Context Helps Determine Spatial Knowledge from Tweets

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

Yan Huang
University of North Texas
yan.huang@unt.edu

Eduardo Blanco
University of Arizona
eduardoblanco@arizona.edu

Abstract

This paper introduces the problem of determining whether people are located in the locations they mention in their Twitter streams. In particular, we investigate the role of context—tweets published before and after a tweet mentioning a location—in this challenging problem. We present a new corpus of Twitter streams with spatial information. Our analyses show that context is key to define the ground truth, as human judgments depend on whether annotators have access to context. Experimental results show that a neural architecture that takes into account context in addition to the tweet mentioning a location yields better results. We also conduct an error analysis to provide insights into the errors made by our best model.

1 Introduction

Information extraction has been popular since decades ago (Cardie, 1997). One of its subareas is geotagging (Cheng et al., 2010; Mahmud et al., 2021; Jurgens et al., 2021), which aims at adding geographical information to an entity. A common example of geotagging is to predict the home location of Twitter users (Elmongui et al., 2015). Geotagging users with their home location is challenging because people change their location often thus they may not tweet from their home location.

Twitter is a social network in which users post short messages known as tweets. Business reports state that the number of daily active users went from 166 million in April 2021 to 229 million in April 2022 (Kemp, 2022). According to a recent report (Dean, 2021), (a) the Twitter app was downloaded over 6 million times in the fourth quarter of 2020, and (b) Twitter users in the US spend on average 158.2 minutes per month on the app. In addition, Beveridge (2021) reports that more than 500 million tweets are published per day.

Users rarely geotag their tweets with their GPS coordinates. Previous studies estimate that only 0.85% of tweets are geotagged (Sloan et al., 2013). Alternatively, tweets mentioning a city can still be rich in spatial information. Mentioning a city, however, does not guarantee that the author was in the city when he tweeted—people often tweet about a city from elsewhere (Han et al., 2016).

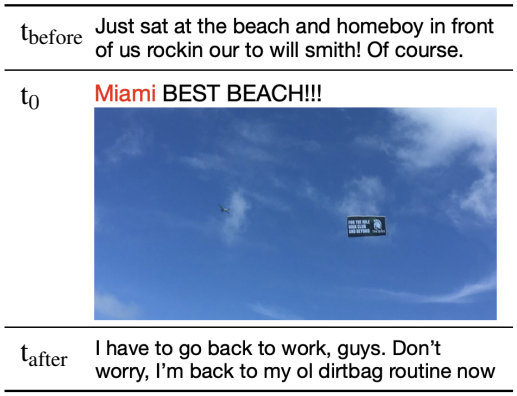

Figure 1: Tweet mentioning *Miami* (t_0) and the tweets published before and after (t_{before} and t_{after}). t_0 by itself is insufficient to determine whether the author was in *Miami*: (a) based on the text, the author could be tweeting about *Miami* from anywhere, and (b) the picture of the sky could be from almost anywhere. Considering context—the tweets published before and after—provides enough evidence to conclude that the author was in *Miami* when he published t_0 as he was there on vacation and eventually had to go back to work.

In this paper, we explore the problem of determining whether people are located in the cities they mention in their tweets. More specifically, we explore the role of context—tweets published before and after the tweet that mentions a city—in this challenging problem. Consider the three tweets in Figure 1, which were published in chronological order from top to bottom. Taking into account only the tweet mentioning *Miami* (t_0), it is unsound to conclude that the author was there when he published this tweet. Indeed, the text only states that *Miami* has the *best beach*, and the picture of the sky could be from almost anywhere. The tweets pub-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 149–160
November 1–4, 2023.

lished before and after, however, provide enough context to conclude that the author was on vacation in *Miami* (*sat in the beach*, *have to go back to work*, etc.) thus he was in *Miami* when he tweeted t_0. Note that t_0 is the only tweet mentioning *Miami* yet taking into account context is necessary to correctly determine whether the author was there.

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, location information can be utilized to facilitate transportation planning to provide positive long-term and sustainable economic impacts on many regions.

The main contributions of this paper are:[1] (a) a corpus of 3,494 tweets mentioning a city (t_0), their context (3 tweets before and 3 tweets after), and annotations indicating whether the author was in the city when he published t_0; (b) analysis showing that the problem requires taking into account context—annotations change depending on whether crowdworkers have access to the context; (c) experimental results showing that taking into account tweets published both before and after the tweet mentioning the city is beneficial; and (d) qualitative analysis providing insights into the errors made by our best model, including characteristics of both t_0 and the tweets in the context.

2 Related Work

Previous works within information extraction have targeted either spatial or temporal information. Works targeting spatial information extraction include event localization (Pustejovsky, 2013), and geolocation prediction (Mousset et al., 2020; Chong and Lim, 2017). Works targeting temporal information extraction include event ordering (Naik et al., 2019; Cassidy et al., 2014; Ning et al., 2020) and event understanding (Ma et al., 2021). In some tasks, such as human interaction prediction (Ke et al., 2016) and facial expression recognition (Zhang et al., 2017), spatial and temporal information are complementary to each other. In these works, spatial/temporal information is used as contextual information.

Various kinds of context have been proved useful in many tasks. Yu et al. (2022) investigate the role of conversational context in annotation and detections of hate speech and counter-hate speech. Wang et al. (2015) use topic, history, and conversation information as context to detect sarcasm in Twitter. Ren et al. (2016) and Vanzo et al. (2014) use similar context to approach the task of sentiment classification. Similar to their works, we also use history tweets as the context. Therefore, the temporal information contained in history tweets is used as the context to extract spatial information.

Most previous works targeting spatial information extraction are centered around events (Mousset et al., 2020; Chong and Lim, 2017; Frisoni et al., 2021). Unlike us, these efforts do not aim at determining spatial information about people. As we shall see, people often mention places despite they are not located there. In addition, knowing the location where an event occurs does not necessarily indicate the location of the participants in the event or people talking about the event. For example, one can state that there is a housing market bubble somewhere without ever stepping foot there. Finally, we note that people change their location often and most events are much shorter than a human lifespan.

More related to our work, Fuchs et al. (2013) analyze personal behavioral patterns using geotagged tweets from the Seattle, Washington area (i.e., tweets published with geographical coordinates within Seattle). Since only 0.85% of tweets are geotagged (Sloan et al., 2013), their work may suffer from lack of generality. We work with tweets mentioning cities across the United States without requiring a geotag. Doudagiri et al. (2018) annotate whether people are located at the locations they tweet about. Their corpus is not publicly available at the time of writing, and experimental results are not presented. The paper presented here is complementary. Xiao and Blanco (2022) work on a task similar to ours, though they investigate the role of modality in the annotation and learning method of spatial information extraction. Differently, we show that context—tweets published before and after—must be taken into account.

3 A Corpus of Tweets and Spatial Knowledge

Our main goal is to investigate whether contextual information is beneficial to infer spatial information between authors of tweets and the locations they mention in their tweets. To our knowledge, we

[1] Corpus and code available at https://github.com/zhaomin1995/aacl2023_repo

are the first to tackle this problem, so we create a new corpus. Doing so allows us to analyze whether human judgments depend on whether they have access to the context (i.e., tweets published before and after the tweet mentioning a location).

Collecting Tweets While one could start with a new collection of tweets, we choose to build upon an existing corpus with spatial annotations disregarding context (Xiao and Blanco, 2022). This corpus contains 6,540 English tweets mentioning a city and crowdsourced annotations indicating whether the author was in the city when he tweeted *based on the content of the tweet and nothing else*. Building upon this corpus allows us to (a) leverage their context-unaware annotations and (b) focus on our main goal: to investigate the role of context.

We refer to the tweets in the existing corpus as *target tweets*. We complement target tweets with the three tweets published immediately before and after within a window of 90 days[2]. We refer to these tweets as *earlier tweets* and *later tweets* respectively. We also refer to both *earlier tweets* and *later tweets* as *context tweets*. An instance thus contains seven tweets: one target tweet mentioning a city, three earlier tweets, and three later tweets (we discarded target tweets without enough context).

Annotation Guidelines We aim at capturing spatial information intuitively understood by humans. To this end, we crowdsource annotations from non-experts asking two questions. The first question is "Was the author of the tweet located in *city* when the target tweet was published?". Crowdworkers choose between two options:

- `yes`: the author of the tweets was located in *city* when the target tweet was published; or
- `no`: I cannot tell whether the author of the tweets was located in *city* when the target tweet was published.

Note that `no` does not guarantee that the author was not in *city*. It rather indicates that the crowdworkers cannot establish that the author was in *city*. The results of our pilot annotation show that annotators cannot determine whether the author of the tweet was in *city* in more than 95% of the instances that are not labeled as `yes`.

The second question is "How confident are you about your answer to the first question?" Crowdworkers choose options from a modified Likert scale with five options: *Extremely confident*, *Very confident*, *Moderately confident*, *Slightly confident*, and *Not confident at all*. The second question complements the first question and allows annotators to bridge the semantic gap between `yes` and `no` answers to the first question. During pilot annotations designed to refine the annotation guidelines, we discovered that asking these two questions results in more reliable annotations than, for example, including *probably yes* and *probably no* as options for the first question and skipping the second question.

Annotation Interface and Process We crowdsource annotations on Amazon Mechanical Turk. Crowdworkers answer the questions above for one instance (the target tweet mentioning a city, three earlier tweets, and three later tweets) before moving to the next instance. In order to ensure crowdworkers read the tweets in the order they were published, the interface displays the earlier tweets, the target tweet, and the later tweets with short delays in between after workers click on the appropriate button. We display screenshots of tweets from the Twitter's website to ensure that all characters, symbols, emojis, and images are correctly displayed.

A total of 329 annotators participated in our annotation task. The hourly pay for crowd workers ranged from $9 to $13 (the federal minimum wage is $7.25). After collecting multiple annotations for each instance (Section 3.1), we replace `yes` labels with `no` if the confidence level is lower than or equal to *Moderately confident*. Our rationale is that `yes` with low confidence level indicates that there is not enough evidence about the author being located in *city* when the target tweet was published, which is actually the definition of `no`.

3.1 Annotation Quality

Ensuring annotation quality is critical in any crowdsourcing effort. Our first defense is to recruit crowdworkers located in the United States with an approval rate above 95%. Second, we do not allow workers to continue to work on our task if the average completing time per Human Intelligence Task in the past (i.e., the average time spent before submitting answers to both questions per instance) is under 8 seconds. We decided on 8 seconds based on observations during pilot annotations—one just cannot read seven tweets and answer two questions in eight seconds. In addition, we split all instances into batches of 500 instances and publish one batch every 8 hours to avoid annotator fatigue.

Our last defense is to collect five annotations per

[2]The 90 days is the maximal time window. If there are more than three tweets within 90 days, we choose the closest three. These seven tweets are published by the same user.

instance and filter out unreliable annotations until we obtain substantial inter-annotator agreement for each batch. We do so using Multi-Annotator Competence Estimation (Hovy et al., 2013, MACE) and Krippendorff's α (Krippendorff, 2011). MACE ranks annotators by their competence scores and adjudicates labels according to these scores. Krippendorff's α is used to measure inter-annotator agreement. $\alpha = 0$ indicates only the agreement expected by chance, while $\alpha = 1$ indicates annotators always agree. Krippendorff's α coefficients at or above 0.6 are considered substantial and above 0.8 (almost) perfect (Artstein and Poesio, 2008). We repeat these steps for each batch until $\alpha \geq 0.6$:

1. Use MACE to calculate competence scores for workers and sort them in descending order.
2. Discard all the annotations by the crowdworker with the lowest score.
3. Calculate the Krippendorff's α coefficient on the remaining annotations.

We discard instances left without annotations after the above steps. The final corpus consists of 3,494 annotated instances with Krippendorff's $\alpha = 0.76$. In the rest of this paper, we work with these instances. We reserve for future work a thorough analysis of the instances with the most disagreements between annotators.

3.2 Ethical considerations

Determining where people are located has the potential to open the door to malicious (or 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 or potential malicious users such as stalkers. 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 context complements target tweets. In order to alleviate the issues above, we implemented the following safeguards:

- Our corpus only contains seven tweets per user published within 90 days thus tracking and surveillance are not enabled by this work. Neither user information nor any metadata is included in the public corpus.
- Our analyses and experiments only take into account the text and image in a tweet. We do not consider user information or any metadata.
- We have designed a takedown request process following Mirowski et al. (2019). People can request us to delete tweets via an online form. We will propagate takedown requests to researchers who download the corpus and the license will require them to delete them.

		after adding context (our annotations)	
		`yes`	`no`
before adding context (original annotations)	`yes`	74.5	25.5
	`no`	60.7	39.3

Table 1: Label percentages depend on whether annotators have access to context (i.e., tweets published before and after). Many labels change if annotators have access to context, especially if the label is `no` (60.7%) before adding context. The *correct* ground truth labels are the ones collected with context.

4 Corpus Analysis

The 3,494 instances in our corpus mention 94 unique cities. The most common cities are *Miami* (17% of target tweets) and *Chicago* (6%); other cities account for at most 5% each. The annotation process results in the following label distribution: Most instances are annotated `yes` (67.7%), and around one-third are annotated `no` (32.3%).

Do labels depend on whether context is available to crowdworkers? Yes, crowdworkers understand substantially different spatial information depending on whether they have access to context. Table 1 compares our annotations and the original annotations that do not consider context. Note that the *correct* ground truth labels are the ones obtained taking into account context.

Most target tweets annotated `no` without considering context are annotated `yes` considering context (60.7%). The percentage of annotation changes for target tweets annotated `yes` without context is smaller but substantial (25.5%). In other words, figuring out whether the author of a tweet is in the city he tweets about requires looking at context.

We show examples of annotation changes in Figure 2. In the first example, annotators understood that the *target tweet* is a news report about *Philadelphia*. As a result, annotators could not determine whether the author was located in *Philadelphia* when the *target tweet* was published. The later tweet, however, implicitly mentions that the author

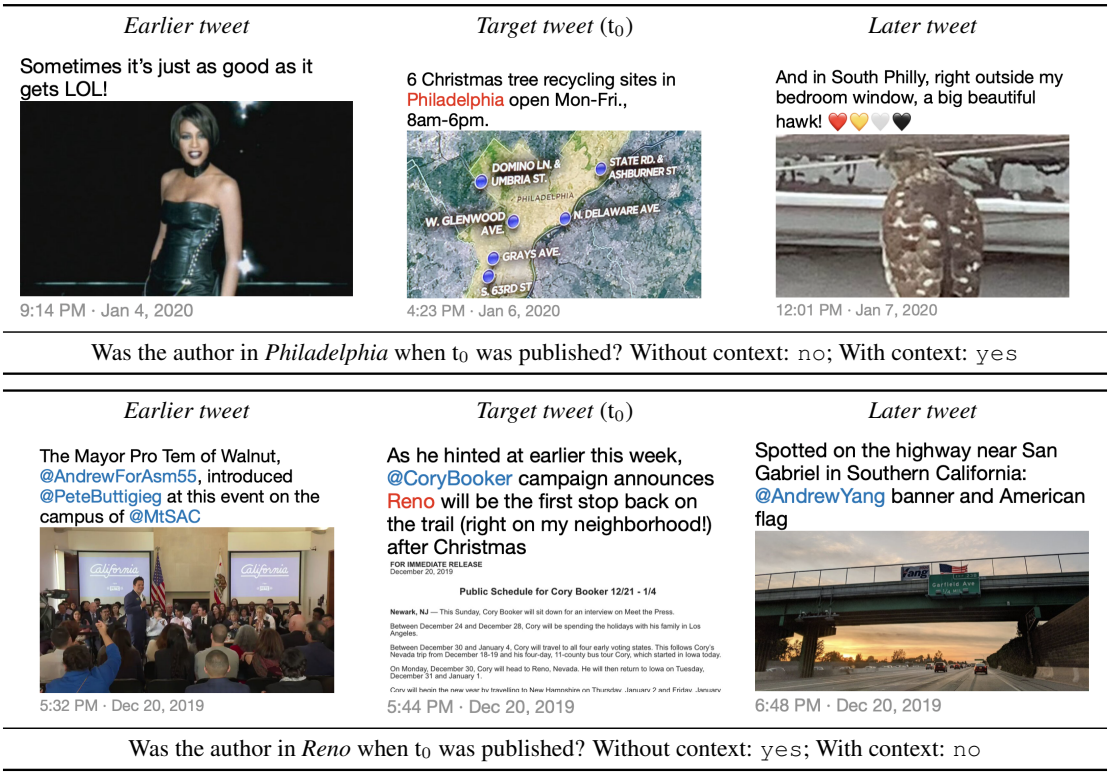

Figure 2: Examples of tweets with different human judgments depending on whether we show context to annotators. Note that the *correct* ground truth is the label obtained when annotators have access to context. We only show selected tweets in the context because of space constraints (recall that we show annotators the three tweets published before and after a target tweet). Annotators understood that the target tweet in the first example was probably a news update and did not conclude that the author was in *Philadelphia* when they did not have access to context. The tweet published after, however, implicitly states that the author lives in that city (his bedroom is in Philly) thus the correct label is `yes`. The target tweet in the second example is about a local election campaign and annotators understood that the author was in *Reno* looking only at the target tweet. Tweets published before and after within an hour, however, mention distant locations and annotators rightfully concluded that the author was not in Reno.

lives in South Philly—his bedroom is there. Since the target and later tweets were published in consecutive days, annotators concluded that the author was in Philadelphia. In this example, context provides evidence that the target tweet is not just a news report about Philadelphia, but a report from someone who lives there.

In the second example, the *target tweet* shares information about a campaign event to take place in Reno, where the author lives (*right on my neighborhood*). Without context, annotators understood that the author was there when he published the target tweet despite the only information available was that he lives there. Looking at earlier and later tweets, however, annotators concluded that he was not in Reno as he tweeted from *San Gabriel* near Los Angeles shortly after.

5 Experiments and Results

We use the 3,494 instances in our corpus to conduct experiments to automatically determine whether authors of tweets are located in the cities they mention in their tweets. Each instance consists of seven tweets published in chronological order. We reduce the problem to a classification task. The inputs to the model are seven tweets. The output is a label indicating whether the author of the tweets was located in the city when the *target tweet* was published (`yes` or `no`). We create stratified training and test splits (80% / 20%) and reserve 20% of the training split for validation. Our models do not take into account user information. They make predictions solely based on the tweets' text. [3] We use Pytorch (Paszke et al., 2019) and the pretrained models (i.e., BERT) released by HuggingFace (Wolf et al.,

[3] Since only part of context tweets contain images, we only take into account tweets' text.

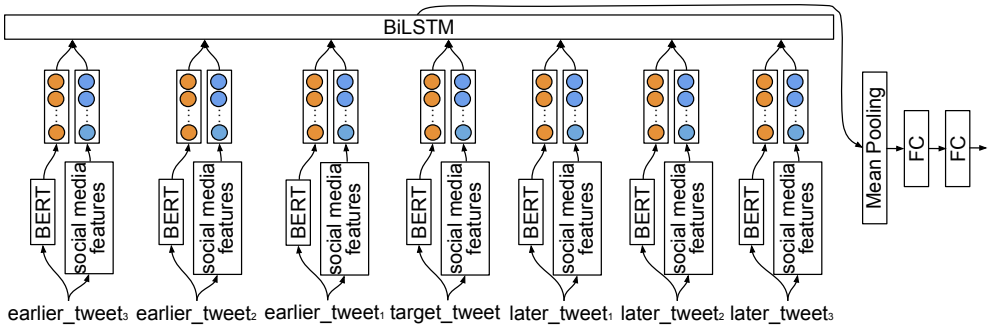

Figure 3: Neural network architecture for determining whether the author of a tweet (target_tweet) is located in the places mentioned in the tweet. We represent the target tweet and context tweets by concatenating their BERT representations and social media features. The BiLSTM processes the sequence of tweets and two fully connected layers after mean pooling make the final prediction.

	no			yes			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1
Majority baseline	0.00	0.00	0.00	0.68	1.00	0.81	0.46	0.68	0.55
Context-Unaware Network (target tweet)	0.00	0.00	0.00	0.68	1.00	0.81	0.46	0.68	0.55
Context-Aware Networks									
earlier + target tweets	0.00	0.00	0.00	0.68	1.00	0.81	0.46	0.68	0.55
target + later tweets	0.00	0.00	0.00	0.68	1.00	0.81	0.46	0.68	0.55
earlier + target + later tweets	0.44	0.28	0.35	0.71	0.83	0.76	0.62	0.65	**0.63**
without social media features	0.39	0.30	0.34	0.70	0.78	0.74	0.60	0.62	0.61

Table 2: Results obtained with the majority baseline, context-unaware network (i.e., only taking into account the *target tweet*), and context-aware networks (three tweets published before and after). The context-unaware network does not outperform the majority baseline, indicating the need of contextual information. The results obtained with context-aware networks demonstrate that it is beneficial to consider (a) both *earlier_tweets* and *later_tweets* (only considering either one does not outperform the baseline), and (b) the social media features.

2020). We train all the neural networks 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.

Context-Unaware Neural Baseline First, we experiment with a neural network that only takes into account the target tweet (t_0) to establish a neural baseline. The inputs of the context-unaware neural network include the tweet text and social media features we extract from the tweet text. The social media features characterize the text in a tweet and include the number of hashtags and features targeting emojis and subjective language. The supplementary materials detail all the features we work with. We use the vector associated with the CLS token returned by BERT [4] (Devlin et al., 2019) as the text representations. We concatenate the text representations and social media features 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).

Context-Aware Neural Network Figure 3 shows the architecture of our context-aware network. The inputs of context-aware networks are the same as the context-unaware network, except that the context-unaware network only uses one tweet (*target tweet*), while the context-aware networks use seven tweets (three *earlier tweets*, one *target tweet*, and three *later tweets*). Since BERT [5] is pretrained using masked language modeling and next sentence prediction—two general tasks not related to extracting spatiotemporal knowledge from social media posts—we extract social media features to enhance the BERT representations. We provide the details

[4] We use bert-base-uncased version of BERT. Note that during the model training process, BERT is held fixed; its parameters are not updated during the training procedure.

[5] Similar to the experiments with the context-unaware neural baseline, BERT is held fixed during the training process.

about social media features in the supplementary materials. The social media features are concatenated with the BERT representations. Note that BERT tokenizer can preprocess the inputs, thus we do not have a data preprocessing step. We use a BiLSTM (size: 1024) (Hochreiter and Schmidhuber, 1997) to encode the sequence of tweets (earlier, target and later tweets). We use a mean pooling layer (Lin et al., 2014) to downsample the feature map and keep the most informative features. We use two fully-connected layers to reduce the dimensionality. The numbers of units in the two fully connected layers are set to 512 and 2, respectively. We use dropout (Srivastava et al., 2014) [6] and ReLU (Nair and Hinton, 2010) in the second-to-last fully connected layer to avoid overfitting and vanishing gradient problems. We experimented with alternative pretrained text encoders and various configurations, including different sizes for the BiLSTM and fully connected layers. Specifically, we have explored BERTweet (Nguyen et al., 2020) and max pooling. However, we do not observe any improvements from these alternative choices.

5.1 Results

Table 2 shows the experimental results with the test split using the majority baseline (always `yes`), the context-unaware neural network, and the context-aware networks. The context-unaware network does not outperform the majority baseline. This indicates that the context tweets are essential for the neural network to extract spatial information.

We present results with the context-aware network using four inputs to represent an instance: (a) *earlier tweets* and *target tweet*, (b) *target tweet* and *later tweets*, (c) *earlier tweets*, *target tweet* and *later tweets* (full input), and (d) same as (c) but excluding the social media features. We note several interesting observations:

- Neither *earlier* nor *later tweets* are beneficial by themselves. Only considering either yields the same results than the majority baseline and the context-unaware network.
- The social media features characterizing the text in the tweet are beneficial (F1: 0.61 vs. 0.63). In other words, these features provide information that is not captured in the distributed representation generated by BERT. Note that social media features are beneficial only when combining *earlier* and *later tweets*. Packing any of the *earlier*, *target*, and *later tweets* with social media features does not yield better results (F1: 0.55 vs. 0.63).

[6] Dropout rate is 0.2.

5.2 Qualitative Analysis

To better understand what kind of mistakes our best-performing model makes, we perform a qualitative analysis. We randomly picked 100 errors made by the best-performing model. In order to analyze the source of the errors, we divide the errors into two parts. The first part (46% of all errors) contains the errors that occur when the *context tweets* mislead the model. We refer to these errors as *context errors*. The second part (54% of all errors) contains the errors that occur when the target tweets mislead the model. We refer to these errors as *target errors*.

When and why did context tweets mislead the model? Figure 4 shows the two most common errors in the *context errors*. We only show one earlier and one later tweet due to space constraints. Note that the percentages are relative to the number of *context errors*. The most common error (63%) occurs when there are multiple references to people and named entities in the context and target tweets. The first example (top) exemplifies this error. The tweets refer to, among others, *Tom Hanks*, *his family*, *Black college student*, and *Peter Liang*. We hypothesize that the model struggles to identify which reference to people, if any, is semantically close to the author. The second most common error (21%) occurs when the publication timestamps is key to make the right prediction. The tweets in the second example (bottom) were published within 5 hours and mention *Denver*, *San Diego*, and *Bahamas*. Taking as a whole, it is not possible to tell whether the author was in any of those locations. The model, however, struggles to figure this out as it does not have access to publication timestamps.

When and why did target tweets mislead the model? Figure 5 shows the most common errors in the *target errors*. Note that the percentages are relative to the number of *target errors*. We only show the most common errors and the *target tweets* due to space constraints. The most frequent error (48%) occurs when the target tweet is a news report. In the example (left), the author is reporting on a news event taking place in *Arlington* and the model wrongly predicts `yes` despite there is no evidence that the author was there. The second most common errors (42%) occurs when the target tweet is

Figure 4: Examples of the most common context-related errors made by the best-performing model (bottom row in Table 2). The most common error (63%) occurs when there are many named entities in the tweets, especially *persons*. The first example (top) illustrates this error type. The many mentions of people (Tom Hanks, his family, Black college students, and Peter Liang) mislead the model. The second most common error (21%) occurs because the model does not have access to tweet timestamps. The second example (bottom) exemplifies this scenario. It is unlikely that somebody is in *Denver*, *San Diego* and *Bahamas* within 5 hours, and the model wrongly predicted that the author was in *San Diego* when he published the target tweet despite earlier and later tweets indicate otherwise.

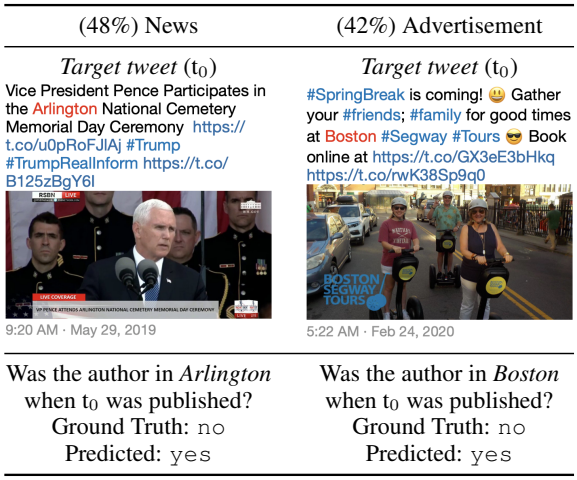

Figure 5: Most common errors made by the best-performing model when it is mislead by the *target tweet*.

an advertisement and the model predicts `yes` despite there is no evidence that the author was in the city at hand, as the example on the right shows. A less common source of errors (10%) occurs when the city mentioned in the target tweet is not a location. For example, *Houston* is sometimes used to refer to the singer rather than the city.

6 Conclusions and Future Work

Context does matter in determining the relationship between the authors of tweets and the places they mention in their tweets. We have demonstrated so by (a) analyzing whether human judgments change depending on whether we show them *context tweets* and (b) investigating whether neural networks benefit from incorporating *context tweets*. We found that 36.9% of human judgments change depending on whether we show *context tweets*. Experimental results demonstrate that networks that incorporate *context tweets* yield better results.

We have created a corpus of 3,494 instances

which consists of seven tweets (three *eariler tweets*, one *target tweet*, and three *later tweets*) published in chronological order. The annotations include whether the authors of tweets are in the locations they mention in a *target tweet*. As part of our future work, we plan to experiment with longer contexts and include tweet timestamps. Also, we plan to address the most common cause of errors: multiple mentions to people and named entities. Our research agenda includes investigating how to differentiate between entities mentioned in a tweet and those who participate in the events described in a tweet (including the author), or more generally, described in social media data.

7 Limitations

The work presented here has several limitations:

Corpus and annotations. First, our corpus only contains tweets written in English and containing a city in the US. While accounting for any location worldwide is outside the scope of the paper, it is possible that the corpus does not generalize to other locales. Second, in order to ensure high agreements, we disregarded annotations of the worse-performing crowdworkers (final corpus size: 3,494 instances). This strategy has the potential to also disregard the most challenging instances (i.e., most ambiguous in the eyes of the annotators) and may simplify the problem of extracting saptial knowledge in general.

Experiments and results. Our models rely on large pre-trained transformer that are not available in all languages and require substantial computation. We note, however, that the social media features we define manually do not have this issue and bring substantial improvements. Additional limitations of the current models are described in the Qualitative Analysis (Section 5.2).

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Claire Beveridge. 2021. 33 Twitter Stats All Marketers Need to Know in 2022. https://blog.hootsuite.com/twitter-statistics/. Accessed: February, 2021.

Claire Cardie. 1997. Empirical methods in information extraction. *AI Magazine*, 18(4):65.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. 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, page 759–768, New York, NY, USA. Association for Computing Machinery.

Wen-Haw Chong and Ee-Peng Lim. 2017. Tweet geolocation: Leveraging location, user and peer signals. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1279–1288, New York, NY, USA. Association for Computing Machinery.

Brian Dean. 2021. How Many People Use Twitter in 2021? [New Twitter Stats]. https://backlinko.com/twitter-users. Accessed: October, 2021.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. 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.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hicham G. Elmongui, H. Morsy, and Riham Mansour. 2015. Inference models for twitter user's home location prediction. *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media + Society*, 4(1):2056305118763366.

Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2021. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757.

Georg Fuchs, Gennady L. Andrienko, Natalia V. Andrienko, and Piotr L. Jankowski. 2013. Extracting personal behavioral patterns from geo-referenced tweets.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

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.

David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. 2021. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):188–197.

Qiuhong Ke, Mohammed Bennamoun, Senjian An, Farid Boussaïd, and Ferdous Ahmed Sohel. 2016. Spatial, structural and temporal feature learning for human interaction prediction. *CoRR*, abs/1608.05267.

Simon Kemp. 2022. Twitter Statistics And Trends. https://datareportal.com/essential-twitter-stats.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Retrieved from https://repository.upenn.edu/asc_papers/43*.

Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network in network. *CoRR*, abs/1312.4400.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. EventPlus: A temporal event understanding pipeline. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 56–65, Online. Association for Computational Linguistics.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2021. Where is this tweet from? inferring home locations of twitter users. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):511–514.

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.

Paul Mousset, Yoann Pitarch, and Lynda Tamine. 2020. End-to-end neural matching for semantic location prediction of tweets. *ACM Trans. Inf. Syst.*, 39(1).

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. 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.

James Pustejovsky. 2013. Where things happen: On the semantics of event localization. In *Proceedings of the IWCS 2013 Workshop on Computational Models of Spatial Language Interpretation and Generation (CoSLI-3)*, pages 29–39, Potsdam, Germany. Association for Computational Linguistics.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 215–221. AAAI Press.

Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. 2013. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological Research Online*, 18(3):74–84.

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.

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for sentiment analysis in Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2345–2354, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Zelin Wang, Zhijian Wu, Ruimin Wang, and Yafeng Ren. 2015. Twitter sarcasm detection exploiting a context-based model. In *Web Information Systems Engineering – WISE 2015*, pages 77–91, Cham. Springer International Publishing.

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.

Zhaomin Xiao and Eduardo Blanco. 2022. Are people located in the places they mention in their tweets? a multimodal approach. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2561–2571, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203.

A Supplementary Materials

Table 3 details the social media features used in our experiments. Experimental results (Table 2) show that these features complement the distributed representation obtained with BERT. In other words, combining the representation from BERT and these social media features is beneficial to extract spatiotemporal knowledge.

	Name	Description
Hashtags	num hashtag	The number of hashtags
	within hashtag	Whether the location is within a hashtag, e.g., *#houston*
Emojis	num emoji	The number of emojis
	most common emoji	The three most common emojis
Subjectivity	num strong_subj	The number of strongly subjective words (e.g., *excoriate*)
	num weak_subj	The number of weakly subjective words (e.g., *say*)
Other	within mention	Whether the location is within a mention, e.g., *@dallasnews*
	num URL	The number of URLs
	num token	The number of tokens
	num elongated	The number of elongated words (e.g., *looooove*)
	num exclamation	The number of exclamation marks

Table 3: Social media features used to complement the BERT representation of text in a tweet (Section 5, Figure 3). Note that the strongly subjective words and weakly subjective words are from MPQA (Deng and Wiebe, 2015). Experimental results show that taking these social media features into account is beneficial. In other words, these features complement the BERT representation when context (earlier and later tweets) are fed to the network.