

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 (Es-hel et al., 2017), and (b) geolocation (Han et al., 2016). The former identifies, among others, loca-

Spectacular Memorial Day
Phoenix weather at 20°F below
the norm! <https://t.co/qFJsH0zHtC>



when my place of employment
tell me it's time to go back to
work and I know my coworkers
went to Atlanta for Memorial Day
<https://t.co/hPoHPkFFh6>



Figure 1: Examples of tweets in which the author is and is not located in the place mentioned in the tweet (*Phoenix* (left) and *Atlanta* (right) respectively).

tion 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:¹ (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).

¹Corpus and code available at https://github.com/zhaomin1995/coling2022_repo

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 (Hon-nibal 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 crowdsource 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 α (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 α 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 α at or above 0.6 are considered *substantial*, and above 0.8 (nearly) *perfect* (Artstein and Poesio, 2008).

		text		image	
		yes	no	yes	no
text + image	yes	74	26	91	9
	no	72	28	81	19

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%).

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:

- text and image: *yes*: 51.09%, *no*: 48.91%
- only text: *yes*: 80.93%, *no*: 19.07%
- only image: *yes*: 69.74%, *no*: 30.26%

Note that the *right* label (i.e., the ground truth) is the one obtained when crowdworkers have access

	P	R	F1
text	0.65	0.66	0.65
image	0.64	0.65	0.64
text_image	0.62	0.68	0.65
text + image + text_image	0.64	0.74	0.68

Table 2: Results obtained with the full network (text + image + text_image) and individual components. Taking into account the three representations is beneficial.

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.

6 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 image?
- 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).




(38%) Image depicts key event	(31%) Advertisements	(14%) Image depicts location
<p>#MIAMI TORNADOES: Some workers are paying it forward as paychecks continue even after businesses remain closed after the Memorial Day storms</p> 	<p>Perfect time to gear up for summer travel #tumi #tumitravel #travel #summervacation #springbreak #houstonpremiumoutlets</p> 	<p>CORONAVIRUS: The spread of the #coronavirus did not appear to dampen spring break plans for a lot of people at #Jacksonville beach today @ActionNewsJax @WOKVNews</p> 
<p>Location: <i>Miami</i> Gold: <i>yes</i>, Predicted_{text}: <i>no</i></p>	<p>Location: <i>Houston</i> Gold: <i>no</i>, Predicted_{text}: <i>yes</i></p>	<p>Location: <i>Jacksonville</i> Gold: <i>yes</i>, Predicted_{text}: <i>no</i></p>

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

(46%) Text describes key event	(27%) Text describes location	(10%) Advertisements
<p>Arlington Spring Break Kids Camp Gets Wild https://t.co/qrFqcClRqH https://t.co/Jp5Wh1d2zb</p> 	<p>My city better than yours! Period #Miami #MemorialDay https://t.co/OwZ4ibAw7r</p> 	<p>🔥 My Orlando Chapter Got Something For Ya! ATL Memorial Day Weekend Flash Sale Orlando Members Only! 🇺🇸🇩🇪 Get Them Before There Gone! 🍷</p> 
<p>Location: <i>Arlington</i> Gold: <i>yes</i>, Predicted_{image}: <i>no</i></p>	<p>Location: <i>Miami</i> Gold: <i>yes</i>, Predicted_{image}: <i>no</i></p>	<p>Location: <i>Orlando</i> Gold: <i>yes</i>, Predicted_{image}: <i>no</i></p>

Table 4: Most common errors fixed by the full network compared to the network that only uses the image component.

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




(56%) Missing Information	(24%) Advertisements	(8%) Sentence Fragments
<p>Ever wondered what happens with all those flags you see in Arlington on #MemorialDay? https://t.co/GYHIGdtXaq https://t.co/qwaAC4NPLu</p> 	<p>50% Off Memorial Day Fireworks Cruise Tickets at 200 N. Breakwater Access #chicago #memorial #fireworks #cruise #tickets #breakwater #access https://t.co/EIYj2JZiwV https://t.co/r6GaWQFSb6</p> 	<p>Knuckle Up: On Today's Episode of Spring Break Miami (VIDEO) - https://t.co/tEJqwwFqBr https://t.co/8bX9juZ3CK</p> 
<p>Location: <i>Arlington</i> Gold: no, Predicted_{full}: yes</p>	<p>Location: <i>Chicago</i> Gold: no, Predicted_{full}: yes</p>	<p>Location: <i>Miami</i> Gold: yes, Predicted_{full}: no</p>

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 be-

tween 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

- Ron Artstein and Massimo Poesio. 2008. [Inter-coder agreement for computational linguistics](#). *Comput. Linguist.*, 34(4):555–596.
- Lakshika Balasuriya, Sanjaya Wijeratne, Derek Doran, and Amit Sheth. 2016. Finding street gang members on twitter. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 685–692. IEEE Press.
- Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. [Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition](#). In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.
- Brandwatch. 2017. 45 Incredible and Interesting Twitter Statistics. <https://www.brandwatch.com/blog/44-twitter-stats/>. Accessed: December 3, 2018.
- Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco. 2019. [Extracting possessions from social media: Images complement language](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 663–672, Hong Kong, China. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. 2018. [Twitter user geolocation using deep multiview learning](#). In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6304–6308. IEEE.
- Vivek Reddy Doudagiri, Alakananda Vempala, and Eduardo Blanco. 2018. [Annotating if the authors of a tweet are located at the locations they tweet about](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. 2010. [A latent variable model for geographic lexical variation](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.
- Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. [Named entity disambiguation for noisy text](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.
- Casey Fiesler and Nicholas Proferes. 2018. [“participant” perceptions of twitter research ethics](#). *Social Media + Society*, 4(1):2056305118763366.
- Aditi Gupta, Hemank Lamba, Ponnuram Kumaraguru, and Anupam Joshi. 2013. [Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy](#). In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM.
- Bo Han, Paul Cook, and Timothy Baldwin. 2012. [Geolocation prediction in social media data by finding location indicative words](#). In *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India. The COLING 2012 Organizing Committee.
- Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. [Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text](#). In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217, Osaka, Japan. The COLING 2016 Organizing Committee.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. [Learning whom to trust with MACE](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association*

- for *Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.
- Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. [Attention-based multimodal neural machine translation](#). In *Proceedings of the First Conference on Machine Translation*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2014. [Adam: A method for stochastic optimization](#). Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
- Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017. Clef 2017: Multimodal spatial role labeling (msprl) task overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 367–376, Cham. Springer International Publishing.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability. Retrieved from https://repository.upenn.edu/asc_papers/43.
- Kevan Lee. 2015. What analyzing 1 million tweets taught us. <https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/>. Accessed: December 3, 2018.
- Chenliang Li and Aixin Sun. 2014. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 43–52. ACM.
- Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2019. [The streetlearn environment and dataset](#).
- Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. [Unifying text, meta-data, 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)*, pages 1260–1272, Vancouver, Canada. Association for Computational Linguistics.
- Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. [Multimodal named entity disambiguation for noisy social media posts](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Pew Research Center. 2019. Social media use. <https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/>.
- Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. [A neural model for user geolocation and lexical dialectology](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 209–216, Vancouver, Canada. Association for Computational Linguistics.
- Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. [Semi-supervised user geolocation via graph convolutional networks](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2019, Melbourne, Australia. Association for Computational Linguistics.
- Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. [Named entity recognition in tweets: An experimental study](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. [Supervised text-based geolocation using language models on an adaptive grid](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510, Jeju Island, Korea. Association for Computational Linguistics.
- Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. 2014. Twitter user gender inference using combined analysis of text and image processing. In *Proceedings of the Third Workshop on Vision and Language*, pages 54–61.
- Danae Sánchez Villegas and Nikolaos Aletras. 2021. [Point-of-interest type prediction using text and images](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7785–7797, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2013. [Linking named entities in tweets with knowledge base via user interest modeling](#). In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 68–76, New York, NY, USA. ACM.
- Karen Simonyan and Andrew Zisserman. 2014. [Very deep convolutional networks for large-scale image recognition](#). *CoRR*, abs/1409.1556.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. [Dropout: A simple way to prevent neural networks from overfitting](#). *Journal of Machine Learning Research*, 15(56):1929–1958.
- Twitter. 2014. The 2014 #yearontwitter. https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html.
- Twitter. 2020. Q3 2020 letter to shareholders. https://s22.q4cdn.com/826641620/files/doc_financials/2020/q3/Q3-2020-Shareholder-Letter.pdf.
- Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. [Categorizing and inferring the relationship between the text and image of Twitter posts](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy. Association for Computational Linguistics.
- Sanjaya Wijeratne, Lakshika Balasuriya, Derek Doran, and Amit Sheth. 2016. Word embeddings to enhance twitter gang member profile identification. *arXiv preprint arXiv:1610.08597*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. [Improving multimodal named entity recognition via entity span detection with unified multimodal transformer](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.
- Xin Zheng, Jialong Han, and Aixin Sun. 2018. A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671.