CMSC396H Final Writeups

Katherine Tootchen
University of Maryland
ktootch@terpmail.umd.edu

Patrick Scott
University of Maryland
pscott1@terpmail.umd.edu

Xinyu Yang
University of Maryland
yxy0302@gmail.com

Abstract

The relationship between sentiment and weather has been studied in the past with research focusing on how the weather and changes in it have affected sentiment on social media. However, this work has largely focused on just the weather, changes of weather, and sentiments of tweets but has not factored in location and instead treats all regions the same despite regional differences in climate and population that may affect the sentiment. This work examines the impact that weather has on online sentiment on a regional basis by adding region as one of the factors in our regression in order to obtain the specific impact of each region on their sentiment based on weather. After doing this analysis, little correlation was found between the temperature and sentiment on twitter in the different regions of the United States.

1 Introduction

People's mood's are impacted by a variety of things such as how stressed they may be. It is commonly thought that the weather is also a larger determinant of people's moods, with people being happier in better weather, and less happy when it is gloomy out. In the past, there have been studies on how the sentiment of tweets is related to the weather, and if we can tell the general mood of people based on the weather. By comparing the sentiment analysis of posts on social media with the weather at the time of a post, researchers examined the correlation between the weather and the general mood people had and wished to share online.

However, one thing that was left off of this research was examining the specific country that the tweets were from, and how cultural, climate, and geographical differences may be impacting the mood that people have on the weather. With the diversity of cultures and climates throughout the world, it would be naive to assume that everyone has the same mood depending on the weather. Some may be happy to see it raining, or even when it is cold. Some may be happy to see the snow once in a while, and others may be frustrated that they have to commute through lots of snow all the time.

We use similar methods to previous research work, but by selecting specific locations we can get more holistic views of how weather affects mood, and can see if the mood depending on the weather significantly varies depending on what country one is in. We can estimate specific effects that the geographical location has on how people feel about the weather, and the specific ways that the sentiment based on the weather varies with one's location.

2 Related Work

We have looked into two related papers for our research.

2.1 Paper 1: Tweetin' in the Rain: Exploring Societal-scale Effects of Weather on Mood

This paper does an exploration of the impact of weather on people's mood, using twitter sentiment

analysis. The authors use a Partial Dependence Plot and Bagged Decision Trees for analysis of correlation. However, it does not include many other factors beyond the weather and twitter sentiment to just make a very general analysis of sentiment. In our research, we include sentiment analysis in different area in the United States, which will yield a more detailed result in analyzing the correlation between weather and sentiments.

2.2 Paper 2: Sunshine with a Chance of Smiles: How Does Weather Impact Sentiment on Social Media?

Similar to Paper 1, this paper explores a similar concept, while looking more at data from Snapchat. This paper examines changes in weather in addition to what the weather is, and location data. It has a solid model for sentiment analysis, and also includes features such as location and time in sentiment analysis as they might be relevant to determining mood as well. For our database, we used Twitter data instead of Snapchat data.

3 Approach/Solution

3.1 Retrieving Twitter Database

For our database, we used data from September 2009 - January 2010 in Cheng-Caverlee-Lee September 2009 - January 2010 Twitter Scrape. It contains 3,844,612 tweets from 115,886 users and locations are labeled by city and state. Each tweet was identified by a unique tweet id, a user id, its text, a timestamp, and the location. The data was provided as a text file with tabs separating values.

3.2 Pre-Processing

Some pre-processing was needed to prepare the data for analysis. One of the difficulties with reading the data in was that tweets could contain newline characters or tabs that interfered with reading the file into a more usable data frame. To fix this, we manually went through a number of particular tweets with issues causing cascading issues with the data. Others tweets whose spacing only impacted their own readability were able to be removed in an automated process that checked for missing information in a row of data. A few tweets with problematic spacing were identified as the postings of bots due to repetition of the same promotional message numerous times and was removed rather than fixed as this study aims to analyze moods and sentiments of which bots do not have.

3.3 Obtaining Regional and Weather Data

We used the geopy package to convert the text location specifying city and occasionally state into latitude and longitude coordinates we would later use to fetch weather data and identify regions. There were a total of 6052 unique location in the data that we processed. The python package meteostat was used in order to retrieve the average temperature corresponding to the location and day a tweet was posted. In the case where a temperature could not be retrieved, a placeholder value of -1000 that would be removed at the time for analysis. Temperatures were recorded in degrees Celsius. Again using the geopy package to get coordinates, we were able to break down the locations of each tweet into the Northeast, Southeast, Midwest, Southwest, West, or not in the United States. The breakdown of these coordinates is shown in table 1. For the area of overlap between the West and Southwest, that was considered the Southwest.

Table 1: US regions based on Latitude and Longitude

	Longitude Bounds	Latitude bounds
Northeast	$-81W$ to $-65W$	$40N$ to $55N$
Southeast	$-95W$ to $-70W$	$20N$ to $40N$
Midwest	$-102W$ to $-81W$	$40N$ to $55N$
Southwest	$-115W$ to $-95W$	$20N$ to $40N$
West	$-130W$ to $-102W$	$20N$ to $55N$

3.4 Analyzing Sentiments in Tweets

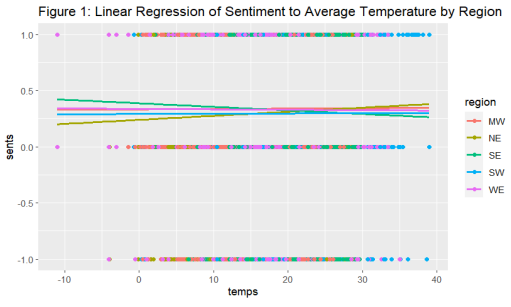
To analyze the sentiment in each tweet, we used a NLTK sentiment documentation at https://www.nltk.org/api/nltk.sentiment.html with Regex to parse the tweet content of each tweet. We chose to use this tool because we have prior experience of using it, even though using Regex is relatively slower, it yields pretty accurate analysis. NLTK's sentiment analysis relies on a lexicon-based approach which includes a wide range of words, each assigned a sentiment polarity score (positive, negative, or neutral). Each sentiment was then given a numerical score, of -1 for negative sentiment, 0 for neutral, and 1 for positive.

4 Results

Table 2: Regression of Regions and Temperature on Sentiments

	Coefficient	P-value
Northeast	−0.0021	0.175
Southeast	0.0007	0.524
Midwest	0.0016	0.322
Southwest	−0.0014	0.227
West	0.0013	0.232

The results of our regression are shown in table 2. As all of the coefficients have high p values, none of them can really be concluded to be very significant. Similarly, this model also had an R^2 value of 0.001 so it is not really a great predictor.



5 Conclusion

Our results have shown that there is a lack of correlation overall between weather and sentiments. This contradicts with previous studies and possible factors that may have contributed to our results include:

- Only average temperature for that day was used to encompass weather conditions
- A different process for sentiment analysis giving a larger range of values could influence results
- Use of a regression model for analysis

An additional note for concern was the presence of bots and business accounts in the set of tweets we analyzed. Posts from such accounts are unlike to have the sentiments of their owners affect the content and could lead to undue inflation of numbers.

Another thing we could change in the future is using a non binary representation of sentiment where instead of simply positive, negative, and neutral, they are on a gradient, making the data more precise and results more interpretable.

To further study the correlation between weather and sentiments in different locations, we would want to incorporate additional weather data, analyze more tweets, and look at regions outside of the United States.

References

[1] Hannak, A., Anderson, E., Feldman Barrett, L., Lehmann, S., Mislove, A., & Riedewald, M. (2021). Tweetin' in the Rain: Exploring Societal-Scale Effects of Weather on Mood. Proceedings of the International AAAI Conference on Web and Social Media, 6(1), 479-482. https://doi.org/10.1609/icwsm.v6i1.14322

[2] Jiang, J., Murrugara-Llerena, N., Bos, M. W., Liu, Y., Shah, N., Neves, L., & Barbieri, F. (2022). Sunshine with a

Chance of Smiles: How Does Weather Impact Sentiment on Social Media?. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 393-404. https://doi.org/10.1609/icwsm.v16i1.19301