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

Sunshine with a Chance of Smiles: How Does Weather Impact Sentiment on Social Media?

Julie Jiang,[1,2]* Nils Murrugara-Llerena,[2] Maarten W. Bos,[2] Yozen Liu,[2] Neil Shah,[2] Leonardo Neves,[2] Francesco Barbieri[2]

[1] Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA
[2] Snap Inc., Santa Monica, CA 90405, USA
juliej@isi.edu, {nmurrugarraller, maarten, yliu2, nshah, lneves, fbarbieri}@snap.com

Abstract

The environment we are in can affect our mood and behavior. One environmental factor is weather, which is linked to sentiment as expressed on social media. However, less is known about how integrating changes in weather, along with time and location contextual cues, can improve sentiment detection and understanding. In this paper, we explore the effects of three contextual features–weather, location, and time–on expressed sentiment in social media. Leveraging a large Snapchat dataset, we provide extensive experimental evidence that including contextual features in addition to textual features significantly improves textual sentiment detection performance by 3% over transformer-based language models. Our results also generalize cross-domain to Twitter. Ablation studies indicate the relative importance of weather compared to location and time. We also conduct correlation analyses on 8 million Snapchat posts to highlight the link between past weather and current sentiment, showing that weather has a lasting impact on mood. Users generally exhibit more positive sentiment in better weather conditions as well as in improved weather conditions. Additionally, we show that temperature's link with mood holds after controlling for time or population density, but there exist geographical differences in how temperature affects mood. Our work demonstrates the effectiveness of including external contexts in linguistic tasks and carries design implications for researchers and designers of social media.

Introduction

How's the weather where you are now? We as authors hope it is warm and sunny because weather can have a significant impact on your mood. Sentiment analysis performance has tremendously advanced following developments in transformer-based language models (Devlin et al. 2019; Barbieri et al. 2020). However, beyond textual cues, there are also easily accessible yet underutilized environmental cues that could help improve content understanding. One important factor in our environment is the weather, which affects our mood (Howarth and Hoffman 1984; Keller et al. 2005; Denissen et al. 2008) and behavior (Chen, Cho, and Jang 2015; Guéguen 2013; Yang, Jhang, and Chang 2016). Several works also suggest a link between online expressed sentiment and weather on social media (Coviello et al. 2014; Li, Wang, and Hovy 2014; Baylis et al. 2018; Wang, Obradovich, and Zheng 2020).

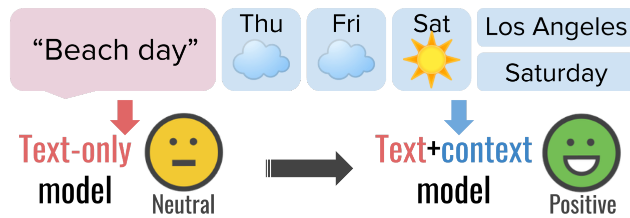

Figure 1: A text-only model misclassifies the sentiment as neutral. The contextual cues of location (Los Angeles), time (a weekend), and weather (the sun coming out after two cloudy days) can help inform that the sentiment is positive.

However, less is known about how integrating *changes* in weather, along with time and location contextual cues, can improve sentiment detection and understanding. Consider the motivating example depicted in Figure 1. Even a state-of-the-art text-based model misclassifies the sentiment as neutral.[1] But a model that is aware of the contextual cues–the example took place in LA (*location*) on a weekend (*time*) and it has been cloudy for the past two days (*historical weather*) before the sun came out (*current weather*)–may correctly classify the sentiment as positive.

The contribution of this work is a comprehensive study of how external contextual factors–weather, location, and time–impact expressed sentiment on social media. We collect a large-scale dataset of 8 million public Snapchat posts in the US between March and September 2020. Snapchat is a popular social media platform known for its ephemeral multi-media messages. We further collect an extensive weather dataset including both real-time weather data and up to eight weeks of historical weather for each post. Additionally, we replicate our main findings cross-domain on a Twitter dataset. We aim to answer the following two research questions:

- **RQ1 Modeling**: Can including external contextual fea-

*Work done when author was an intern at Snap Inc.

[1] e.g., https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

tures (weather, location, and time) in language models improve sentiment detection?

- **RQ2 Analysis**: How do contextual factors, and particularly historical and current weather, impact expressed sentiment?

Though past work showed that weather can predict expressed sentiment, our work addresses some of their limitations and offers additional insights. First, large-scale online sentiment has not been studied with the latest language models (Devlin et al. 2019). Second, to the best of our knowledge, ours is the first work that models historical weather in addition to current weather and analyzes how past weather and changes in weather affect sentiment. Third, we include location and time in our analysis to isolate geographical and temporal confounding factors in studying weather's impact on sentiment. Finally, while prior work on social media only considered urban areas, our data is geographically more diverse, encompassing substantial data from rural areas ((35%), which is a factor that may impact community sentiment (Christenson 1979). Our contributions in this paper are summarized as follows:

- Contextually-aware models *significantly* outperform text-only models by *3%* on F1 and Pearson metrics.
- Ablation studies show that short-term historical weather ($\approx$ three days) is the most important weather factor in predicting sentiment.
- We demonstrate that sentiment is sensitive to current weather, past weather, and changes in weather.
- The effect of weather remains after controlling for time (hourly and weekly) or population density, though it does exhibit geographical differences.

Our work demonstrates the promising use of external contextual features of time, location, and in particular weather, to improve social media language modeling frameworks. Since such contextual cues are easy to obtain and integrate into existing frameworks, without imposing on user privacy, our work carries many design implications for social media platforms and content understanding frameworks.[2]

Related Work

Dey and Abowd (1999) conceptualized four primary types of contexts in context-aware computing: location, identity, time, and activity. In this work, user identity is anonymized and the activity is public content-sharing. We focus on location and time, and the secondary context they induce: weather.

Weather Affects Mood and Behavior

Extensive literature suggests a link between weather–sunlight, temperature, humidity, wind, etc.–and mood (Howarth and Hoffman 1984; Denissen et al. 2008; Kööts, Realo, and Allik 2011). Time spent outdoors in nice weather is also associated with a better mood (Keller et al. 2005). Though the effect sizes of weather's impacts on mood are small (Keller et al. 2005; Denissen et al. 2008), they are significant and are consistent with the literature on the seasonal effects of weather on mood (Rosenthal 1984; Harmatz et al. 2000).

Weather also affects behavior. It can predict the stock market (Chang et al. 2008; Yang, Jhang, and Chang 2016), the housing market (Hu and Lee 2020), university admissions (Simonsohn 2007), daily activity decisions (Doksæter Sivle and Kolstø 2016), and shopping decisions (Parsons 2001; Busse et al. 2015). Weather influences social interactions as well. High temperatures are linked to aggression (Anderson et al. 2000; Zhong and Zhou 2012). When the weather is nice, people are more generous (Cunningham 1979) and courtship success is higher (Guéguen 2013). While people generally prefer summers to winters (Rosenthal 1984; Harmatz et al. 2000), temperature is not linearly associated with good behavior. Cunningham (1979) found that people's helping rates dropped when temperature deviates from the optimal temperature of 19°C (65°F).

Observable Effects of Weather on Social Media

Weather effects on sentiment have also been observed on social media (Hannak et al. 2012; Li, Wang, and Hovy 2014). In a Twitter study, Li, Wang, and Hovy (2014) showed that the average temperature difference between two consecutive days contributed to mood, with people generally appreciating cooler weather up to a certain point. They also showed that rain, snow, and hail induced negative sentiments (Li, Wang, and Hovy 2014). Other studies found that rainfall, extreme temperatures, precipitation, humidity, and cloudiness were associated with negative sentiment on Facebook and Twitter (Coviello et al. 2014; Baylis et al. 2018). Wang, Obradovich, and Zheng (2020) found that extreme weather worsened expressed sentiment on Weibo posts in China. However, all of these works examined only urban areas and did not factor in historical weather.

Location and Time

The *where* and the *when* are two of the most fundamental contextual cues (Dey and Abowd 1999). The impact of location on mood and behavior can be characterized by aspects of geography, ecology, socioeconomy, and cultural divides (Mitchell et al. 2013; Van de Vliert and Van Lange 2019). Previous research on Snapchat Stories found the decision to share content publicly (via *Our Story*) or privately to friends (via *My Story*) may depend on the location and the time (e.g., public sharing is less likely when users are at home) (Habib, Shah, and Vaish 2019). Many works also recognized the impact of time on mood: from analyzing diurnal activity patterns, researchers found that people are generally in a better mood on weekends (Dodds et al. 2011; Golder and Macy 2011; Li, Wang, and Hovy 2014). Emotions also vary throughout the day. People's moods are usually better in the morning but often deteriorate as the day progresses (Dodds et al. 2011; Golder and Macy 2011). Hannak et al. (2012) demonstrated that combining location and time in models can contextualize weather cues. Following their suggestion, we incorporate location and time to further understand the effects of current and historical weather on sentiment.

[2]The code and the Twitter dataset are available at https://github.com/snap-research/sentiment-weather-impact.

Language Models

Pre-trained transformer-based language models (LMs) such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) have altered the state of natural language applications including text classification and natural language inference (Wang et al. 2018). In the case of standalone and short social media texts such as ours, RoBERTa is preferred since it does not use the next sentence prediction objective (Liu et al. 2019; Barbieri et al. 2020). In particular, RoBERTa with additional in-domain retraining has brought considerable improvement over prior methods in Twitter sentiment classification, among other social media NLP initiatives (Barbieri et al. 2020).

Data

Snapchat

Well-known for its ephemeral chat messages, Snapchat launched the *Our Story* feature in 2014. *Our Story* is a collection of user-submitted public Snaps about specific events, places, or topics (Anderson 2015; Snapchat 2018). Using an internal API, we sampled 8M public *Our Story* Snaps with textual captions geolocated in the US between March and September 2020. We utilize the user-generated captions, timestamps, and locations of the Snap. No user personal information is used. We also do not include audiovisual signals, since previous research on Snapchat suggests that incorporating visual cues in addition to textual cues did not significantly improve textual sentiment detection (Chaudhuri 2019). By linking the city and states to the list of urban areas from the 2010 census (US Census Bureau 2021), we found that 35% of our dataset comes from rural areas.

We use two types of features: textual and contextual. Textual features derive from the user-generated Snap captions. The contextual features are *location*, *time*, and *weather*. The location features include the city, state, and altitude (available for 95% of the data) of the Snap. We also match the location to the 2010 census database to obtain the population density.[3] We one-hot encode the US state and the city and impute missing altitudes with the mean from each state. For each Snap, we compile a set of localized time features including the day of the year, weekday, and hour of the day. We discuss the collection and processing of weather features below.

Weather features. We use OpenWeatherMap[4] to collect weather data. Weather is queried using an hourly timestamp and location (city and state). For each query, OpenWeatherMap returns the exact hourly *temperature*, *"feels like"*, *atmospheric pressure*, *humidity*, *windspeed*, *cloudiness* (%), *rain* volume, and *snow* volume. It also returns a single string description of the weather type, such as *clear* or *cloudy*. We one-hot encode the four major weather types: *clear* (32%), *cloudy* (53%), *rainy* (13%), and *foggy* (2%).

We collect the exact hourly weather at the time of the Snap as well as the aggregated daily (24-hour period of the day) and historically averaged weather during the three days, one week, two weeks, four weeks, and eight weeks periods prior to the day of the Snap. To aggregate, we compute the mean, maximum, and minimum of the hourly *temperature*, *"feels like"*, *pressure*, *humidity*, *windspeed*, and *cloudiness*. The *rain* and *snow* volumes are summed up. We also calculate the proportions of the four major weather types over the aggregation period.

[3]https://simplemaps.com/data/us-cities

[4]https://openweathermap.org/

Dataset	IRA	Fleiss' κ	G&K's γ
Snapchat	0.69	0.49	0.88
Twitter	0.47	0.25	0.75

Table 1: Both the Snapchat and Twitter datasets have average inter-rater Gammas over 0.70, indicating good reliability of the annotations.

Twitter (and Instagram)

As an additional dataset for experimental validation, we collect 2 million geo-tagged original Tweets from October 2015 to February 2017 using the Twitter API. Upon examining the source of the Tweets, we find that more than 90% of our data were also concurrently posted to Instagram. We matched the data collection with an hourly weather dataset of 30 major US and Canadian cities (Beniaguaev 2017). Due to the sparsity of the collected weather data, there is insufficient historical weather data, so we use only hourly weather features.

Sentiment Annotations

A random subset of 4,000 Snap captions and 5,000 Tweets were annotated for sentiment. The Snap captions were annotated by four in-house expert annotators on a three-point scale of negative, neutral, and positive sentiment. The Tweets were annotated through Amazon Mechanical Turk[5] on a five-point scale of very negative, somewhat negative, neutral, very positive, and somewhat positive. The extension to a five-point scale is to improve the reliability of non-expert ratings (Preston and Colman 2000). Following Snow et al. (2008), we employ five (non-expert) workers per Tweet to approximate the performance of experts. After removing unqualified instances (e.g., spam, non-English) and instances with insufficient annotations, we obtain a dataset of 2,885 Snaps and a dataset of 4,469 Tweets. Details of the annotation process can be found in the Appendix.

Agreement and reliability. We evaluated the reliability of annotations using the inter-rater agreement (IRA) score, Fleiss' Kappa (Fleiss 1971), and Goodman and Kruskal's Gamma (Goodman and Kruskal 1979) in Table 1. IRA is the average agreement (i.e., equivalent annotations) between every pair of annotators, and Fleiss' Kappa (Fleiss 1971) is traditionally used for multi-rater, ordinal categories of annotation tasks. We primarily use the Goodman and Kruskal's Gamma (Goodman and Kruskal 1979) correlation coefficient to assess annotation reliability following recommendations by Amidei, Piwek, and Willis (2019), who suggested that correlation in agreement is preferred over an exact

[5]https://www.mturk.com/

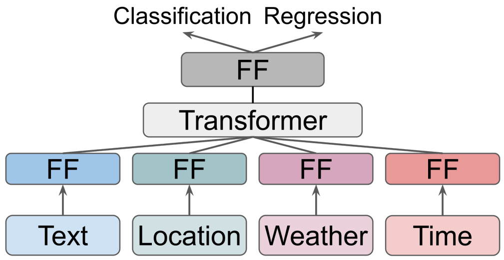

Figure 2: Illustration of our proposed model architecture. FF: blocks of feed-forward layers.

agreement in the strictest sense as human variations in language interpretations make it difficult to obtain exact agreement. Both datasets have an average inter-rater Gamma over 0.70, indicating good reliability of the annotations (Rosenthal 1996; Amidei, Piwek, and Willis 2019).

Annotation aggregation. To capture nuances and ambiguities that naturally arise when subjectively annotating sentiment, we do not dismiss any instances with low agreement (Poesio and Artstein 2005; Pavlick and Kwiatkowski 2019; Fornaciari et al. 2021). We instead take a two-pronged approach to aggregate annotations by using both hard and soft labels. Our approach is inspired by Fornaciari et al. (2021), which proposed predicting soft labels (a probability distribution over the hard labels) in addition to the hard labels for various NLP tasks.

The hard labels are the annotators' majority votes. Since we have four annotators and three possible classes for the Snapchat dataset, at least one class will receive two votes (the pigeonhole principle). If the votes are split half-half between two classes, we assign the neutral label. Instances in which two voted for positive and two voted for negative are extremely rare ($< 1\%$). For the Twitter dataset, we combine very/somewhat positive as positive and very/somewhat negative as negative and select the majority vote out of the three remaining categories. Instances for which there is no majority vote are already discarded (see Appendix).

The soft labels are aggregated by averaging all of the annotators' votes. The labels are converted to numeric values with 0.00 = (very) negative, 0.25 = somewhat negative, 0.50 = neutral, 0.75 = (somewhat) positive, and 1.00 = (very) positive.

Method

We adopt a modular deep learning framework, as shown in Figure 2. There are four input modules—text, location, weather, and time. Each module individually undergoes several feed-forward (FF) layers before being combined and undergoing several more FF layers. The textual feature input is the output of an LM (e.g., *BERT* (Devlin et al. 2019)). In essence, we freeze the LM and train additional layers on top. We do not finetune the LMs because freezing and finetuning lead to comparable performance (Peters, Ruder, and Smith 2019; Pfeiffer et al. 2020), yet freezing is substantially cheaper and computationally more stable (Devlin et al. 2019). Importantly, since we aim to find if adding contextual features improve prediction performance, choosing the absolute best LM base model is peripheral to our goal.

We make two important architectural choices through experimentation. First, we experiment with two ways to incorporate weather features. One is simply to include weather features from all timeframes as one long, flattened array, denoted *W(one)*. The other is to add each timeframe of weather feature as a separate module, denoted *W(sep)*. Second, we explore several concatenation strategies, including linear concatenation, many-to-one attention (Luong, Pham, and Manning 2015), and a transformer block (Vaswani et al. 2017).

Since we aggregated two types of annotation labels—hard and soft—we employ a multi-task learning strategy, optimizing both for classification and regression using categorical cross-entropy and mean squared error, respectively. Both losses contribute equally to the final loss function.

Results

Experimental Results (RQ1)

To answer RQ1, we conduct extensive experimental studies on both the annotated Snapchat and Twitter datasets. We fix the train/val/test sets to be randomly stratified splits of 80%/10%/10%. The classification (hard) labels are evaluated with F1 score.[6] Following Mohammad et al. (2018), the regression (soft) labels are evaluated with the Pearson correlation coefficient. The final score for evaluation is $\frac{1}{2}$(F1 + Pearson). We select the best hyperparameters using the validation set (see the Appendix for hyperparameter search space). We test each model configuration ten times with ten different random seeds, which we use to determine the statistical significance between any two models with a two-sample t-test.

Including contextual cues. We compare text-only models with text and context models on the Snapchat dataset. The contextual features used are weather, location, and time. Weather features include all current (hourly and daily) and historical (three days to eight weeks) timeframes. We use the following LMs as text-only baselines:

- *RoBERTa-base* and *RoBERTa-large* (Liu et al. 2019) from HuggingFace (Wolf et al. 2020).
- *Snap-RoBERta* (base), a *RoBERTa* model fine-tuned on 40M Snap captions.

Table 2 presents our main results on the Snapchat dataset. For text-only models, *Snap-RoBERTa* achieves the best performance compared to off-the-shelf LMs, possibly due to in-domain finetuning. Using *Snap-RoBERTa*, we compare the effects of adding contextual features with various concatenation and weather input strategies. All of our text and context models show *significant* improvement over the text-only model. Furthermore, we find that using transformer blocks outperforms attention layers, which in turn outperforms simple concatenation. Between using weather timeframe as separate modules *W(sep)* and as one module *W(one)*, *W(one)*

[6] Macro- and micro-F1 scores are equivalent on stratified datasets.

Model		%↑	F1 & Pearson	Pearson	F1	Neg. F1	Pos. F1
Text-only	RoBERTa-large	-	69.88	72.46	67.30	63.26	78.81
	ROBERTa	-	65.07	65.43	64.72	55.32	76.93
	Snap-RoBERTa	-	**74.23**	79.18	69.29	66.81	80.21
Text + context	Snap-RoBERTa+W(one)+L+T: Concat	2.0%*	75.74	81.29	70.19	67.25	82.42
	Snap-RoBERTa+W(one)+L+T: Attention	2.5%*	76.08	79.17	72.99	70.76	82.44
	Snap-RoBERTa+W(sep)+L+T: Transformer	2.7%*	76.22	81.75	70.69	69.14	81.06
	Snap-RoBERTa+W(one)+L+T: Transformer	3.2%*	**<u>76.64</u>**	81.36	71.93	70.66	82.39

W(one): weather as one module; W(sep): weather as separate modules by timeframes; L: location; T: time

Table 2: Experimental results comparing text-only models with text + context models show that adding contextual features bolsters performance for the Snapchat dataset. The best F1 & Pearson score is bolded for each category and the best score overall is underlined. We compute the % improvement over the *Snap-RoBERTa*'s score. * indicates statistically significant ($p < 0.05$) improvements.

Model	F1 & Pearson	F1	Pearson
Random	15.62	32.57	-1.33
Majority	16.29	32.57	0.00
W(one)+L+T	**19.73**	31.16	8.29

Table 3: The context-only model outperforms a naive random or majority model. Pearson Correlation coefficient is set to 0 when one variable is constant.

Context	Only	Without
Weather	**76.46 (3.0%)**	**75.90 (2.2%)**
Location	75.85 (2.2%)	76.82 (3.5%)
Time	75.77 (2.1%)	76.60 (3.2%)

Table 4: Ablation studies show that weather is the most important contextual feature. Using only weather features achieves the highest score (Only) and not using weather achieves the lowest score (Without). Scores are F1 & Pearson (% improvement over *Snap-RoBERTa*).

performs better. This added improvement could be attributed to the model associating current with historical weather and responding to long-term trends in weather. The best text and context model is markedly better (*3.2%*) than the text-only model, and the improvement is statistically significant.

Using only contextual cues. We also compare a model using only context features with a pseudo-random (based on the distribution of the labels) model and a model that only predicts the majority label in Table 3. The context-only model achieves a score of 19.73 and is over *20%* (also significantly) better than a pseudo-random or majority model. This striking improvement shows that even when the model is not given text data, it can still extract clues from the contextual features to predict the textual sentiment.

Context ablation. Table 4 reports the ablation study results of the relative importance of each type of context. We use the Transformer concatenation strategy and the *Snap-RoBERTa* model for text. We experiment with including only one type of context (*'Only'*), in which a higher score indicates higher relative importance for that context, as well as with removing one context (*'Without'*), in which a lower score indicates higher importance. In both experiments, we observe that weather features stand out as the most important feature. Only adding weather leads to the highest performance gain and removing it leads to the lowest performance gain.

Weather ablation. We additionally conduct an in-depth investigation of weather features. The questions we want to answer are: (i) how much (historical) weather information do we need to achieve good performance? (ii) which weather timeframe is the most predictive for sentiment analysis? To this end, we experiment with incrementally adding more historical weather data and only using one timeframe of weather data in Figure 3. The left subfigure indicates that adding historical weather ranging from three days to two weeks is beneficial. Adding too *much* historical weather (e.g., eight weeks) can actually hurt performance, possibly because historical weather that goes too far back is not as informative. The right subfigure shows that adding three days of historical weather in isolation leads to the best performance. Both experiments agree that using three days of historical weather is the most beneficial to improving model performance.

However, we do note that no timeframe of historical weather data is substantially better than the rest. In fact, using historical weather that goes back as far as eight weeks is just as useful in enhancing the model prediction as current weather, underscoring the importance of historical weather. From the ablation study of incrementally adding historical weather, we observe that using (short-term) historical weather in addition to current weather leads to the largest performance boost.

Generalizability of findings. To demonstrate the generality of our findings across domains, we replicate our experiments on the Twitter dataset. In Table 5, we present results for text-only and text and context models using the following LMs:

- *Twitter-RoBERTa-base-sentiment* (Barbieri et al. 2020), pretrained on 58M Tweets and fine-tuned for sentiment analysis

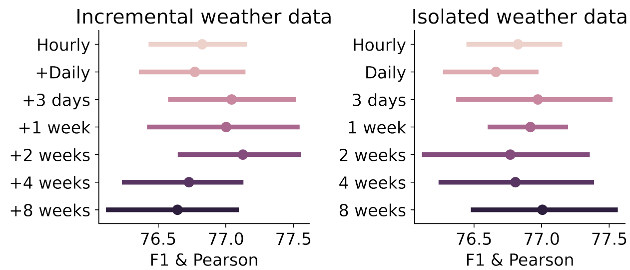

Figure 3: Weather data ablation studies suggest that short-term (≈ three days) historical weather is the most important. Left: incrementally adding more (historical) weather data. Right: only using one timeframe of weather data.

Model	Context	↑%	Score
Twitter-RoBERTa-base-sentiment			75.85
	+WLT	1.34%*	**<u>77.19</u>**
Snap-RoBERTa			63.87
	+WLT	8.82%*	**72.69**
BERTweet-base			60.35
	+WLT	7.74%*	**68.09**

WLT: weather, location, and time

Table 5: On Twitter, models consistently perform better with contextual features. Models are scored with F1 & Pearson. *indicates significant improvement ($p < 0.05$).

- *BERTweet-base* (Nguyen, Vu, and Nguyen 2020), pre-trained on 850M Tweets
- *Snap-RoBERTa*, described above

There are consistent, significant improvements when adding context features, regardless of the LM used, indicating that our findings hold across domains.

Summary: Empirical experiments on both Snapchat and Twitter demonstrate that contextual features, particularly weather, improve textual sentiment detection. Recent historical weather (≈ three days) is the most important historical weather feature.

Impact of Contextual Factors on Sentiment (RQ2)

Next, we analyze the impact of contextual factors on sentiment. We apply the best text-only model, *Snap-RoBERTa* trained on our annotated subset, to the full 8 million dataset to obtain predicted sentiment labels. We choose the text-only model to avoid circular dependencies in our analysis. For ease of interpretability, we use the predicted hard classification labels, with 0.0 being negative, 0.5 being neutral, and 1.0 being positive.

Manual validation. To ensure the predicted sentiment labels are of good quality, an expert annotator (see Appendix for qualifications) annotated a held-out subset of random instances. We held out 200 instances to allow for a larger sample size. Our predictions achieve 73.0 point macro-F1, comparable to the test scores in Table 2.

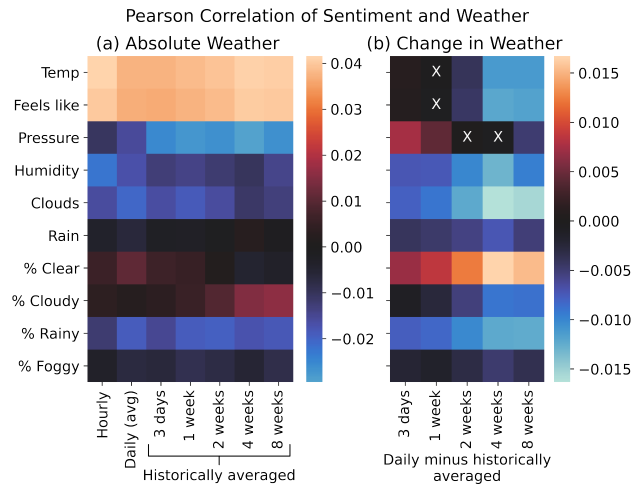

Figure 4: Weather and changes in weather are Pearson correlated with sentiment (a) The correlation coefficient of absolute weather values and sentiment (b) The correlation coefficient of changes in weather values (daily minus historical) and sentiment. All coefficients are significant ($p < 0.05$) except for those marked "x".

Current and historical weather correlate with sentiment. In Figure 4(a), we visualize the Pearson correlation between the recorded weather features (*temperature*, *rain* volume, etc.) and the predicted sentiment across different timeframes (hourly, daily averaged, and historically averaged). Consistent with previous literature, the correlation magnitude is small (Keller et al. 2005; Denissen et al. 2008) but *statistical significant*, indicating the presence of weak sentiment signals from weather. Of all of the weather features, *temperature*, *"feels like"*, and proportion of *clear* weather are positively correlated with sentiment, whereas *Pressure*, *humidity*, *cloudiness*, and proportion of *rain* weather are negatively correlated with sentiment.[7]

Our findings additionally suggest that past weather correlates with sentiment just as much with current weather. Comparing correlation values for each row (i.e., each weather variable), we find that the colors of the squares do not change significantly. Each row is either largely red (positively correlated) or blue (negatively correlated) and this correlation changes minimally across timeframes. In other words, current and historical weather data correlate with current sentiment almost equally. Some weather variables have varying degrees of correlations depending on the timeframe. For example, historical pressure values are more negatively correlated with current sentiment than current pressure values. Such a result would imply that *changes* in weather could have an effect on sentiment, which motivates the following analysis.

[7]The apparent contradiction between the percentage of *clouds* and proportion of *cloudy weather* is likely because the weather is only labeled *cloudy* if the majority of the sky is covered in clouds (Donegan 2016). Therefore, using *cloudiness* more accurately depicts the relationship between cloudy weather and sentiment.

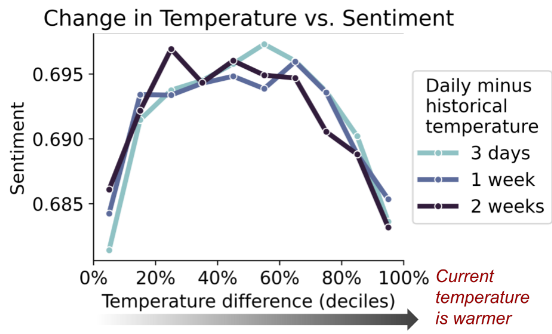

Figure 5: Short-term differences in temperature exhibit concave relationships with sentiment: sentiment declines when the temperature becomes too high or too low. Each family of weather differences is binned into deciles.

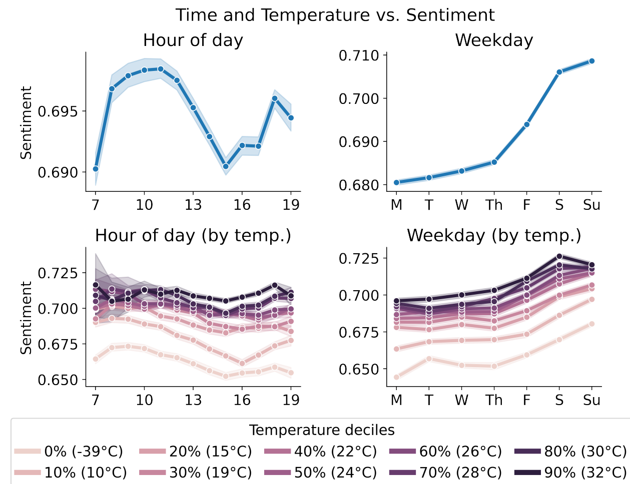

Figure 6: Weather effects on mood exist after controlling for time. Top: sentiment as a function of various time granularities. Bottom: same plots, but disaggregated by the distribution of temperature in deciles (lower bound temperature). Darker lines mean warmer temperatures.

Changes in weather correlate with sentiment. In Figure 4(b), we calculate the difference between the weather on the day of the Snap and historical weather and we show how they correlate with sentiment. We find that sentiment generally changes correspondingly to improvements or declines in the weather. For instance, people experience more positive sentiment when the weather is *clearer*, but more negative sentiment in *humid*, *cloudy*, or *rainy* weather.

For most of these variables, the magnitude of the correlation with weather declines with a longer duration of historical time frames (row-wise in Figure 4(b)), suggesting that recent weather has a larger impact on sentiment than earlier weather. A notable distinction is the top two rows, which show that recent changes in *temperature*/*"feels like"* are not linearly correlated with the weather.

Because our data spans the summer season, a significantly hotter temperature compared to long-term historical temperature (four to eight weeks prior) is negatively correlated with sentiment possibly because of the presence of heat waves, which could decrease sentiment. As such, we next consider short-term temperature changes to monitor mood changes due to weather changes.

Sentiment declines with extreme temperature changes. To explore the dynamics between short-term temperature changes and sentiment, we plot the relationship between the change in temperature and the average sentiment in Figure 5. A large temperature difference means that the present temperature is warmer than before. For each set of computed differences, we bin the values into 10 equally-sized deciles. We see that short-term changes in the temperature exhibit concave relationships with the sentiment: sentiment drops both when the weather becomes too hot or too cold, suggesting that users' expressed sentiment may be sensitive to extreme weather changes, similar to findings in prior works (Baylis et al. 2018; Wang, Obradovich, and Zheng 2020).

Sentiment varies with time. Time is an important factor of consideration when modeling sentiment. The top row of Figure 6 displays the average sentiment for each hour of the day and day of the week. We exclude hours after 7 p.m. and before 7 a.m. to focus on the 12-hours with the most Snapchat usage traffic. From the hour of the day plot (top left), we see that sentiment rises over the course of the morning, dips in the afternoon ("afternoon slump"), and rises back up as the day progresses into the night. The weekday plot (top right) demonstrates that sentiment consistently increases over the week, with Monday having the lowest and the weekends having the highest sentiment. These findings agree with past work showing that people are generally happier during weekends or after work (Dodds et al. 2011; Golder and Macy 2011; Li, Wang, and Hovy 2014).

Sentiment varies with weather when controlling for time. Since we observe such a clear trend in sentiment with regards to time, we want to isolate the impact of time from the weather. Temperature can change dramatically from night to day and from season to season, therefore any relationships we observe with respect to hourly weather may be confounded by time. In the bottom row of Figure 6, we display how sentiment changes with respect to time, disaggregated by the temperature values binned into 10 equally-sized deciles. We observe that when the temperature is warmer (darker lines), the sentiment is consistently higher than if it was cooler (lighter lines), at *every time point*. This indicates that the link between weather (temperature) and sentiment exists even when controlling for time.

Geographical differences in mood sensitivity to weather. Lastly, we explore whether people in different locations react differently to weather conditions. Figure 7(a) shows the Pearson correlation coefficient of temperature and sentiment for each US state. We notice a distinct geographical trend, with people in the Southeast (i.e., Florida, possibly due to the tropical climate) least affected by changes in temperature and people on both coasts relatively more affected.

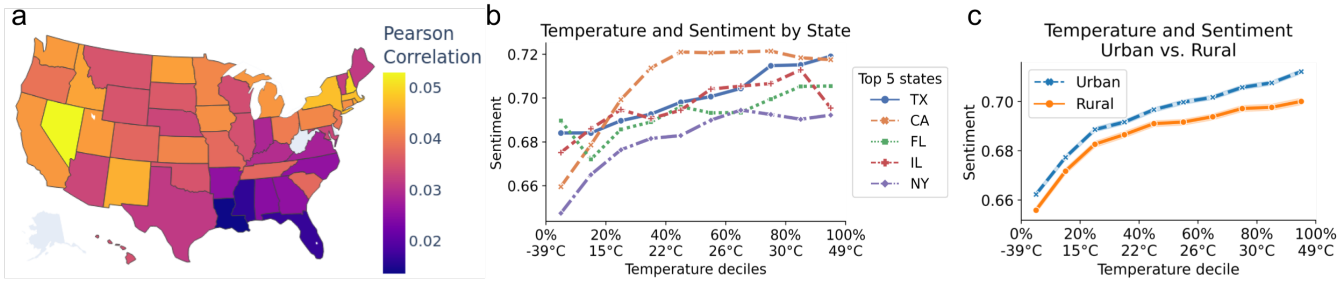

Figure 7: Geographical location plays a role in mood sensitivity to temperature but not population density. (a) Choropleth of the Pearson correlation coefficient of sentiment and temperature for every state. Only significant coefficients ($p < 0.05$) are shown. (b)-(c): Sentiment for each temperature decile (lower bound as displayed) (b) for the top 5 states by data size and (c) for urban and rural areas (according to the 2010 census list of urban areas).

We also conduct a detailed analysis of the top five states for which we have the most data points in Figure 7(b). We observe that the relationship between temperature and sentiment is mostly linear. Compared with other states, New Yorkers experience lower sentiment than all other states regardless of weather. Florida has the flattest line, meaning that Floridians' sentiment is least affected by temperature compared to other states (also confirmed by the low correlation between temperature and sentiment for Florida in Figure 7(a)). One notable exception is California, where people have overwhelmingly higher expressed sentiment than the rest of the country even when the temperature is only slightly above average. However, Californians also experience the steepest decline in sentiment when the temperature falls below average. Indeed, we find that cities in California are most impacted by adverse changes in the temperature and by rainfall (see the Appendix).

Urban vs. rural. Figure 7(c) explores the relationship between temperature and sentiment in urban vs. rural areas. We find that Snaps from rural areas consistently score lower than urban areas in sentiment. However, mood detected in Snaps from both urban and rural areas seems similarly affected by changes in temperature.

Summary. We find that sentiment is correlated with current, historical, and changes in weather, implying that past weather has a lasting effect on mood. We also find that links between weather and sentiment exist even when controlling for time and population density, but there are geographical differences in the sensitivity of mood to temperature

Discussion

RQ1. *Can including external contextual features (weather, location, and time) in language models improve sentiment detection?* Through experiments, we demonstrate that including contexts in addition to text significantly improves sentiment detection performance. Weather is the most important context feature. In particular, we highlight the importance of historical weather, with three-day historical weather being the most important historical timeframe of weather. We also replicate these findings on a Twitter dataset.

RQ2. *How do contextual factors, and particularly historical and current weather, impact expressed sentiment?* Echoing previous psychological literature (Howarth and Hoffman 1984; Keller et al. 2005; Denissen et al. 2008) as well as empirical studies on social media (Li, Wang, and Hovy 2014; Coviello et al. 2014; Baylis et al. 2018; Wang, Obradovich, and Zheng 2020), we find evidence that weather is linked to mood. The correlations we found between weather and sentiment, while small, are statistically significant. We further underscore the importance of historical weather data on current mood, emphasizing the lasting impact of weather. Finally, we connect weather features with time and location features to show that weather effects are observable regardless of time or location. There are geographical patterns in mood sensitivity to weather, possibly due to climate variations or other factors.

Even though we show that weather is associated with the expressed sentiment, we find that the texts themselves mostly do not pertain to weather. During our annotation process, we find that weather was seldom explicitly stated in the text ($< 1\%$). This implies that weather might be impacting users' emotional states *implicitly* rather than *explicitly*. There is evidence to suggest that weather may have an impact on individual happiness and subjective well-being beyond momentary affective states (Tsutsui 2013; Schwarz and Clore 1983). We leave further research on this topic for future work.

Limitations. We discuss several limitations of our work. Our work considers *expressions* of sentiment on social media, which we use as a proxy for actual user mood. In studying historical weather, we also assume that users are relatively immobile. The timeframe of our study coincides with the COVID-19 pandemic, during which Snapchat users exhibited limited movement (Yang et al. 2021). Furthermore, we believe our dataset is large enough to compensate for a small subset of users traveling at any time. While the pandemic may also have affected users' emotional states, the Twitter dataset was from an earlier period and validates our work. Another limitation is the possibility of seasonal bias in our Snapchat dataset. To alleviate this concern, we replicated our findings on the Twitter dataset, which was collected over a longer period over an entire year. Finally, we recognize that our methodology is not causal. Since weather is extremely unlikely to be causally affected or confounded by other factors, any correlations we observe between weather and sentiment are very likely caused by the weather's influence on expressed sentiment.

Ethics statement. Any research studies that deal with user information are at risk of distributing sensitive user information and thus must be conducted with care. We took steps to ensure that the research is carried out ethically and that the data is handled securely. The examples used in this paper are fictional and not based on users' personal information. In accordance with Snap's commitment to protecting user data privacy, all analyses for this paper are restricted to public social media posts. Additionally, all modeling and analyses were done through Snap's secure internal data storage and no data was stored on local computers or systems outside of Snap at any point. The Twitter data will only be identified by Tweet IDs, which complies with Twitter's data sharing policy (Twitter 2020).

Broader impact. For the broader ICWSM community, our work highlights the promising direction of migrating content understanding models from using *user-specific* to *user-agnostic* features. We demonstrate the potential of including external contextual features, rather than user-level features, to improve content understanding. Weather, location, and time are contextual signals that can be readily available but have remained relatively underutilized in content understanding. As protecting user privacy is an important goal in AI-driven research, we believe our work is of value for researchers in this field.

Our findings also suggest that online social platforms can moderate their content based on contextual signals such as weather conditions. One application of this is utilizing weather signals to improve or enhance user mood. For instance, in less than ideal weather conditions, social media platforms can display more positive content to help lift users' moods. Alternatively, when the weather is good, platforms can also encourage users to go outside to enjoy the weather: Keller et al. (2005) showed that the beneficial effects of warm weather are, perhaps not surprisingly, only observed when the participants are spending time outdoors. Finally, weather-related cues can also be integrated into social media as interactive features to make the weather experience—positive or negative—more enjoyable.

Another direction of application surrounds recent efforts to design technology for situational visual impairments (Tigwell, Menzies, and Flatla 2018), which refers to environmental conditions that negatively affect one's ability to complete tasks. We hope our work motivates developers to consider supporting interfaces that are appropriate for rare and extreme weather conditions.

Conclusion

Our work explores how current weather and historical weather, combined with other external features of location and time, impact users' expressed sentiment through a large-scale empirical study of 8 million Snapchat *Our Story* posts in the US. We highlight the added value (3% improvement) of including contextual features, especially short-term historical weather, in predicting user sentiment. Our results also generalize cross-domain to Twitter. Furthermore, we find that historical weather has a lasting impact on sentiment. The effects of weather on sentiment exist even after controlling for time and location, though there are geographical differences in mood sensitivity to weather. Our work shows the impact weather has on our emotions and on how we express ourselves online, which could be better utilized in content understanding models.

Appendix

Annotation Details

Snapchat annotations. The expert annotators were screened for annotation quality in English and are highly experienced in annotating textual and visual social media data. They were asked to label each Snap caption as negative, neutral, positive, or n/a. N/a is used for captions that they deemed non-English, clearly spam, or whose sentiment is unclear. Captions labeled n/a by any annotator were discarded.

Twitter annotations. Twitter annotations were done through Mturk in HITs (batches) of 40 Tweets. Each HIT is annotated by five unique workers because previous work showed that five non-experts can approximate expert quality in sentiment (valence) detection (Snow et al. 2008). To qualify, workers must be located in the US, have more than 100 approved HITs, and have a HIT approval rate $> 95\%$. Workers are given instructions and a few examples. To improve the reliability of the non-expert ratings, we extend the three-point scale to an easier five-point scale (Preston and Colman 2000): very negative, somewhat negative, neutral, somewhat positive, and very positive. N/A can be assigned to any Tweets that are spam, non-English, and undetermined. Workers are compensated $0.60 for each HIT of 40 Tweets.

We prune any annotations that did not pass our reliability checks. Following Amidei, Piwek, and Willis (2019), we compute the Goodman and Kruskal's Gamma (Goodman and Kruskal 1979), a correlation coefficient designed for ordinal annotations, for every pair of annotators. We first remove annotators whose average Gamma is below 0.5 (large correlation (Rosenthal 1996)) and then remove HITs/batches whose average inter-rater gamma is below 0.5. We also remove Tweets that are labeled n/a by any annotators. Finally, we remove Tweets for which there is no majority sentiment consensus on positive (including very and somewhat positive), neutral, and negative (including very and somewhat negative).

Sentiment distribution. Consistent with prior work on Snap *Our Story* data (Alghamdi et al. 2020), the sentiment distribution is imbalanced, with the majority of Snaps being positive (52% positive, 33% neutral, and 15% negative). We found the same imbalance in the Twitter/Instagram data (64% positive, 29% neutral, and 7% negative).

	Temperature ↑	Pressure ↓	Humidity ↓	Rain ↓	Clouds ↓
1	Los Angeles, CA	Fort Lauderdale, FL	Las Vegas, NV	San Diego, CA	San Diego, CA
2	Seattle, WA	Miami, FL	Austin, TX	Los Angeles, CA	Las Vegas, NV
3	Cleveland, OH	Newark, NJ	Memphis, TN	Houston, TX	Los Angeles, CA
4	Memphis, TN	Tampa, FL	San Antonio, TX	San Francisco, CA	San Antonio, TX
5	Philadelphia, PA	Los Angeles, CA	Detroit, MI	Anaheim, CA	Cleveland, OH

Table A1: The top five cities where the sentiment is most positively ↑ or negatively ↓ correlated with changes in weather.

Experimental Details

The hyperparameters we searched are $L2$ regularization: $\{0.01, 0.001, 0.00001\}$, dropout (Srivastava et al. 2014): $\{0.0, 0.2\}$, hidden dimensions (of all layers): $\{16, 32, 64, 128, 256\}$, number of layers (after combination): $\{2, 4, 8\}$, number of layers (before combination): $\{4, 8, 16, 32\}$, and whether to apply layer normalization after every dense layer. We fix the activation function to be ReLU. The optimizer is Adam (Kingma and Ba 2015) with a learning rate of 0.001. The learning rate 0.001 is chosen from $\{0.01, 0.001, 0.0005, 0.0001\}$ from a initial smaller subset of hyperparameters. All models are trained until no improvement on the validation set after 20 epochs.

City-level Analysis

To conduct city-level analysis, we take the top 50 cities with the most number of data points, which covers 41% of all data. The five cities with the most positive average cities are in Southern California and Texas—Anaheim, CA, San Diego, CA, Burbank, CA, Arlington, TX, Dallas TX—whereas the five cities with the most negative average sentiment are in Ohio and the East Coast—Cincinnati, OH, Newark, NJ, Brooklyn NY, Cleveland, OH, and Boston MA. To examine which cities are most impacted by changes in the weather, we rank the cities by their Pearson Correlation coefficient between their sentiment and weather features. Table A1 shows that cities in California (Los Angeles, San Diego, Anaheim, and San Francisco) are most positively impacted by warmer temperatures and most negatively impacted by rain and clouds. Florida is most negatively impacted by high atmospheric pressure, and Texas is negatively impacted by humidity.

Acknowledgements

The authors thank the reviewers for their thoughtful comments as well as Emilio Ferrara for his useful feedback.

References

Alghamdi, N.; Alageeli, N.; Sharkh, D. A.; Alqahtani, M.; and Al-Razgan, M. 2020. An eye on Riyadh tourist season: Using geo-tagged Snapchat posts to analyse tourists impression. In *ICCIS '20*, 1–6.

Amidei, J.; Piwek, P.; and Willis, A. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *INGl '19*, 344–354.

Anderson, C. A.; Anderson, K. B.; Dorr, N.; DeNeve, K. M.; and Flanagan, M. 2000. Temperature and aggression. In *Advances in Experimental Social Psychology*, volume 32, 63–133. Academic Press.

Anderson, K. E. 2015. Getting acquainted with social networks and apps: Snapchat and the rise of ephemeral communication. *Library Hi Tech News*.

Barbieri, F.; Camacho-Collados, J.; Neves, L.; and Espinosa-Anke, L. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of EMNLP '20*, 1644–1650.

Baylis, P.; Obradovich, N.; Kryvasheyeu, Y.; Chen, H.; Coviello, L.; Moro, E.; Cebrian, M.; and Fowler, J. H. 2018. Weather impacts expressed sentiment. *PloS one*, 13(4): e0195750.

Beniaguaev, D. 2017. Historical Hourly Weather Data 2012-2017. Kaggle.

Busse, M. R.; Pope, D. G.; Pope, J. C.; and Silva-Risso, J. 2015. The psychological effect of weather on car purchases. *The Quarterly Journal of Economics*, 130(1): 371–414.

Chang, S.-C.; Chen, S.-S.; Chou, R. K.; and Lin, Y.-H. 2008. Weather and intraday patterns in stock returns and trading activity. *Journal of Banking & Finance*, 32(9): 1754–1766.

Chaudhuri, A. 2019. *Visual and text sentiment analysis through hierarchical deep learning networks*. Springer.

Chen, X.; Cho, Y.; and Jang, S. Y. 2015. Crime prediction using Twitter sentiment and weather. In *2015 Systems and Information Engineering Design Symposium*, 63–68.

Christenson, J. A. 1979. Urbanism and community sentiment: extending Wirth's model. *Social Science Quarterly*, 60(3): 387–400.

Coviello, L.; Sohn, Y.; Kramer, A. D.; Marlow, C.; Franceschetti, M.; Christakis, N. A.; and Fowler, J. H. 2014. Detecting emotional contagion in massive social networks. *PloS one*, 9(3): e90315.

Cunningham, M. R. 1979. Weather, mood, and helping behavior: Quasi experiments with the sunshine samaritan. *Journal of personality and social psychology*, 37(11): 1947.

Denissen, J. J.; Butalid, L.; Penke, L.; and Van Aken, M. A. 2008. The effects of weather on daily mood: A multilevel approach. *Emotion*, 8(5): 662.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT '19*, 4171–4186.

Dey, A. K.; and Abowd, G. D. 1999. Towards a better understanding of context and context-awareness. Technical Report GIT-GVU-99-22, Georgia Institute of Technology.

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, 6(12): e26752.

Doksæter Sivle, A.; and Kolstø, S. D. 2016. Use of online weather information in everyday decision-making by laypeople and implications for communication of weather information. *Meteorological Applications*, 23(4): 650–662.

Donegan, B. 2016. Weather terms you've heard, but probably don't know what they mean. *The Weather Channel*. Accessed: 2021-09-27.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Fornaciari, T.; Uma, A.; Paun, S.; Plank, B.; Hovy, D.; and Poesio, M. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *NAACL-HLT '21*, 2591–2597.

Golder, S. A.; and Macy, M. W. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051): 1878–1881.

Goodman, L. A.; and Kruskal, W. H. 1979. *Measures of Association for Cross Classifications*. Springer.

Guéguen, N. 2013. Weather and courtship behavior: A quasi-experiment with the flirty sunshine. *Social Influence*, 8(4): 312–319.

Habib, H.; Shah, N.; and Vaish, R. 2019. Impact of contextual factors on Snapchat public sharing. In *CHI '19*, 1–13.

Hannak, A.; Anderson, E.; Barrett, L. F.; Lehmann, S.; Mislove, A.; and Riedewald, M. 2012. Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *ICWSM '12*, 479–482.

Harmatz, M. G.; Well, A. D.; Overtree, C. E.; Kawamura, K. Y.; Rosal, M.; and Ockene, I. S. 2000. Seasonal variation of depression and other moods: a longitudinal approach. *Journal of Biological Rhythms*, 15(4): 344–350.

Howarth, E.; and Hoffman, M. S. 1984. A multidimensional approach to the relationship between mood and weather. *British Journal of Psychology*, 75(1): 15–23.

Hu, M. R.; and Lee, A. D. 2020. Outshine to outbid: Weather-induced sentiment and the housing market. *Management Science*, 66(3): 1440–1472.

Keller, M. C.; Fredrickson, B. L.; Ybarra, O.; Côté, S.; Johnson, K.; Mikels, J.; Conway, A.; and Wager, T. 2005. A warm heart and a clear head: The contingent effects of weather on mood and cognition. *Psychological Science*, 16(9): 724–731.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR '15*.

Kööts, L.; Realo, A.; and Allik, J. 2011. The influence of the weather on affective experience. *Journal of individual differences*.

Li, J.; Wang, X.; and Hovy, E. 2014. What a nasty day: Exploring mood-weather relationship from Twitter. In *CIKM '14*, 1309–1318.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP '15*, 1412–1421.

Mitchell, L.; Frank, M. R.; Harris, K. D.; Dodds, P. S.; and Danforth, C. M. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5): e64417.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, 1–17.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. In *EMNLP '20: System Demonstrations*, 9–14.

Parsons, A. G. 2001. The association between daily weather and daily shopping patterns. *Australasian Marketing Journal*, 9(2): 78–84.

Pavlick, E.; and Kwiatkowski, T. 2019. Inherent disagreements in human textual inferences. *TACL*, 7: 677–694.

Peters, M. E.; Ruder, S.; and Smith, N. A. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 7–14.

Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; and Gurevych, I. 2020. Adapterfusion: Non-destructive task composition for transfer learning. arXiv:2005.00247.

Poesio, M.; and Artstein, R. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, 76–83.

Preston, C. C.; and Colman, A. M. 2000. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1): 1–15.

Rosenthal, J. A. 1996. Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research*, 21(4): 37–59.

Rosenthal, N. E. 1984. *Seasonal affective disorder: a description of the syndrome and preliminary findings with light therapy*. US Department of Health and Human Services, National Institutes of Health.

Schwarz, N.; and Clore, G. L. 1983. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3): 513.

Simonsohn, U. 2007. Clouds make nerds look good: Field evidence of the impact of incidental factors on decision making. *Journal of Behavioral Decision Making*, 20(2): 143–152.

Snapchat. 2018. Our Story. Accessed 2021-09-17.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In *EMNLP '08*, 254–263.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1): 1929–1958.

Tigwell, G. W.; Menzies, R.; and Flatla, D. R. 2018. Designing for situational visual impairments: Supporting early-career designers of mobile content. In *DIS '18*, 387–399.

Tsutsui, Y. 2013. Weather and individual happiness. *Weather, Climate, and Society*, 5(1): 70–82.

Twitter. 2020. Developer Agreement and Policy. Accessed 2022-03-28.

US Census Bureau. 2021. Urban Area Facts. Accessed 2021-10-04.

Van de Vliert, E.; and Van Lange, P. A. 2019. Latitudinal psychology: an ecological perspective on creativity, aggression, happiness, and beyond. *Perspectives on Psychological Science*, 14(5): 860–884.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS '17*, 5998–6008.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP '18 Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355.

Wang, J.; Obradovich, N.; and Zheng, S. 2020. A 43-million-person investigation into weather and expressed sentiment in a changing climate. *One Earth*, 2(6): 568–577.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP '20: System Demonstrations*, 38–45.

Yang, C.-Y.; Jhang, L.-J.; and Chang, C.-C. 2016. Do investor sentiment, weather and catastrophe effects improve hedging performance? Evidence from the Taiwan options market. *Pacific-Basin Finance Journal*, 37: 35–51.

Yang, Q.; Wang, W.; Pierce, L.; Vaish, R.; Shi, X.; and Shah, N. 2021. Online communication shifts in the midst of the Covid-19 oandemic: A case study on Snapchat. In *ICWSM '21*, 830–840.

Zhong, B.; and Zhou, Y. 2012. "Under the weather": The weather effects on US newspaper coverage of the 2008 Beijing Olympics. *Mass Communication and Society*, 15(4): 559–577.