Are People Located in the Places They Mention in Their Tweets?  
A Multimodal Approach

Zhaomin Xiao  
University of North Texas  
zhaominxiao@my.unt.edu

Eduardo Blanco  
University of Arizona  
eduardoblanco@arizona.edu

Abstract

This paper introduces the problem of determining whether people are located in the places they mention in their tweets. In particular, we investigate the role of text and images to solve this challenging problem. We present a new corpus of tweets that contain both text and images. Our analyses show that this problem is multimodal at its core: human judgments depend on whether annotators have access to the text, the image, or both. Experimental results show that a neural architecture that combines both modalities yields better results. We also conduct an error analysis to provide insights into why and when each modality is beneficial.

1 Introduction

Twitter is a social network in which users post short messages known as tweets. While statistics vary depending on the source and publication time, official reports state that 187 million users logged in daily in the third quarter of 2020 (Twitter, 2020), and 500 million tweets were published worldwide on a daily basis in 2014—the last year the number was made public (Twitter, 2014). According to a recent report (Pew Research Center, 2019), 24% of all Americans use Twitter (45% between 18 and 24 years of age), and 46% of them use it at least once a day (26% more than once). Tweets contain not only text (including hashtags, links, emojis, etc.), but also multimedia content such as images and videos. Indeed, 42% of tweets have images (Lee, 2015), and marketing research reveals that having an image improves user engagement: 18% more click throughs, 89% more likes and 150% more retweets (Brandwatch, 2017).

When it comes to noisy user-generated content and spatial information, most previous work falls under two main topics: (a) named entity recognition (Baldwin et al., 2015) and disambiguation (Eshel et al., 2017), and (b) geolocation (Han et al., 2016). The former identifies, among others, location named entities and links them to a knowledge base without specifying who is there. The latter determines one location per user—even if it is not explicitly mentioned. For example, place of residence can be inferred, at least to a certain degree, from the locations of other users and language usage patterns. In this paper, we tackle a complimentary problem: to determine whether people are located in the places they mention in their tweets.

Extracting this kind of spatial information is challenging. First, people often mention places in their tweets even though they are not located there. Second, one must often rely on nuances in both the text and images to make a decision. Consider the tweets in Figure 1. The author of the tweet on the left was in Phoenix when the tweet was published. Note that the text alone could arguably be enough to conclude so, but the image provides additional evidence: the background is compatible with the Phoenix area (desert landscape, mountains, etc.), and the person in the picture is (most likely) enjoying the weather there during a short trip for Memorial Day. The author of the tweet on the right, on the other hand, was not in Atlanta when the tweet was published. In this example, the image together with the text provides evidence that the
author is working rather than enjoying Memorial Day in Atlanta with coworkers.

While the work presented here could be considered fundamental research, it opens the door to several applications. For example, emergency management systems could issue customized alerts to individuals who were, are, or are about to be located near a natural disaster. Similarly, eyewitness verification could benefit as the locations of people and the events they claim to witness must be compatible (within some temporal bounds).

The main contributions of this paper are: (a) a corpus of 6,540 tweets with annotations indicating whether the author was in the places mentioned in the tweets; (b) analysis demonstrating that this is a multimodal problem: the ground truth changes depending on whether annotators have access to the text, the image, or both; (c) experimental results showing that taking into account both modalities is beneficial; and (d) qualitative analysis providing insights into (d.1) when are the text and image beneficial, and (d.2) the remaining sources of errors.

1.1 Ethical Considerations

Determining where people are located has the potential to open the door to malicious (or just unwanted) tracking and surveillance. For example, applications that track location data may turn around and sell that data, revealing someone’s every movement—whether it is to a retail store, an abortion clinic, or a gay bar. Equally important, Twitter users may not be aware that their tweets can be used for research purposes (Fiesler and Proferes, 2018). We are not interested in tracking people or surveillance. Instead, we are interested in investigating the very definition of the problem and analyzing whether and how language and images complement each other.

In order to alleviate the issues above and preserve privacy, we implemented these safeguards. First, our corpus (a) only contains one tweet per user thus we do not enable user tracking or surveillance. Second, our analyses and experiments only take into account the text and image in a tweet—we do not take into account user information or any metadata. Third, we have designed a take-down request process via an online form following Mirowski et al. (2019).

2 Connections to Related Work

Extracting spatial information from social media and tweets in particular has received substantial attention (Zheng et al., 2018). For example, the tasks of named entity recognition (i.e., identifying, among others, location named entities mentioned in text) and disambiguation (i.e., linking named entities to entries in a knowledge base) have been explored in this noisy user-generated domain (Ritter et al., 2011; Baldwin et al., 2015; Shen et al., 2013; Eshel et al., 2017). Unlike us, these efforts do not aim at determining spatial information about authors of tweets. As we shall see, people often mention places where they are not located thus identifying and disambiguating locations tell us what places people tweet about—not the places where they are located when they tweet.

Geolocating twitter users consists in assigning one location to a user (e.g., place of residence). Existing corpora calculate the ground truth (i.e., the location for each user) from the geotags attached to tweets. For example, GeoText (Eisenstein et al., 2010) and Twitter-US (Roller et al., 2012) select the geotag of the first geotagged tweet from each user, and Twitter-World (Han et al., 2012) and W-NUT’16 (Han et al., 2016) select the majority city after mapping geotags to city centers. State-of-the-art models take as their input a user’s Twitter stream, and combine the text in the tweets, metadata and the social network structure with a neural architecture (Miura et al., 2017; Rahimi et al., 2017, 2018; Do et al., 2018). Unlike the work presented here, geolocating assigns one location per user thus it disregards that people participate in events and as a result their locations change. In this paper, we determine whether people are located in the places they mention in their tweets—even if they only mention the place once and regardless of how long and how often they are there.

More related to our work, Li and Sun (2014) determine whether people have visited, are currently at, or will soon visit points of interest (e.g., monuments, train stations). In their corpus, 47.3% of points of interest are invalid, resulting in little spatial information. More recently, Doudagiri et al. (2018) annotate whether people are located at the locations they tweet about (corpus size: 1,000 tweets), but they do not present experimental results. These two corpora were not publicly available at the time of writing. The work presented here complements these efforts. First, we target any city...
mentioned in a tweet, not predefined points of interest. Second, we show that both text and images must be taken into account. Indeed, the ground truth changes depending on which modalities annotators have access to, and experimental results show that models benefit from both modalities. Third, we release a new corpus of 6,540 tweets.

Finally, we note that coupling language and vision has been proposed for, among others, machine translation (Huang et al., 2016) and spatial role labeling (Kordjamshidi et al., 2017). Within social media, some examples include determining the relationship between text and images (Vempala and Preotiuc-Pietro, 2019), point-of-interest type prediction (Sánchez Villegas and Aletras, 2021), multimodal named entity recognition (Yu et al., 2020), named entity disambiguation (Moon et al., 2018), identifying fake news (Gupta et al., 2013), extracting possessions (Chinnappa et al., 2019), revealing demographic attributes (Sakaki et al., 2014), determining account types (Wijeratne et al., 2016), and detecting user groups (Balasuriya et al., 2016). Our work is inspired by these efforts, but to our knowledge we are the first to target spatial information about authors of tweets using both text and images.

3 A Corpus of Tweets and Spatial Information about the Authors

Our main goal is to understand what kind of spatial information one can infer between authors of tweets and the places they mention in their tweets. To our knowledge, we are the first to tackle this problem, so we create a new corpus. This allows us to explore whether human judgments change depending on whether annotators have access to the text, image or both (Section 4) as well as conduct experiments to automate the task (Section 5).

Collecting tweets We collected 10,000 tweets suitable for our purposes using the criteria below:

1. Each tweet contains both text and an image.
2. The text in each tweet:
   (a) is written in English and has at least five tokens;
   (b) mentions an event that occurred within 14 days of the tweet publication date; and
   (c) mentions a city.

We work with tweets that contain both text and images because we want to explore how spatial information depends on the interpretation of these modalities. We identify the language in which a tweet is written with langdetect\(^2\) and spaCy (Honnibal et al., 2020). The list of events we consider include the following: Christmas, Spring Break, Thanksgiving, Election Day, Labor Day, Memorial Day, and Veteran’s Day. Note that the Twitter search engine does not simply match keywords, thus small variations such as #veteransday are also matches. Finally, we use a list of the 100 most populous cities in the U.S.,\(^3\) This list includes large cities such as Los Angeles and Chicago as well as smaller cities such as Irving, TX and Richmond, VA (populations below 220,000).

We acknowledge that the events and cities we work with make our corpus US-centric. We believe, however, that the conclusions we reach are not US-centric. In particular, our analyses and experiments are not grounded on the specific events or cities that we work with. A corpus that covers all countries and events—assuming that doing so is possible—is outside the scope of this paper.

Annotation guidelines We aim at capturing spatial information intuitively understood by humans. To this end, we crowdsource human judgments from non-experts by asking a simple question. More specifically, we show crowdworkers one tweet at a time and ask them “Was the author of the tweet located in city when the tweet was published?,” where city is one of the cities identified in the tweet during the collection process. Crowdworkers choose between two options:

- yes: the author of the tweet was in city when the tweet was published; or
- no: I cannot tell if the author of the tweet was in city when the tweet was published.

Note that no does not guarantee that the author was not in city, it rather indicates that the crowdworker cannot establish that the author was in city.

3.1 Annotation Process

We crowdsource annotations using Amazon Mechanical Turk. The annotation interface includes instructions and examples. Crowdworkers provide answers to the question above for one (tweet, city) pair before moving to the next one. The interface

\(^2\)https://github.com/Mimino666/langdetect
\(^3\)https://gist.github.com/Miserlou/11500b2345d3fe850c92
displays a screenshot of the tweet as shown on the Twitter’s website (desktop version). Doing so ensures that special characters, symbols, and images are displayed properly.

We collected annotations in three independent phases: showing annotators (a) the original tweet (text and image) (b) only the text, and (c) only the image. There was no overlap between the crowdworkers involved in each phase to avoid potential biases. For example, we avoid the possibility that a crowdworker remembers the image in the original version of the tweet when the interface only displays the text. The three annotation phases allow us to analyze whether crowdworkers understand different spatial information if they cannot see the text or image in the original tweet. We created 30,000 annotation tasks (Human Intelligence Tasks in Mechanical Turk parlance; 3 versions per tweet), and crowdsourced five annotations for each. The hourly pay ranges from $9 to $13 (the US federal minimum wage is $7.25).

3.2 Annotation Quality

Ensuring annotation quality is critical in any crowdsourcing effort. Our first defense is to recruit crowdworkers located in the United States and with previous approval rate above 95%. Additionally, we do not allow workers to continue working on our tasks if the average completion time per Human Intelligence Task in the past (i.e., the average time spent prior to submitting) is under 3 seconds. We decided on the minimum time required to complete our task based on observations during pilot annotations.

Our second defense is to collect five annotations per Human Intelligent Task and filter out bad annotations until we obtain substantial inter-annotator agreement. We do so using Multi-Annotator Competence Estimation (Hovy et al., 2013, MACE) and Krippendorff’s $\alpha$ (Krippendorff, 2011). MACE is designed to rank annotators by their competence scores assessing their reliability. The adjudicated labels are determined based on these scores—the most frequent label is not always a good option. Krippendorff’s $\alpha$ is a coefficient indicating inter-annotator agreement when several annotators complete different annotation tasks, as is common in crowdsourcing. $\alpha = 0$ indicates only the agreement expected by chance, and $\alpha = 1$ indicates that annotators always agree. Krippendorff’s $\alpha$ at or above 0.6 are considered substantial, and above 0.8 (nearly) perfect (Artstein and Poesio, 2008).

We ensure $\alpha \geq 0.6$ as follows:

1. Calculate the MACE score of all crowdworkers and sort them by decreasing MACE score.
2. While Krippendorff’s $\alpha < 0.6$:
   (a) Drop all the annotations by the crowdworker with the lowest MACE score.
   (b) If a Human Intelligent Task is left without annotations, republish it.

We republish Human Intelligent Tasks (Step 2b) at most twice in order to keep the crowdsourcing costs reasonable. The final corpus consists of 6,540 annotated tweets with Krippendorff’s $\alpha = 0.61$. In the rest of this paper, we work with these tweets.

4 Corpus Analysis

The 6,540 tweets in our corpus mention 96 unique cities. The most frequent cities are Miami (17% of tweets) and Chicago (6%); other cities account for at most 5% of tweets each. The tweets mention all the events we target (Section 3). The most common event is Spring Break (37% of tweets) followed by Memorial Day (27%). Other events account for between 5% and 10% of tweets except Election Day, which accounts for 3% of tweets.

4.1 Do labels depend on the information available to crowdworkers?

Yes, crowdworkers understand substantially different spatial information depending on whether we show them the original tweet (text and image), the text only, or the image only. The label distribution is as follows for each combination:

<table>
<thead>
<tr>
<th>Information Available</th>
<th>Text Yes (%)</th>
<th>Text No (%)</th>
<th>Image Yes (%)</th>
<th>Image No (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text + Image</td>
<td>74</td>
<td>26</td>
<td>91</td>
<td>9</td>
</tr>
<tr>
<td>Only Text</td>
<td>80</td>
<td>19</td>
<td>81</td>
<td>19</td>
</tr>
<tr>
<td>Only Image</td>
<td>72</td>
<td>28</td>
<td>81</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 1: Percentage of label changes depending on the information available to annotators. Many labels change if the text or image is unavailable, especially if the label when both are available is no (72% and 81%).

Note that the right label (i.e., the ground truth) is the one obtained when crowdworkers have access
Figure 2: Examples of annotations depending on what information is available to annotators (text, image, or both text and image). We only show the adjudicated label after adjudicating the crowdsourced labels. Annotations change substantially (Table 1); the image or text alone often misleads annotators (examples on the left and middle).

Figure 3: Neural network for determining whether people are located in the places they mention in their tweets. The output layer combines individual representations of the text and image (top and bottom, text and image component), and a joint representation of the text and image (middle, text_image component).

Table 1 shows that when the right label is yes and we only show the text or image, crowdworkers most often do not change the label (74% and 91% respectively). When the right label is no, however, it is usually the case that crowdworkers are tricked when they are shown only the text or image (72% and 81% respectively). These percentages demonstrate that both the text and image must be taken into account to determine spatial information about the author of a tweet.

We show examples of annotation changes in Figure 2. In the example on the left, the image alone is insufficient to make any spatial inference between the author of the tweet and Miami. Indeed, it is hard to make any connection between Miami and the basketball court. The text alone (“Happy Spring Break from Miami”), however, is enough to understand that the author is in Miami. The tweet in the middle exemplifies how not having access to the image can trick annotators. When crowdworkers only have access to the text, they understand that the author was in Raleigh celebrating Memorial Day. When they are also shown the image, however, they realize that it is an advertisement and do not conclude that the author is in Raleigh. The tweet on the right shows an example in which the annotations do not change regardless of whether crowdworkers have access to the text, image, or both. The text indicates that the author was in Chicago (“Spring break in the chilly Chicago weather”), and the image also facilitates the same conclusion (cold weather, Cloud Gate in Chicago). Showing both the text and images provides further evidence to conclude that the author was in Chicago when the tweet was published.

5 Experiments and Results

Armed with our corpus (Section 3), we conduct experiments to automatically determine whether authors of tweets are located in the cities they mention in their tweets. We reduce the problem to a classification task. The input to the model is a (tweet, city) pair, and the output is a label indicating
whether the author of the tweet was located in the city when the tweet was published (yes or no). We create stratified training and test splits (80%/20%), and reserve 20% of the training split for validation. If the tweet includes more than one image (it only applies to a handful of tweets), we only feed to the classifier the first image. Our models do not take into account network or user information. They make predictions based exclusively on the content of tweets (the text and image).

### Neural Network Architecture

We build a neural network consisting of three main components (Figure 3): a component to represent the text (top), a component to represent the image (bottom), and a component to jointly represent the text and image (middle). The three components use pre-trained neural networks combined with a trainable fully connected layer to reduce the dimensionality of each representation individually (size: 512). Then, we concatenate the three representations (size: $3 \times 512 = 1536$) and apply two trainable fully connected layers (sizes: 512 and 2) to make the final prediction (yes or no). We use dropout (Srivastava et al., 2014) in the second-to-last fully connected layer (rate: 0.2). We tried different sizes for the fully connected layers during the tuning process, but we did not observe benefits.

The text component is BERT (Devlin et al., 2019) and the image component is VGG16 (Simonyan and Zisserman, 2014). We use the pre-trained models released by HuggingFace (Wolf et al., 2020) and Pytorch (Paszke et al., 2019). We train the neural network for up to 100 epochs using the Adam optimizer (Kingma and Ba, 2014), categorical cross entropy as the loss function, and batch size 8. We stop the training process before 100 epochs if there is no improvement in the validation set for 10 epochs. We implement the neural network with PyTorch (Paszke et al., 2019).

### Results

Table 2 shows the results with the test split using several variations of the neural network: only the text component, only the image component, only the text_image component, and all of them. We observe that the three components by themselves obtain roughly the same results (F1: 0.64–0.65). Combining the three components, however, yields a slightly higher F1 (0.68), which is mostly due to an increase in Recall (0.74 vs. 0.65–0.68). These results show that the three components of the network are beneficial. In particular, incorporating the individual representations for the text and image in addition to the joint representation (text_image) is beneficial.

### Qualitative Analysis

To better understand why and when the text and image are most beneficial, we perform a qualitative analysis of the errors made by each model. More specifically, we answer the following questions:

- When does the image complement the text?
- When does the text complement the text?
- When does the task remain challenging?

#### When does the image complement the text?

We start the qualitative analysis providing insights into when is the image beneficial to solve the task. Table 3 exemplifies the most common errors made by the text component that are fixed by the full network (text + image + text_image).

The most frequent error that benefits from taking into account the image (38%) occurs when the image (apparently) does not have a connection with the location at hand. Instead, it (visually) depicts some event that (a) occurred in the location at hand and (b) is mentioned in the text. Consider the example on the left (Table 3). The text is about tornadoes in Miami, but the image is not a common Miami scene—it shows the destruction caused by the tornado. The text component alone is unable to make the connection, but the full network makes the connection and predicts that the author was in Miami when the tweet was published.

The second most common error fixed by the full network (31%) occurs when the tweet is an advertisement and the text component alone wrongly predicts yes (e.g., middle tweet in Table 3). In this case, taking into account the image allows the full network to identify the tweet as an advertisement and predict no. We note that crowdworkers generally annotate advertisements with no unless there is a connection between the author of the tweet and the location (e.g., My Orlando Chapter Got Something For Ya! […] , right tweet in Table 4).
The third most common error that benefits from taking into account the image (14%) occurs when the image depicts a typical scene of the location at hand. For example, in the right tweet in Table 3, the picture depicts (presumably) Jacksonville beach.

**When does the text complement the image?**

We continue the qualitative analysis providing insights into when is the text beneficial to solve the task. Table 4 exemplifies the most common errors made by the image component that are fixed by the full network (text + image + text_image).

The most frequent error (46%) occurs when (a) the image could have been taken in several places and (b) the text describes an event that occurred in the location at hand and is depicted in the image. The tweet on the left (Table 4) exemplifies this scenario. Indeed, the indoor picture could have been taken in many indoor spaces, but it shows an event described in the text (i.e., the Kids Camp).

The second most common error fixed by the full network (27%) occurs when (a) the image is compatible with the location at hand and (b) the text provides further evidence that the author was there. Consider the middle tweet in Table 4. The model that takes into account only the image fails to identify that the author was in Miami. Taking into account the text (“My city better than yours! Period #Miami […]”), however, allows the full network to make the right prediction (yes).

The third most common error (10%) addressed

<table>
<thead>
<tr>
<th>Location: Miami</th>
<th>Location: Houston</th>
<th>Location: Jacksonville</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold: yes, Predicted&lt;sub&gt;image&lt;/sub&gt;: no</td>
<td>Gold: no, Predicted&lt;sub&gt;ext&lt;/sub&gt;: yes</td>
<td>Gold: yes, Predicted&lt;sub&gt;image&lt;/sub&gt;: no</td>
</tr>
</tbody>
</table>

Table 3: Most common errors fixed by the full network compared to the network that only uses the text component.

<table>
<thead>
<tr>
<th>Location: Arlington</th>
<th>Location: Miami</th>
<th>Location: Orlando</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold: yes, Predicted&lt;sub&gt;image&lt;/sub&gt;: no</td>
<td>Gold: yes, Predicted&lt;sub&gt;image&lt;/sub&gt;: no</td>
<td>Gold: yes, Predicted&lt;sub&gt;image&lt;/sub&gt;: no</td>
</tr>
</tbody>
</table>

Table 4: Most common errors fixed by the full network compared to the network that only uses the image component.
Table 5: Most common errors made by the full network (comparing with the ground truth).

<table>
<thead>
<tr>
<th>Location: Arlington</th>
<th>Location: Chicago</th>
<th>Location: Miami</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold: no, Predicted\textsuperscript{full}: yes</td>
<td>Gold: no, Predicted\textsuperscript{full}: yes</td>
<td>Gold: yes, Predicted\textsuperscript{full}: no</td>
</tr>
</tbody>
</table>

Table 5: Most common errors made by the full network (comparing with the ground truth).

when the text is taken into account are again advertisements. As is usual with screenshots and advertisement, the image component alone predicts no. Taking into account the text allows the full network to realize that the author most likely was in Orlando (My Orlando Got Something for Ya! […]).

Which tweets remain challenging? We close the qualitative analysis with the most common errors made by the full network (Table 5). To do so, we look at the errors made by the full network. (text + image + text_image).

The most common error (56%) occurs when (a) neither the text nor image contains enough information to determine whether the author was in the location at hand, and (b) crowdworkers annotated the tweet with no. Consider the left tweet in Table 5. Crowdworkers did not indicate that the author was in Arlington (no), as there is no evidence that the author was there when the tweet was published. We hypothesize that the full network makes a connection between the flags in the image and “all those flags” from the text, and as a result, it predicts yes.

The second most common error (24%) are again advertisements. Consider the middle tweet in Table 5. Neither the text or image provide much evidence of the author being in Chicago, as indicated by the crowdworkers. The full network, however, predicts yes, most likely because it recognizes an urban environment in the picture.

Finally, the full network struggles when the text is not a complete sentence and the connection between text and image is rather nuanced. For example, the text in the tweet on the right (Table 5) is a sentence fragment, and the picture depicts a fight in a beach. The full network is unable to make the connection between (a) Miami and “the beach,” and (b) the fight and the sentence fragment (“Knuckle Up: On Today’s Episode of Spring Break […]”).

7 Conclusions

We have introduced the task of determining whether people are located in the places mentioned in their tweets. Going beyond named entity recognition and disambiguation, this problem is about figuring out whether the authors of tweets are located in the places mentioned their tweets. Our new corpus (6,540 tweets) shows that people often mention cities in their tweets even though they are not located there (48.9% of city mentions)—or at least there is not enough evidence in the tweet for crowdworkers to conclude so.

Importantly, we have shown that human judgments change substantially depending on whether crowdworkers have access to the text, the image, or both. These changes in human judgments indicate that when it comes to understanding spatial information about the authors of tweets, the text and images complement each other. To our knowledge, our corpus (Krippendorf’s $\alpha = 0.61$) is the first to tackle this challenging problem.

Experimental results show that the task can be automated although our neural network obtains
modest results. In particular, coupling independent representations of the text and image (2 representations) with a joint representation of the text and image yields the best results. These empirical results mirror the observation that human judgments change depending on which modalities crowdworkers have access to. We have also presented a qualitative analysis providing insights into how the image and text complement each other. In summary, they are usually beneficial if they provide additional details about the location at hand or an event that occurred in the location at hand.

Acknowledgements

This research was made possible by funding from the NGA University Research Initiatives (NURI) and a Bloomberg Data Science Research Grant.

References


