregon while it is 8 for Iowa and 16 for Wisconsin which has the largest number of cities.
4. **Strong performance in coverage:** The coverage of our approach, i.e., percentage of tweets without

1 A gazetteer is a database of place names along with their geographic coordinates. Gazetteers typically include points of interest (POIs) such as names of popular landmarks.

2 Jasoos means spy in several languages including Urdu and Hindi.

gazetteer words for which location is inferred is good to excellent depending on the state. Our approach covers the majority of tweets, 75% for tweets in Colorado, 42% for tweets in Iowa, 50% for tweets in Wisconsin, and 62% for tweets in Oregon. These correspond to 439,637 located tweets for Colorado, 123,641 for Iowa, 98,790 for Wisconsin and 552,149 located tweets for Oregon.

5. **Design and evaluation of countermeasures:** To help counter this location inference attack, we design and evaluate several countermeasures based on JASOOS . The results show that these countermeasures can effectively degrade the accuracy of the location inference attack.

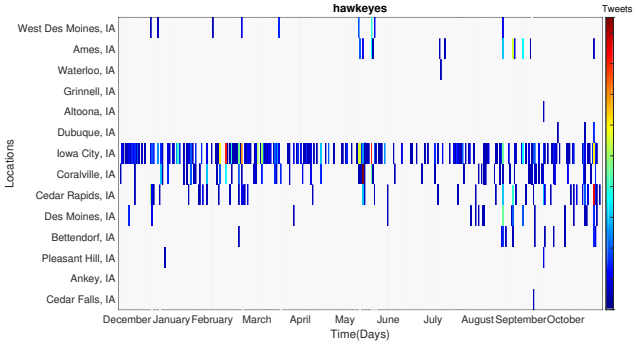
The rest of the paper is structured as follows. In the next section we formulate the problem and illustrate our key insight. Following this in section 3 we detail our proposed approach. Our experiments and results are in section 4, related research in section 5 followed by the last section presenting concluding remarks.

2 Problem Formulation

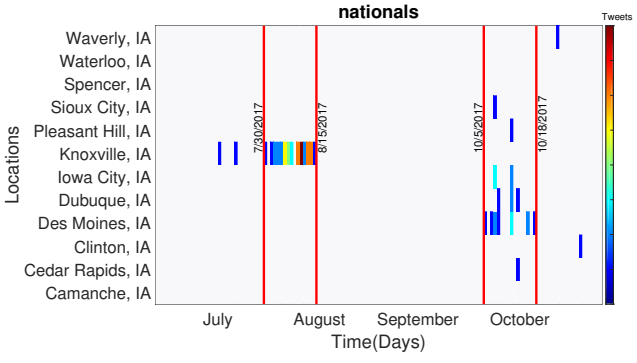
Threat Model. We present the location inference problem as modeling an attack where an adversary has collected a dataset of social media posts over a time period and is interested in inferring the location of as many posts as possible. We assume that there are social media users who are privacy conscious and take due precautions. Thus, for posts from such users, the adversary cannot simply look for gazetteer words such as location names and points of interest (POIs). The adversary also will not have access to location hints in meta-data such as the user's profile (e.g., home location, language) and social network (e.g., friends, followers, likes). Thus, we assume a strong threat model where the adversary only has access to a post's text. We do assume that the adversary's dataset includes geotagged posts, including those that contain gazetteer words, from some privacy neutral users. The goal for the adversary is to infer the location of posts of privacy conscious users that are neither geotagged nor do they contain gazetteer words. To the best of our knowledge, this strong threat model targeting privacy conscious social media users has not been investigated in prior literature.

Problem Statement. In this paper, we focus our attention on Twitter because it has public APIs to facilitate large-scale collection of social media posts of which a sample are geotagged. There are two flavors of location inference problems studied using Twitter: inferring a user's home location and inferring tweet location. The latter, which is more challenging, is our focus here and we aim to achieve this using only tweet text without the assistance of gazetteers. Formally, our goal is to predict the location l of a tweet where $l \in L = \{l_1, l_2 ... l_k\}$ using just the tweet text. Let $T = \{t_1, t_2 ... t_m\}$ be a dataset of tweets that do not contain gazetteer words, where $t \in T$ is posted at time θ_t. We estimate l from the tweet t's feature set $\{f_1, f_2 ... f_n\}$ extracted from its text. We consider the location inference problem as a ranking problem where our goal is to rank locations in L (defined at the city-level) based on the likelihood of being the correct city.

Illustration of our Insight. Our key insight is that non-gazetteer words appearing in a tweet can acquire location revealing properties depending on their usage pattern. We show two examples based on the Iowa dataset described in section 4.1. For example, Figure 1a shows the spatio-temporal distribution of the word "hawkeyes," which refers to the University of Iowa's athletics based in Iowa City. Note that the word "hawkeyes"



(a) "hawkeyes"



(b) "nationals"

Fig. 1. Spatio-temporal distributions of two location-revealing words

is location-revealing because the tweets which contain it are spatially concentrated near Iowa City (including nearby cities such as Cedar Rapids and Coralville). Figure 1b shows the spatio-temporal distribution of the word "nationals." Between July and August, these tweets tend to be concentrated in Knoxville (because of a racing event held at Knoxville July 30 to August 15). Whereas in October these tweets are spread across multiple cities because they refer to a popular, non-local baseball series played between the Washington Nationals and the Chicago Cubs. This demonstrates that some words become location-revealing only at certain points in time because of time-bound local events. As we discuss next, our system JASOOS exploits both temporal and atemporal location-revealing words.

3 Proposed Approach

The machine learning algorithm underlying JASOOS is a standard Naive Bayes which estimates a likelihood for each location given the features of a test tweet. We intentionally employ a standard algorithm, as our focus is more on the threat model. The probabilistic framework underlies many approaches such as in Hulden et al. [24] and Hahmann et al.[20] However, there are several novel aspects in our approach. First, we use Naive Bayes in two flavors: an atemporal model and a temporal model (run at an interval of a day), and make decisions only on tweets for which the two models agree. A key point to note is that temporal approaches for location estimation are themselves rare ([39, 49] are some exceptions). We combine models in order to raise the confidence in our location estimates. For example, in the case of "nationals" in Figure 1b, agreement between the temporal and atemporal models is less likely in the October time period than in the July-August time period. The second unique angle in our approach is that we use a novel *maxwordNB* variant of Naive Bayes. This captures the idea (and our experiments provide support) that exploiting a single location revealing feature is more effective than considering multiple/all features as in the Naive Bayes algorithm. Another novel aspect is that, given our emphasis on privacy leakages, we assess our Naive Bayes based approach under the strict condition of excluding test tweets with gazetteer words.

Next, we briefly describe the Naive Bayes algorithm, our *maxwordNB* variant, the temporal and atemporal models, and the specifics of how we combine them to make location estimations using tweet text.

Naive Bayesian Estimations. We predict location at a city level granularity, but our formulation holds for location prediction at varying granularity. Formally, given a tweet t with feature set $\{f_1, f_2...f_n\}$, we want to calculate $argmax_l P(l|t)$ with $l \in L$. Using Bayes theorem we have:

$$P(l|t) = \frac{P(t|l)}{P(t)} P(l) \tag{1}$$

Assuming independence between tweet features we have,

$$P(t) = P(f_1)P(f_2)...P(f_n) \tag{2}$$

and

$$P(l|t) = \frac{P(f_1|l)}{P(f_1)} \frac{P(f_2|l)}{P(f_2)} ... \frac{P(f_n|l)}{P(f_n)} P(l) \tag{3}$$

where:

$$P(f_i|l) = \frac{|f_i^l|+1}{\sum_{j=1}^{n} |f_j^l| + |L|}$$

Here, $P(l|t)$ is the likelihood of a location l given a tweet t. The likelihood is normalized to get the probability of a location l given the tweet t, $|f_i^l|$ is the number of times f_i appears in tweets originating from location l in the training data while $\sum_{j=1}^{n} |f_j^l|$ is the total number of occurrences of all features appearing in l. We use Laplacian smoothing as shown above.

MaxwordNB Variant. We use a variant of Naive Bayes that we call *maxword Naive Bayes*, or *maxwordNB* for short. For this variant the $\frac{P(t|l)}{P(t)}$ component of equation 1 is redefined as:

$$\frac{P(t|l)}{P(t)} = max\left(\frac{P(f_1|l)}{P(f_1)}, \frac{P(f_2|l)}{P(f_2)}, ... \frac{P(f_n|l)}{P(f_n)}\right)$$

Using this we can modify equation 3 as:

$$P(l|t) = max\left(\frac{P(f_1|l)}{P(f_1)}, \frac{P(f_2|l)}{P(f_2)}, ... \frac{P(f_n|l)}{P(f_n)}\right) P(l) \tag{4}$$

By doing so, we make our location decision for t using a single, most indicative, constituent feature $argmax_f \frac{P(f_n|l)}{P(f_n)}$.

The intuition here is that a single feature with strong location revealing properties should not be drowned out by other less location specific ones. For example, consider the tweet *'Just parked the car. With family seeing the Hawkeyes win at football'*. Assume that the correct location is *Iowa City* and that $P(Iowa\ City|Hawkeyes)$ has the highest probability value. If, instead of *maxwordNB*, we consider all features

as is standard in Naive Bayes then, because of the multiplication with smaller probabilities, the final estimate for Iowa City will be lowered. This raises the risk of the estimates favoring a different city.

Atemporal and Temporal Models. The key difference is whether a tweet timestamp is considered when defining training data. The atemporal model captures latent spatial features of words that hold independent of time. Here the training data is split into N folds by user and not by time. That is, all tweets posted by a user are placed in the same fold so as to avoid any contamination. As is standard practice, all but one of the folds are combined to form training data in each cross validation run. The remaining fold is used as test data. By swapping folds around, each tweet is included in a test set once.

The temporal model captures time specific, latent spatial features of words. In this model we use a shifting window (of one day). The exact same user based N folds created for the atemporal model are used here, additionally sliced temporally into day chunks. The same cross validation strategy is used with each day's data. Since this is retrospective analysis (as opposed to online), tweets posted later than a test tweet t (but in the same day) may be included in t's training data. This design is deliberate since the discussion of the topic in t may continue beyond its timestamp. To clarify, if a tweet has a timestamp at 2016-12-12 12:00:00, training data can be obtained from 2016-12-12 00:00:00 (12 hours before) to 2017-12-13 00:00:00 (12 hours after). It should be noted that the timestamp is not used as a feature, only to select appropriate training data.

Decision Strategy. Given a test tweet t, both the temporal and atemporal models independently rank all locations in L using features of t. Our strategy for making decisions is to do so *only* for those tweets where both models agree. Otherwise a location decision is not made for that tweet. Thus coverage, or the number of decisions made, depends on agreement between the atemporal and temporal models. This strategy limits location decisions to those that can be made confidently using a combination of atemporal and temporal perspectives. The same principle can be extended if additional location inference perspectives (such as from metadata and social networks) were to be later introduced into the location detection system.

If each model identifies a single location for a tweet then agreement is simple to define. However, given that each model outputs a ranking of a set of locations, agreement is a more complex notion. Specifically, we define agreement as dependent on a rank cut off R. Given two rankings of locations for tweet t, $Agreement_{@R}$ holds if there is at least one location in common when the two rankings are limited to the top ranked R locations. In the trivial case with $R = 1$, the two ranks agree if they have the same location ranked at the top. We can expect that the number of tweets with location decisions, i.e., coverage, increases with increasing cut off R. For tweets where there is agreement, we combine the temporal and atemporal rankings by calculating an average rank for each location. Locations are re-ranked by this average.

Features. We use the two main features found in tweets: hashtags and nouns. Hashtags with their specific semantics reduce the ambiguity challenge. The downside is their sparseness, which can lead to insufficient training data and low coverage. Nouns are more prevalent and so the problem of insufficient training data is possibly avoided. However, nouns may be more ambiguous relying on context for interpretation. As compared to other parts-of-speech, we choose nouns as they are more content bearing, i.e., informative. We also investigate hashtags combined with nouns in order to jointly leverage their individual merits.

Exclusion of Gazetteer Features. We remove tweets with gazetteer words from the test set. However, we do keep such tweets in training data as we can use the probability estimates for their non-gazetteer words. For example, the tweet *Chicago Gospel Music Festival was a grand event* will never be a test tweet because it has *Chicago*. However, we can use it to increase robustness of our estimates for its other words such as 'Gospel', 'grand' and 'event'. This design also reflects the real world where the adversary has access to tweets from both privacy seeking and privacy neutral individuals. To reiterate, all results reported are on tweets that do not have gazetteer words which is consistent with our aim to explore inadvertent privacy leakages. The gazetteer we use is a well maintained one provided by the U.S. government [1]. It has explicit city names as well as POIs such as names of parks, cemeteries, bridges, schools, streams, creeks, churches, post offices, hospitals, libraries, farms, etc. We also exclude hashtags that are made simply by combining location names such as *#IowaCity*.

State	Iowa (IA)	Colorado (CO)	Wisconsin (WI)	Oregon (OR)
Number of cities	416	277	548	274
Number of tweets	544,934	987,444	346,710	1,402,344
Tweets w/ hashtags	15.5%	18.7%	21.7%	19.2%
Tweets w/ nouns	98.4%	98.0%	98.5%	98.3%
Tweets w/o gazetteer words	58.5%	60.8%	59.3%	63.8%
Number of users	6,949	16,524	11,175	16,097
Users w/o gazetteer tweets	5,516	13,071	8,517	13,210
Total hashtags	151,996	388,985	150,194	526,058
Unique hashtags	39,878	90,601	45,450	145,621
Total nouns	2,362,756	4,234,728	1,295,720	6,037,733
Unique nouns	25,353	31,743	23,388	35,100

Table 1. Summary of Twitter Data Collected

4 Experimental Evaluation

4.1 Dataset

We collected the Twitter handles of users from August 1, 2017 to August 10, 2017 whose tweets were geotagged as originating from any of the 50 states in the USA using Twitter's Streaming API. Over the span of these 10 days, we collected the handles of 368,552 unique users. For each user, in November 2017, we collected up to 3200 (the limit set by Twitter's API) of the most recent tweets that they had posted. In total, we managed to collect 843,635,243 tweets for all 368,552 users. These tweets spanned over 11 years with the oldest tweet being from September 2006 and the most recent one from November 2017. We are able to identify 140,721,139 geotagged tweets from the corpus of 843,635,243 tweets. When testing, we ignore the geotagged location in geotagged tweets and only use the tweet text to evaluate prediction accuracy.

It is noteworthy that we can only use geotagged tweets to train and test supervised machine learning approaches (including JASOOS) because we do not have location ground truth for non-geotagged tweets. It is unclear whether we can expect a location-inference approach trained and tested on geotagged tweets to perform comparably on non-geotagged tweets. For instance, geotagged tweets are more likely to be sent from GPS-capable mobile devices than non-geotagged tweets. Thus, there may be differences in vocabulary of geotagged and non-geotagged tweets [19] that potentially impact the generalizability of a location-inference approach.

Prior literature on location inference has also faced the challenge of lack of ground truth for non-geotagged tweets. Han et al. [21] showed that a location-inference approach trained and tested on geotagged tweets performs comparably when applied to only non-geotagged tweets. To investigate this further, we conducted our own experiments comparing the vocabulary of geotagged and non-geotagged tweets in our collection. Specifically, we computed Spearman's rank correlation coefficient between random samples of geotagged and non-geotagged tweets (considering only nouns and hashtags) in our dataset. Each sample contained 50,000 tweets and the experiment was repeated 100 times. Table 2 shows that the average correlation between geotagged and non-geotagged tweets (0.714) is comparable to the average correlation between geotagged and geotagged tweets (0.728) and non-geotagged and non-geotagged tweets (0.677). We conclude that the vocabulary of geotagged and non-geotagged tweets are similar in our dataset. Thus, in line with prior research [21], we expect location inference approaches trained and tested on vocabulary of geotagged tweets to perform comparably on non-geotagged tweets.

Datasets	Rank Correlation
Geotagged - Geotagged	0.72766
Non-geotagged - Non-geotagged	0.67726
Geotagged - Non-geotagged	0.71416

Table 2. Comparison of vocabulary in geotagged and non-geotagged tweets

In order to evaluate the effectiveness of our proposed approach in predicting tweet location, we chose four states spanning different geographical regions and demographics. We chose Colorado (CO) from the Mountain region, Oregon (OR) from the West region, and Iowa (IA) and Wisconsin (WI) from the Midwest region. Table 1 provides a summary of the geotagged data set with respect to these four states. There are 416 city

level locations in Iowa, 548 in Wisconsin, 274 in Oregon, and 277 in Colorado suggesting that it may be harder to locate Iowa/Wisconsin tweets because of their larger numbers of cities. There are 544,934 tweets from Iowa, 346,710 tweets from Wisconsin, and almost double the number, 987,444 tweets from Colorado, and the highest number, 1,402,344, from Oregon. This also points to the Iowa and Wisconsin being possibly more challenging because there is relatively less training data. Around 60% of the tweets for these states do not have gazetteer words [1]. Most have nouns as expected while less than 22% have hashtags; thus we anticipate low coverage with the latter.

4.2 Experimental Setup

As indicated earlier, we use N fold cross-validation design. We split our data set based on users with $N = 10$. Folds are first made independent of tweet timestamp ensuring that all tweets from a user are placed in the same fold. Additionally, for the temporal model we process the data set with a day-level time window. The cross validation strategy is identical as described above, except that the folds are limited to the day of interest. While we process the data set in date order, there is no dependency in processing between any pair of days.

4.3 Evaluation Metrics

Recall that our decision strategy works on the basis of agreement between the atemporal and temporal ranking decisions, i.e., $Agreement_{@R}$. A decision is made only if both temporal and atemporal have at least one location in common by the rank cutoff of R. Thus our evaluation metrics can be calculated at different values of R. We experiment with $R = 1$ and $R = 10$. Evaluation at $R = 1$ refers to a strict configuration and $R = 10$ refers to a more lenient configuration for Jasoos. $R = 1$ is an obvious choice, as it requires both system to be absolutely certain in a decision. The maximum value of R is the total number of locations, which for any state is in the order of hundreds. In comparison to this maximum value of R, $R = 10$ is a strict constraint. Other choices of R could have been made besides $R = 10$, however, $R = 10$ seemed reasonable, as it is an order of magnitude higher than $R = 1$.

Rank error is measured as the number of locations predicted incorrectly, i.e., with higher likelihood, than the correct location. For example, if a system predicts three other locations with higher (or equal to) probabilities than the correct location, the rank error is 3. In the case of a tie, all predictions following the tie are decreased by the number of locations in that tie. For instance, if two locations are tied for rank 2, then the next predicted location will have rank 4 (not 3).

When combining the temporal and atemporal rankings for a tweet, the rank is the average of the two rankings. Table 3 illustrates this with data for an example test tweet. In the table the letters represent locations. Let us assume that the correct location for this tweet is B. The average rank for B between temporal and atemporal is 2.5 whereas for location Z it is 3.5. Locations are re-sorted by this average rank and this new order is used to generate final ranks. Thus the correct city ends up at rank 2.

Temporal		Atemporal		Combined		Combined Sorted	
Loc.	RK	Loc.	RK	Loc.	RK	Loc.	RK
A	1	B	1	A	2	A	1
C	2	Z	2	B	2.5	B	2
D	3	A	3	C	3	C	3
B	4	C	4	D	4	Z	4
Z	5	D	5	Z	3.5	D	5

Table 3. Example illustrating combining two rankings (RK = Rank), (Loc. = Location)

Coverage. Given that our decision model is selective about tweets for which location will be predicted, we calculate tweet coverage. This is calculated as the percentage of tweets – without gazetteer words – for which a prediction is made. Thus, tweet coverage, is defined as:

$$\frac{\#\text{ of tweets for which location is predicted}}{\#\text{ of tweets without gazetteer words}}$$

User coverage & User Precision While our focus is on performance at the level of tweets we can also calculate parallel measures at the user level. For example, user coverage is:

$$\frac{\#\text{ of users for whom a tweet location is predicted}}{\#\text{ of users in tweets without gazetteer words}}$$

Likewise, if a correct prediction is made for at least 1 tweet posted by 80 users and predictions are made for tweets from 100 users, then user precision is 0.8.

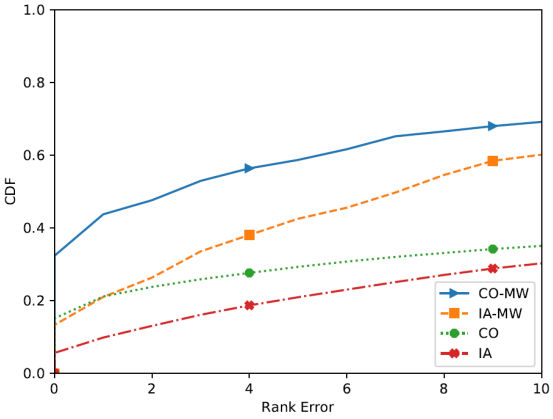


Fig. 2. Comparison of *maxwordNB* (MW) and standard Naive Bayes for Iowa and Colorado using Hashtags+Nouns as features. The y-axis represents percentage of tweets predicted within particular rank error (x-axis)

4.4 Results

We test JASOOS on tweets collected from four different states Iowa (IA), Colorado (CO), Wisconsin (WI), and Oregon (OR) using three different feature sets (Hashtags, Nouns, Hashtags+Nouns), and test our intuition about *maxwordNB*.

***maxwordNB* outperforms *NB*.** As a preliminary step we tested both a regular Naive Bayes system as well as our *maxwordNB* system experimenting with all three feature sets with data from two states. We consistently find *maxwordNB* to be superior. An example plot demonstrating this can be found in Figure 2. As the figure shows, *maxwordNB* outperforms *NB* in Iowa and Colorado states. Since Wisconsin shares a similar tweet-to-city count as Iowa and Oregon similar to Colorado, only Iowa and Colorado are shown. The smallest difference is found at rank error 0 (8%) while the largest difference is found at rank error 9 (33%). Past rank error 10, the difference grows further. This confirms our intuition about the use of one best feature being more effective than using all features. It is possible that tweets located with low rank error are dominated (in probability) by a single feature that tends to point to the correct location. As the rank error increases the probability associated with the best feature may be diminishing. While this remains a win for the single feature seeking *maxwordNB*, this probability likely weakens further when multiplied by the probabilities of the other tweet words for standard Naive Bayes.

Performance is strong for both $R1$ and $R10$ configurations. Recall that JASOOS only makes a prediction on a tweet if the temporal approach and atemporal approach both contain a city in common within a rank cutoff R. Results are in Table 4 for $R1$ and Table 5 for $R10$. Overall, the trends are consistent across both tables.

The setting with Hashtags+Nouns provides the best coverage and rank error scores for each state. Rank error at the first quartile is between 0 and 3. Median rank error achieved for Colorado, for example, with feature set hashtags+nouns, is 3. This means that for 50% of the 598,535 located tweets, rank error is 3 or less. Moreover, the correct city is at rank 1 for 187,515 Colorado tweets. Similarly, the median rank error for Oregon is also 3 for 894,331 tweets. For Iowa the median rank error is 8 for the 306,244 tweets located. For Wisconsin, the median rank error is 16 for 205,566 tweets. Wisconsin is also the state with the largest number of cities and the smallest number of tweets. Coverage achieved is strong to excellent, being between 40 and 75%, depending on state.

Hashtags+Nouns is the most effective. Tables 4 and 5 compare performance across feature sets. Coverage shows the biggest difference between feature sets. Upwards of 75% coverage is achieved by the combination of Hashtags+Nouns for Colorado, and 62% coverage is achieved for Oregon. As expected, because of their sparsity, the Hashtags feature set consistently has very low coverage. The Nouns feature set achieve coverage that is in between. Rank error differences are comparatively small, except in the case of Wisconsin (75%). Overall, this measure appears resilient to the feature set used.

Oregon and Colorado locations are easier to predict. As seen in Tables 4 and 5, rank errors show a slight preference for Colorado and Oregon. For example, rank errors are between 0 to 4 for Colorado, 0 to 3 for Oregon, 3 to 16 for Wisconsin and 2 to 9 for Iowa when considering the first two quartiles. Oregon and Colorado's slightly better coverage might be attributed to a couple of reasons. First, as seen in Table 1, Oregon and Colorado have over 50% more tweets than Iowa or Wisconsin. A higher number of tweets means more training data, which could allow JASOOS to obtain the necessary information to make better decisions. Second, Iowa and Wisconsin each have over 66% more cities than Colorado or Oregon. A higher number of cities means a higher number of locations from which a tweet can be identified, increasing the difficulty of the problem.

	Feature Set	Rank Errors			Coverage	User Precision	User Coverage
		25%	50%	75%	[%]	[%]	[%]
IA	Hashtags	2	9	22	0.74	26.67	13.32
	Nouns	3	8	20	41.39	24.25	80.51
	Hashtags+Nouns	2	8	20	40.21	25.16	94.91
CO	Hashtags	0	3	19	2.42	47.36	19.29
	Nouns	0	3	16	74.14	50.50	90.82
	Hashtags+Nouns	0	3	16	73.22	51.25	97.87
WI	Hashtags	0	6	57	1.04	43.54	6.63
	Nouns	3	17	53	29.53	35.10	53.48
	Hashtags+Nouns	3	16	51	48.06	36.45	62.59
OR	Hashtags	0	0	10	1.94	59.09	18.91
	Nouns	0	3	18	47.62	62.03	69.13
	Hashtags+Nouns	0	3	17	61.74	63.73	73.46

Table 4. Results for JASOOS with rank 1 cutoff

	Feature Set	Rank Errors			Coverage	User Precision	User Coverage
		25%	50%	75%	[%]	[%]	[%]
IA	Hashtags	2	9	22	1.79	55.00	22.32
	Nouns	3	8	20	92.71	70.17	97.43
	Hashtags+Nouns	2	8	20	93.53	50.73	99.60
CO	Hashtags	0	4	20	3.58	60.75	23.45
	Nouns	0	3	16	93.96	65.47	97.87
	Hashtags+Nouns	0	3	16	94.91	56.54	99.86
WI	Hashtags	0	9	56	1.80	53.42	10.11
	Nouns	3	16	51	46.99	48.56	62.55
	Hashtags+Nouns	3	16	51	53.63	43.35	66.04
OR	Hashtags	0	1	13	2.43	67.61	21.67
	Nouns	0	3	17	59.85	75.58	73.34
	Hashtags+Nouns	0	3	17	63.45	68.09	74.69

Table 5. Results for JASOOS with rank 10 cutoff

This combination of reasons might be responsible for a harder problem for Iowa and Wisconsin as compared to Colorado and Oregon.

Location-revealing words are not necessarily gazetteer words. Table 6 provides examples of tweets correctly located with the *maxwordNB* feature responsible underlined. We note that JASOOS is able to correctly predict the location of tweets using features which are unlikely to be included in a gazetteer. For example, "#BeatTexas" is not a gazetteer word (for Iowa) but instead behaves like one during football matches between the University of Iowa based in Iowa City, IA and Texas State University. Similarly "#cyclONEgrad" does not refer to any point of interest in Ames, IA but instead it refers to a graduate from Iowa State University; the athletics team from Iowa State University is referred to as Cyclones. Even though these words are not like gazetteer words, their strong spatial affinity to locations facilitates the identification of locations just as gazetteer words would.

Table 7 provides examples from the Iowa dataset illustrating how the temporal model operates. We see nouns such as 'semester, 'night', 'stream', 'today' as maxwords. We see maxwords related to other events ('today' referring to college football taking place that day). We see topics of local interest such as the hashtag '3synodia17', 'dis2017' and we see maxwords that point to location names, but in ways that cannot be anticipated ('stream'). All of these maxwords are *temporally* location indicating. Moreover, all of these could potentially be spatially diffuse at other times or could be maxwords for some other locations at still other points in time. A noun like 'semester', maxword for Ames, IA on December 21, 2016, could show up for another city experiencing containing a university whose semester ends at a different time. Also it may be that on a given date 'semester' is spatially diffuse (multiple universities end-

Tweet Text	City	Temp Prob.	Atemp Prob.
Congrats, Nettie! #cyclONEgrad	Ames, IA	0.152	0.057
#dis2017 follow me pls	Ames, IA	0.140	0.036
Congrats @karter_schult #UNI https://t.co/MmT8w3GClj	Cedar Falls, IA	0.059	0.0388
Colt Cadets!#TournamentOfDrums https://t.co/8E2V3CNvSd	Cedar Rapids, IA	0.193	0.041
#LetKyleRace https://t.co/IKxJqpDzk8	Knoxville, IA	0.109	0.028
Katie Davis never disappoints. #5HourNats	Knoxville, IA	0.236	0.153
Simply. Awesome. #SolheimCup2017 @abba https://t.co/UEKk8w6hok	West Des Moines, IA	0.140	0.147
Hurray for #fedexguy ! https://t.co/zUcNTdIBy0	Urbandale, IA	0.055	0.052
Jack Trice it's 2019s been awhile #BeatTexas	Ames, IA	0.110	0.141

Table 6. Examples of tweets correctly located by JASOOS using the underlined maxwords.

ing on the same day) and therefore not show up as a maxword for any location.

Table 8 provides the a list of the top ten highest probability maxwords identified by both the temporal system and atemporal system for correctly located tweets. Though some overlap of maxwords is seen between temporal and atemporal, the different systems identify different maxwords overall. The temporal system is able to identify words that can be temporally diffuse, for example 'seahawks' (which refers to the Seattle football team) appears in Wisconsin, this is most likely due to the Green Bay, Wisconsin Packers playing the Seahawks during this time. The atemporal system is able to identify words that are more spatially diffuse, for example 'brewers' is identified in Milwaukee and is the professional baseball team for this city.

The takeaways from this analysis are that (i) nouns play a significant role, (ii) many of the nouns that give away location have no inherent spatial semantics and are not likely to be ever seen in a gazetteer and (iii) therefore these are difficult to identify proactively. Instead they would have to be identified contemporaneously with the tweet being written that needs location protection.

Conjecture. From this analysis we conjecture the following about the role of common nouns in our model. As the spatial and temporal granularities of the location problem considered become more fine-grained an increasing number of common nouns are likely to become location revealing. In other words if instead of taking all cities in Iowa as our dataset, we had taken just the cities in the south west region, then the number of location indicating common nouns will likely increase. Likewise, if we decrease the spatial and temporal granularities of the problem, taking for example all cities of the mid western states in one dataset, then the power of these common nouns at location detection will decrease. Instead the method would have to rely more on city level atemporal features that more durably (over time and space) identify location. This is a conjecture that likely applies to other location detection methods that rely on the probabilistic model as well and we will test it in follow up research. This type of analysis of the features behind location detection would enrich cross study comparisons that vary in their location detection granularities.

Privacy is at risk. Our results indicate a strong possibility of location privacy risk even for users who take precautionary measures. Our system with the best configuration predicts the correct location for $\sim$ 12% to $\sim$ 25% of the test tweets depending on the state. This is out of large sets: 416 cities for Iowa, 277 for Colorado, 548 for Wisconsin, and 274 for Oregon.

These located tweets correspond to 1,102 users in Iowa, 6,006 users in Colorado, 6,184 users in Oregon, and 1,943 users in Wisconsin. These users are at risk in terms of location privacy. Moreover, the decisions made by JASOOS when triangulated with additional evidence from other location inference methods, (say from social network data) - which we have intentionally avoided - offers the potential to reduce the rank error further.

4.5 Comparison to Other Systems

To gauge how effective JASOOS is in comparison with state-of-the-art approaches we implemented and tested two other systems within the same threat model. First, we replicated the temporal approach proposed by Paraskevopoulos and Palpanas [38] which is based on the vector space model with tf*idf weighted vectors representing each location at a period of time. We replicated their best temporal method and applied it to our

Tweet Text	Timestamp	City	Temp Prob.
How to we strengthen ministry of all the baptized? #3synodIA17	2017-09-25 18:14:40	West Des Moines, IA	0.293
Fundraising - the art of the ask @LSTChicago #3synodIA17	2017-09-25 19:02:21	West Des Moines, IA	0.643
vamos! #3synodIA17	2017-09-26 16:47:59	West Des Moines, IA	0.590
Congrats to this #cyclONEgrad! https://t.co/yCeBvz1rlf	2016-12-17 19:20:18	Ames, IA	0.191
Congrats, @ISUnettie! #cyclONEgrad https://t.co/6o9gZqKQsW	2016-12-17 21:59:03	Ames, IA	0.153
#DIS2017 https://t.co/xH7sD83IK7	2017-08-20 01:08:51	Ames, IA	0.162
#dis2017 follow me pls	2017-08-20 01:28:23	Ames, IA	0.140
#dis2017 #dabforisaac	2017-08-20 01:30:11	Ames, IA	0.140
3 hours left on my last 10 hour shift of the week and the local dive bar is calling for me	2017-08-26 00:35:33	Urbandale, IA	0.194
I got to see Izzy and Link and WILLIE and POPPY last night I'm so hap	2017-08-26 19:02:05	Urbandale, IA	0.160
I hate when people undermine the difficulties of other people's activities like sports, when they have never experienced them before.	2017-08-05 03:56:25	Urbandale, IA	0.127
I love when people tell me they appreciate me just for being me because most of the time I feel like I'm not doing anything right.	2017-08-05 06:22:43	Urbandale, IA	0.112
@realDonaldTrump I like how you have to point this out. Must be since you so seldom actually work.	2017-08-05 22:38:27	Urbandale, IA	0.123
I wish somebody video taped me going Super Saiyan on these hoes last night	2017-09-04 06:04:00	Urbandale, IA	0.221
So I had a #Nightmare last night;I was a #server again and I was horrible at everything i will not go back to that again! #NeverGoingBack	2017-09-04 14:13:44	Urbandale, IA	0.224
Another semester	2016-12-16 15:54:56	Ames, IA	0.110
I GOT A 4.0 THIS SEMESTER I AM SO HAPPY	2016-12-21 01:15:01	Ames, IA	0.064
My semester GPA at DMACC was a 3.91!!!	2016-12-21 06:08:44	Ames, IA	0.067
@pirmas697 and @raysngs with the cameos on the stream #WatchCity	2017-07-29 00:42:32	Davenport, IA	0.050
@PolskaKrolowa Stream said 7:30 KO	2017-07-29 02:04:59	Davenport, IA	0.050
But seriously how bleeping awesome of a feeling it is to wake up today...Cccooollleeegggeee Fffooottttbbbaaallll ba https://t.co/jCwrh91bfp	2017-09-02 13:41:47	Urbandale, IA	0.152
@UNIFootball good luck today guys #UNIFight #BeatState	2017-09-02 16:32:31	Urbandale, IA	0.123
Hey @GanassiChip, #LetKyleRace... https://t.co/OEduqkUTK3	2017-08-10 09:25:24	Knoxville, IA	0.104
#LetKyleRace https://t.co/ylixBRRZy1	2017-08-10 13:57:24	Knoxville, IA	0.109

Table 7. Examples of tweets correctly located using the underlined temporal maxwords. (Iowa dataset, Hashtags + Nouns, Rank 1 cutoff)

Iowa dataset maintaining the same cross-validation design as for JASOOS. Results are shown in table 9.

As a second comparison, we designed and implemented a temporal deep learning approach. The model

Temporal							
Iowa		Colorado		Wisconsin		Oregon	
Maxword	Probability	Maxword	Probability	Maxword	Probability	Maxword	Probability
3synodia17	0.6430	drinking	0.8909	seahawks	0.8121	block	0.9145
fair	0.5766	family	0.7761	bbc17	0.7470	year	0.8690
womensmarch	0.4199	american	0.6838	cubs	0.6571	ye	0.8504
des	0.4087	oscars2017	0.6829	state	0.6133	police	0.7867
state	0.4069	amp	0.6273	gophers	0.5745	rctid	0.7680
temperature	0.3902	rockies	0.6265	brewers	0.5530	drinking	0.6911
total	0.3783	grammys	0.6264	packers	0.5459	cold	0.6901
traffic	0.3773	god	0.6183	thisismycrew	0.5399	person	0.6424
ia	0.3670	womensmarch	0.6109	saints	0.5341	baonpdx	0.6388
amp	0.3375	man	0.6094	marchmadness	0.4817	photo	0.6158
Atemporal							
Iowa		Colorado		Wisconsin		Oregon	
Maxword	Probability	Maxword	Probability	Maxword	Probability	Maxword	Probability
total	0.8646	collins	0.8841	seahawks	0.4337	pdx	0.8153
des	0.7856	es	0.7163	osh17	0.4258	police	0.8142
west	0.7856	copolitics	0.6756	art	0.4204	se	0.8139
hay	0.6169	maga	0.6596	park	0.4178	ne	0.8132
merle	0.6142	coleg	0.6455	ramp	0.4120	block	0.8126
justiceleague	0.5960	downtown	0.6334	brewers	0.4108	medical	0.8032
hour	0.5793	traffic	0.6050	thisismycrew	0.3937	airport	0.7691
temperature	0.5270	powers	0.5952	bbc17	0.3823	rctid	0.7681
issue	0.4651	quality	0.5826	bucks	0.3741	c	0.7622
3synodia17	0.4469	steamboat	0.5657	mke	0.3724	baonpdx	0.7599

Table 8. Top 10 Maxwords (duplicates removed) appearing in correctly identified tweets for datasets (Hashtags + Nouns)

we implemented was the CNN model designed for text classification problems by Kim [27]. This CNN first constructs feature based tweet representations of size k x d. k is chosen as the max number of features (nouns) found in training tweets, while d is the chosen dimension of the word embedding (in our case d = 100). Pre-trained word embeddings are obtained from Glove[3]. The tweet representations are fed into 3 sets of convolutional windows of length (3 x d, 4 x d, 5 x d) each window having 100 filters each. We use max pooling and finally a dense layer followed by a softmax output layer (of size = number of cities). Similar to JASOOS we use the same decision strategy which is to combine the decisions made by the temporal and atemporal CNN models (see Decision Strategy in the Proposed Approach section). Again the cross validation design is used with the same folds of data. The temporal model is constructed by first training the CNN on 2 folds of atemporal data. Next, given a day's collection of test tweets (these are taken from a third fold) the model is retrained on the remaining 7 folds of data for that same day before applying it to the test tweets. Since hashtags are sparse, word embeddings for hashtags are rare causing the CNN to break. Because of this we focused on nouns alone as features for the CNN. We present results for the Iowa dataset in table 10.

Both the vector based model and the CNN are evaluated under the same condition of no gazetteer words. Comparisons were made using the rank 1 cutoff setting.

Analysis of comparison results:

JASOOS **achieves higher performance over the vector based approach of [38].** Selecting nouns as the best version for the vector based system (because of its high coverage) we find that the rank error for JASOOS is better by 7 points for the 25% quartile (hashtags+nouns) and this difference increases with each higher quartile (19 points and 75 points at 50% and 75% quartiles respectively). Coverage varies between system, with JASOOS achieving higher coverage in Hashtags + Nouns, while [38] achieve higher coverage in with nouns.

CNN outperforms JASOOS **in coverage.** The CNN approach is almost identical to JASOOS. Rank error is one better for the CNNs for the first quartiles. The largest difference seen is in Coverage, which is 92% compared to JASOOS's 40%. Thus the CNN performs the same as JASOOS while providing superior coverage.

3 https://nlp.stanford.edu/projects/glove/

However, it should be noted that JASOOS has the added advantage of interpretability which is useful for further understanding this threat model and constructing countermeasures.

4.6 Countermeasures

Next, we implement and evaluate the three different countermeasures:

1. *deletion*: delete the maxword identified by JASOOS from the tweet;
2. *addition*: add a maxword identified by JASOOS representing another location; and
3. *deletion and addition*: delete the maxword identified by JASOOS from the tweet and add a maxword identified by JASOOS representing another location.

Both deletion and addition countermeasures degrade JASOOS's accuracy. Table 11 reports the results for these three countermeasures for nouns and nouns + hashtags on IA data set. We note that rank error increased substantially for both deletion and addition countermeasures. Specifically, the 25th percentile rank error increases by at least 9 for both deletion and addition. The 50th percentile and 75th percentile rank error exhibit even greater increases in rank error, increasing by 27 and 71 respectively. We also note drop in coverage for both deletion and addition countermeasures in the nouns feature set. However, we note an increase for the addition countermeasure in the hashtags + nouns feature set. This increase in coverage combined with the increase in rank error means that JASOOS is making more wrong predictions when the addition countermeasure is applied. Finally, we note 18% drop in user precision and 64% drop in user coverage when countermeasures are applied. Note that combining the two countermeasures (deletion and addition) does not provide any additional benefit.

Naive Bayes is resilient to deletion, but not addition. Since JASOOS's location inference is based on one feature (i.e. maxword), it might be more susceptible to the aforementioned countermeasures than a standard Naive Bayes algorithm. To investigate this hypothesis, we evaluate the countermeasures on a standard Naive Bayes algorithm. Table 12 shows that rank error generally improves slightly with the deletion strategy. However, the addition countermeasure causes large increases in rank error: 19 for the 25th percentile, 23 for the 50th percentile, and 8 for the 75th percentile. We conclude that there are feasible countermeasures to both JASOOS (based on maxword) and standard Naive Bayes. While there are a couple of counter strategies (of the ones tested) for maxword Naive Bayes there is only one for standard Naive Bayes.

4.7 Limitations

Here we discuss the current limitations of our system and expand on possible future solutions.

States are run in isolation. Currently, JASOOS is trained on and predicts each state separately. This means that predicting for Iowa, the system knows it must occur in one of the 416 list of cities. This is an obvious limitation of our system as an adversary might not necessarily know which state the tweet occurs in. One way to possibly address this issue, is by first generalizing our system to predict state. Then after a prediction for a state is made, further classify the tweet using the set of cities from the predicted state. Future work would implement and test this model.

Rank error does not guarantee proximity. Currently, one of the main metrics used to evaluate the systems is rank error. However, having a good rank error (e.g rank error = 5) does not necessarily mean that our system predicted in a closer proximity to the true location. For example, if the true location of a tweet is Des Moines, and our system predicts it as the 5th most likely city, the four cities which occur before it, may be large distances from Des Moines. One advantage of rank error versus distance, however, is that rank error remains consistent when changing states/countries. Error distance will be much lower when predicting in Rhode Island, compared to Texas, whereas rank error does not fluctuate based on physical size.

5 Related Work

In this section, we review prior literature on tweet location inference. As discussed in a very recent review by Zheng et al. [52], three types of location detection problems have been considered: user home location, tweet location and locations mentioned in a tweet.While our focus is on tweet location, we note that there are many papers with home location inferences as the goal (e.g., [42] [43] [34] [40] [45]). [26] A variety of machine learning and statistical methods have been explored, our main focus is on the feature sets that they employ.

	Feature Set	Rank Errors			Coverage	User Precision	User Coverage
		25%	50%	75%	[%]	[%]	[%]
JASOOS	Hashtags	2	9	22	0.74	26.67	13.32
	Nouns	3	8	20	41.39	24.25	80.51
	Hashtags+Nouns	2	8	20	40.21	25.16	94.91
Vector	Hashtags	77	94	103	2.44	18.52	27.01
	Nouns	9	27	95	91.59	33.97	97.44
	Hashtags+Nouns	8	28	96	7.77	17.14	44.85

Table 9. Results for JASOOS and the vector based temporal model [38] with rank 1 cutoff (Iowa dataset)

System	Rank Errors			Coverage	User Precision	User Coverage
	25%	50%	75%	[%]	[%]	[%]
JASOOS	3	8	20	41.39	24.25	80.51
CNN	2	8	20	92.23	25.04	97.61

Table 10. Results for JASOOS and the CNN based model with rank 1 cutoff (Iowa dataset), Nouns used as features

	Feature Set	Rank Errors			Coverage	User Precision	User Coverage
		25%	50%	75%	[%]	[%]	[%]
None	Nouns	3	8	20	41.39	24.25	80.51
	Hashtags+Nouns	2	8	20	40.21	25.16	94.91
Del.	Nouns	12	35	91	26.48	6.56	26.95
	Hashtags+Nouns	12	34	93	39.81	6.56	29.29
Add	Nouns	12	35	91	30.01	6.69	27.89
	Hashtags+Nouns	13	34	93	49.04	6.66	30.89
Both	Nouns	12	35	91	26.53	6.56	26.93
	Hashtags+Nouns	12	35	94	41.17	6.56	29.62

Table 11. Results for JASOOS after applying countermeasures with rank 1 cutoff (Iowa data set)

	Feature Set	Rank Errors			Coverage	User Precision	User Coverage
		25%	50%	75%	[%]	[%]	[%]
None	Nouns	35	94	147	71.19	10.06	34.31
	Hashtags+Nouns	37	96	147	72.79	9.28	34.61
Del.	Nouns	36	89	145	71.24	9.53	34.34
	Hashtags+Nouns	33	91	145	72.79	9.13	34.61
Add	Nouns	54	117	155	71.24	9.13	34.34
	Hashtags+Nouns	56	119	154	72.79	7.74	34.61
Both	Nouns	56	123	157	71.24	8.49	34.34
	Hashtags+Nouns	50	114	152	72.79	9.37	34.61

Table 12. Results for Naive Bayes after applying countermeasures (Iowa data set)

5.1 Gazetteer Approaches

Prior research using gazetteers for tweet location inference simply looks for the presence of gazetteer words in tweets. It should be noted that gazetteers can be extended to include non-standard name phrasings as described next.

One angle has been to identify location names in tweets using different Named-Entity Recognition (NER) tools. Gelernter and Mushegian [16] used the Stanford NER tool to identify location names mentioned in tweets about disaster events. They reported that the Stanford NER tool did not identify names of non-standard location phrasings. Lingad et al., [30] compared the effectiveness of existing NER tools such as StanfordNER and OpenNLP for geolocating tweets in disaster events. They showed that retraining existing NER tools on Twitter data improved their effectiveness by helping them identify non-standard location phrasings.

Another line of research has focused on identifying location names in tweets using external gazetteers of crowdsourced location-based services such as Foursquare. For example, Li and Sun proposed extracting location names from tweets using Foursquare's crowdsourced POI inventory containing standard and many non-standard location phrasings [29] [4]. Schulz et al., used an external NER service DBpedia Spotlight as well as crowdsourced location-based services such as UberSocial, TrendsMap, Flickr, Roketatchi, and Foursquare [47].

The gazetteer lists from these work include not only standard locations such as states or city names (referred to as "standard" gazetteers) but also more fine-grained ones such as landmarks, stadiums or restaurants (referred to as POIs or "extended" gazetteers). We expect privacy seeking users to avoid these words in their tweets. For this reason, our threat model specifically targets tweets that do not include location names found in a well reputed gazetteer [1]. Our method aims to identify other words in tweet text that can reveal location. Next, we discuss prior literature on automatically identifying such location revealing words in tweet text.

5.2 Tweet Content

Many examples of prior research utilize unigram and n-gram features extracted from tweet content alongside different prediction models. For example, Flatow et al. [15] and Priedhorsky et al. [41] used Gaussian models,Paule et al. use a voting strategy with tweets similar to the test tweet voting on location [18], Chong and Lim [11] combined a variety of probabilistic models while Ozdikis et al. use probabilities estimated using kernel density estimations [37]. Liu et al. [31] used a Hidden-Markov model, Hulden et al. [24] used kernel densities with a Naive Bayes model, Cheng et al. [10] leverage a maximum likelihood estimate along with various smoothing techniques. Hahmann et al. [20] used Naive Bayes and Maximum Entropy models, Iso et al. [25] used a Convolutional Mixture Density Network (CMDN), Hong et al. [23] used a modified Sparse Additive Generative (SAGE) model, and Zhang et al. [51] use a random forest classifier. The paper by Zhang et al. [51] is noteworthy in that they also examine loss in privacy due to the use of hashtags. However, they do not eliminate hashtags derived from gazetteer entries and they do not consider temporal factors. In general, there are very few attempts at considering temporal features. One example is that of Paraskevopoulos and Palpanas [39] who created vectors from important extracted keywords and updated them by sliding windows. A second example is by Yamaguchi et al., [49] who generated word-level location distributions that change over time.

It is noteworthy that these prior works do not exclude gazetteer words. Thus their performance independent of gazetteer words is an unknown. In contrast, we specifically focus on tweets by privacy seeking users that do not include gazetteer words. In addition, our approach incorporates both atemporal and temporal features from tweet content to capture non-gazetteer words such as non-standard location phrasings for one-off events (e.g. a book festival) and locally interesting topics (e.g., a term referring to graduating from a local College).

5.3 Tweet Metadata

Prior literature has also targeted location inference with tweet meta-data features such as social network, interaction history with friends/followers, profile information, and third-party sources (e.g., external links). Researchers have trained a dynamic Bayesian network to estimate a user's location through the locations of friends in the social network [46]. Chong and Lim showed that the similarity of tweet content can help infer a user's location history [12]. Cao et al., exploited geographic information from a user's social network to analyze the embedded social relations of POIs [8]. Bakerman et al., explore a hybrid approach of combining tweet content and user network information [6].

In addition to social network information, prior literature has used meta-data from user profiles such as timezone and third-party content such as hyperlinks to infer tweet location. Schulz et al., used tweet meta-data such as timezone, user's profile location, and external hyperlinks to generate a spatial distribution of tweet location [47]. Priedhorshy et al., exploited spatially significant n-grams extracted from tweet meta-data such as user description, location, and timezone to infer tweet location [41]. Zubiaga et al., used tweet meta-data such as language to infer tweet location at the country level [53]. Dredze et al., used tweet timestamp and timezone information in addition to tweet text to infer tweet location [14].

In contrast to these works we do not consider tweet meta-data because our focus is on privacy seeking users who are careful enough to not include location hints in their meta-data. In sum, we focus on estimating location from just the tweet text, without any meta-data, under

the strict condition of not having access to any obviously location indicative words.

6 Concluding Remarks

We presented a novel threat model of inadvertent location privacy leaks on Twitter even when a user deliberately avoids using location indicating words from a gazetteer. We presented JASOOS, built on a standard Naive Bayes approach, to demonstrate the existence of this threat model. We used a popular probabilistic approach as the basis for JASOOS in order to determine the extent to which a standard location detection approach can threaten privacy. Our system covers between 40% to 74% of the tweets depending on the state with average rank error (across states) of 1.25 for the top quartile of tweets and average rank error of 7.5 at median point.

In comparison to JASOOS, while our replication of a state-of-the-art temporal vector based approach proposed in [38] achieves much better coverage (92%), it yields considerably lower rank errors (for Iowa) of 8 for 25% quartile and 28 for median. In comparison to JASOOS, a CNN based temporal approach that we designed, also achieves the same high coverage of 92% as the vector approach while giving the best top quartile rank error = 2 with rank 1 cutoff. Rank errors for the median and third quartile are identical to JASOOS.

These results demonstrate the presence of a new threat model that could target location privacy. We observe that the threat to privacy may increase if the decisions made by JASOOS are combined with additional evidence from other location inference methods (say from social network metadata) – which we have intentionally avoided. Such combinations of evidence offer the potential to reduce the rank error further.

The key insight in JASOOS is that non-gazetteer words, and we find these include common nouns, which do not carry location semantics, acquire location revealing properties due to their usage patterns amongst users who do not keep their location private. We offer a conjecture for future research based on our analysis of location revealing words as to conditions under which common words reveal location especially using probabilistic models. The unpredictable nature and dynamics of vocabulary sharing between privacy seeking users, who do not disclose their location or use obvious location revealing words, and those who disclose their location makes it challenging for users to anticipate words responsible for leaking their location. However, we also propose countermeasures that show good potential at protecting privacy. In future work JASOOS can be leveraged using such countermeasures to develop a warning system that would ingest tweets to inform users about the usage of potential location revealing words in their tweets.

Since JASOOS relies on the availability of geotagged tweets to train a supervised machine learning algorithm, the higher the prevalence of geotagged tweets, the greater the threat faced by privacy conscious users seeking to keep their location private. As part of our future work, we are interested in evaluating JASOOS in different countries with varying (more/less) fraction of geotagged tweets [48]. We believe that JASOOS can be adapted to explore privacy considerations with other attributes (e.g., gender, ethnicity) as well. As long as a sufficient number of tweets tagged with the target attribute are available, our approach can capture latent correlations between tweet words and the target attribute.

Acknowledgement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] United States Board on Geographic Names - Domestic and Antarctic Names, U.S. Geological Survey. https://web.archive.org/web/20180912182706/https://geonames.usgs.gov/docs/stategaz/AllStates_20180801.zip.

[2] DHS' Pilots for Social Media Screening Need Increased Rigor to Ensure Scalability and Long-term Success. https://www.oig.dhs.gov/sites/default/files/assets/2017/OIG-17-40-Feb17.pdf, 2017.

[3] Social Media Fact Sheet, Pew Research Center. http://www.pewinternet.org/fact-sheet/social-media/, 2018.

[4] B. Ağır, K. Huguenin, U. Hengartner, and J.-P. Hubaux. On the privacy implications of location semantics. *Proceedings on Privacy Enhancing Technologies*, 2016(4):165–183, 2016.

[5] M. Allen. Health Insurers Are Vacuuming Up Details About You – And It Could Raise Your Rates, NPR. https://www.npr.org/sections/health-shots/2018/07/17/629441555/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates, 2018.

[6] J. Bakerman, K. Pazdernik, A. Wilson, G. Fairchild, and R. Bahran. Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(3):34, 2018.

[7] T. Brewster. Beyond Cambridge Analytica – The Surveillance Companies Infiltrating And Manipulating Social Media, Forbes. https://www.forbes.com/sites/thomasbrewster

/2018/04/18/cambridge-analytica- and-surveillance-companies-manipulate -facebook-and-social-media/6fced4e84053, 2018.

[8] B. Cao, F. Chen, and D. Joshi. Inferring crowd-sourced venues for tweets. In *2015 IEEE Int. Conf. on Big Data*, 2015.

[9] A. Chaabane, G. Acs, and M. A. Kaafar. You Are What You Like! Information Leakage Through Users' Interests. In *NDSS*, 2011.

[10] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[11] W. Chong and E. Lim. Tweet geolocation: Leveraging location, user and peer signals. In *ACM Conf. on Information and Knowledge Management*, 2017.

[12] W.-H. Chong and E.-P. Lim. Tweet geolocation: Leveraging location, user and peer signals. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1279–1288, New York, NY, USA, 2017. ACM.

[13] N. Confessore. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far, The New York Times. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html, 2018.

[14] M. Dredze, M. Osborne, and P. Kambadur. Geolocation for Twitter: Timing Matters. In *NAACL-HLT*, 2016.

[15] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza. On the accuracy of hyper-local geotagging of social media content. In *ACM Conf. on Web Search and Data Mining*, 2015.

[16] J. Gelernter and N. Mushegian. Geoparsing Messages from Microtext. *Transactions in GIS*, 2011.

[17] C. Gibbons. The FBI Is Setting Up a Task Force to Monitor Social Media. https://www.thenation.com/article/the-fbi-is-setting-up-a-task-force-to-monitor-social-media/, 2018.

[18] J. D. Gonzalez Paule, Y. Moshfeghi, J. M. Jose, and P. V. Thakuriah. On fine-grained geolocalisation of tweets. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 313–316. ACM, 2017.

[19] S. Gouws, D. Metzler, C. Cai, and E. Hovy. Contextual bearing on linguistic variation in social media. In *Workshop on Language in Social Media (LSM)*, 2011.

[20] S. Hahmann, R. S. Purves, and D. Burghardt. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *J. Spatial Information Science,*, 2014.

[21] B. Han, P. Cook, and T. Baldwin. Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, 2014.

[22] B. Hecht, L. Hong, B. Suh, and E. Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[23] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Conf. World Wide Web*, 2012.

[24] M. Hulden, M. Silfverberg, and J. Francom. Kernel density estimation for text-based geolocation. In *AAAI Conf. on Artificial Intelligence*, 2015.

[25] H. Iso, S. Wakamiya, and E. Aramaki. Density estimation for geolocation via convolutional mixture density network. *arXiv:1705.02750*, 2017.

[26] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu, and D. Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *ICWSM*, 2015.

[27] Y. Kim. Convolutional neural networks for sentence classification. *arXiv:1408.5882*, 2014.

[28] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 2013.

[29] C. Li and A. Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 43–52, New York, NY, USA, 2014. ACM.

[30] J. Lingad, S. Karimi, and J. Yin. Location extraction from disaster-related microblogs. In *22nd international conference on World Wide Web companion International World Wide Web Conferences Steering Committee*, 2013.

[31] Z. Liu and Y. Huang. Where are you tweeting?: A context and user movement based approach. In *ACM Conf. on Information and Knowledge Management*, 2016.

[32] J. Mahmud, J. Nichols, and C. Drews. Where Is This Tweet From? Inferring Home Locations of Twitter Users. In *AAAI Conference on Weblogs and Social Media*, 2012.

[33] H. Mao, X. Shuai, and A. Kapadia. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *ACM Workshop on Privacy in the Electronic Society*, 2011.

[34] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1260–1272, 2017.

[35] R. Nauen. Number of Employers Using Social Media to Screen Candidates at All-Time High, Finds Latest CareerBuilder Study, PR Newswire. https://www.prnewswire.com/news-releases/number-of-employers-using-social-media-to-screen-candidates-at-all-time-high-finds-latest-careerbuilder-study-300474228.html, 2017.

[36] L. Newman. Feds Monitoring Social Media Does More Harm Than Good, Wired. https://www.wired.com/story/dhs-social-media-immigrants-green-card/, 2017.

[37] O. Ozdikis, H. Ramampiaro, and K. Nørvåg. Locality-adapted kernel densities for tweet localization. In *41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[38] P. Paraskevopoulos and T. Palpanas. Fine-grained geolocalisation of non-geotagged tweets. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2015.

[39] P. Paraskevopoulos and T. Palpanas. Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Soc. Netw. Anal. Min.*, 2016.

[40] A. Poulston, M. Stevenson, and K. Bontcheva. Hyperlocal home location identification of twitter profiles. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 45–54. ACM, 2017.

[41] R. Priedhorsky, A. Culotta, and S. Y. D. Valle. Inferring the origin locations of tweets with quantitative confidence. In *ACM Conf. on Computer Supported Cooperative Work and Social Computing*, 2014.

[42] A. Rahimi, T. Baldwin, and T. Cohn. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. *arXiv:1708.04358*, 2017.

[43] A. Rahimi, T. Cohn, and T. Baldwin. A neural model for user geolocation and lexical dialectology. *arXiv:1704.04008*, 2017.

[44] L. Rainie. Americans' complicated feelings about social media in an era of privacy concerns, Pew Research Center. http://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/, 2018.

[45] E. Rodrigues, R. Assunção, G. L. Pappa, D. Renno, and W. Meira Jr. Exploring multiple evidence to infer usersâĂŹ location in twitter. *Neurocomputing*, 171:30–38, 2016.

[46] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 723–732, New York, NY, USA, 2012. ACM.

[47] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser. A Multi-Indicator Approach for Geolocalization of Tweets. In *ICWSM*, 2013.

[48] L. Sloan and J. Morgan. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS One*, 2015.

[49] Y. Yamaguchi, T. Amagasa, H. Kitagawa, and Y. Ikawa. Online user location inference exploiting spatiotemporal correlations in social streams. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1139–1148, New York, NY, USA, 2014. ACM.

[50] F. Zamal, W. Liu, and D. Ruths. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *AAAI Conference on Weblogs and Social Media*, 2012.

[51] Y. Zhang, M. Humbert, T. Rahman, C.-T. Li, J. Pang, and M. Backes. Tagvisor: A privacy advisor for sharing hashtags. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 287–296. International World Wide Web Conferences Steering Committee, 2018.

[52] X. Zheng, J. Han, and A. Sun. A Survey of Location Prediction on Twitter. In *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[53] A. Zubiaga, A. Voss, R. Procter, M. Liakata, B. Wang, and A. Tsakalidis. Towards Real-Time, Country-Level Location Classification of Worldwide Tweets. In *IEEE Transactions on Knowledge and Data Engineering*, 2017.