Quantifying Correlation Between Weather and Twitter Public Sentiment

Frederico Araujo
UNI: fca2118

Uday Mukhija
UNI: um2158

Yewen Zhou
UNI: yz4175

Abstract

In this project we try to quantify the correlation between weather parameters and social media sentiment, specifically on Twitter. In order to do that we build regression models based on data collected through streaming APIs and later on compare those models to find the one most suitable for future predictions, in the sense that its error when tested is the lowest. Sentiment classification is done by a combination of two different classifiers and 5 different regression models will be build and their errors measured in order to find the best one.

1. Introduction

In the past few years, we have seen a huge growth in the use of microblogging platforms such as Twitter. Twitter currently has over 350 million users, with its major purpose being news articles, conversation and public relations. Spurred by this growth, a lot of companies and media organizations are now using data to analyze what people think and feel about their services, in order to create better market strategies. These analyses help businesses and researchers define goals, identify and find the audience better. It helps in improving competitor analysis and improving brand awareness. This domain has reshaped market research and the way businesses work. Recommendation of items provided by recommendation systems are a great example of applications of sentiment analysis.

This project is aimed at creating an automatic sentiment analysis classifier that can work with large amounts of data, the accurate and automatic sentiment classification of an unknown tweet stream as well as finding correlation between Twitter data and the weather in New York. We picked Twitter because it offers a better approximation of sentiment compared to conventional reports like blogs or articles. This type of sentiment analysis is useful for those who are trying to research a product or service, or for businesses researching public opinion of their company.

2. Related Works

In this section we analyze three papers which also tried to find a relationship between weather and social media sentiment. We analyze how their methods differed from what we try to achieve in this project and how their results reflect those different methods.

2.1. Weather impacts expressed sentiment [1]

This paper also tried to find a relationship between social media sentiment and weather, however they used static data from both Facebook and Twitter for that purpose, while our project relies solely on streamed Twitter data. We made this choice because limiting the data to one platform reduces the variables one has to control to make a meaningful analysis, and also because we have found that the weather information gathered in real time is more reliable than one from a dataset possibly created years ago.

The paper also uses more concrete weather parameters in their regression, notably percentage cloud cover and relative humidity, while one of the parameters our project uses is weather status, an abstract description of the current weather condition. While the more concrete approach can give results that showcase more precisely how those parameters affect sentiment, we believe that a general description is helpful to determine the sentiment for especially different weather situations, e.g. sunny vs. rainy.

The paper also takes information from the entirety of the continental United States, while we choose to focus on the city of New York. We limited our dataset this way because we believe different cities could have considerably different results, and thus one global model would be unreliable.

Finally, the results showcased in the paper were based on one linear regression model based on their parameters, while we chose to test multiple models, including the more robust Support Vector Regression to model our data, and compare their results. Having only one simple model could potentially not result in a correct assignment, and so we made the decision to search for the best.

2.2. Tweetin' in the rain [2]

In this paper, the authors examined the external factors of weather on aggregate twitter sentiment by using machine learning techniques. They used a corpus of over 1.5 billion tweets, which are cleaned manually using existing sentiment inference algorithms based on token lists, and aggregated into 20 largest U.S. metropolitan areas according to the U.S. Census Bureau. Then, the researchers collect weather data by using Mathematica's WeatherData package for each metropolitan area. Finally, they created a Decision Tree classification algorithm so that for every hour, the model predicts the aggregate sentiment.

The approach adopted in this paper differs from our approach in the following ways. First, they used a large corpus of historical twitter data instead of streaming tweets alive. Their techniques are offline whereas our implementation is online. Second, the model they used is a Decision Trees classifier while in our approach, we will be using regularized linear regression models and also multi-layer perceptrons to hopefully push the boundary of accuracy into the next level. Thirdly, they used Amazon Mechanical Turks and paid human readers to evaluate the sentiment on tweets and compared the results to their machine learning models. However, due to economic reasons, we will be cross-referencing different sentiment inference models to validate our results.

Their models also have advantages and drawbacks. Because tweets are limited to 140 characters, many sentiment inference techniques will fail. Their approach to construct a Twitter-specific token list by using tweets themselves proved to be robust during the evaluation process. However, it is not a convincing result when they only implemented the decision trees model and the only evaluation metrics they used was the ROC curve. To amend the possible limitation incurred by using only the ROC curve, we will adopt a more comprehensive approach to not only evaluate our model using common statistics such as accuracy, precision, recall, but also include graph metrics such as ROC curve and PR curve.

2.3. Twitter sentiment analysis introduction [3]

In this paper, authors classified sentiment into "Positive", "Neutral" and "Negative". Twitter API was used to collect the data and it was classified manually. Three classifiers were used, a Naïve Bayes classifier, Maximum Entropy and Support Vector Machine. Stanford Classifier and OpenNLP packages were used for the MaxEnt implementation. The Stanford Classifier gave bad results with the default parameter settings, and it did not improve much by changing the smoothing constants. However, MaxEnt performed considerably better with the OpenNLP package. Features were picked for each class based on the frequency of occurrence, due to the large number of different features. When extended to handle three classes, the Naive Bayes classifier only obtained 40% accuracy.

Some other problems faced while deploying the algorithm used in this paper were that this classified the overall sentiment of the tweet. The polarity changes depending on whose perspective one is viewing the tweet from. In our approach, to obtain better accuracy, we are not using a single classifier, but combining multiple classifiers. We are also ensuring that our training data has cleaner labels.

3. Current Progress Including Data, Methods, System

The data acquisition method to build the model correlating weather and Twitter public sentiment is composed of 5 steps: acquisition of tweets, processing of tweets, acquisition of weather data, combination of tweet and weather data and sentiment classification, and finally building the regression model. This process is outlined in the figure below, which shows the Airflow DAG to build it.

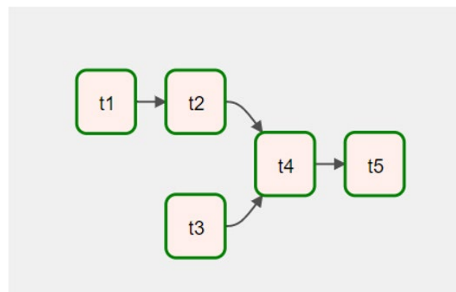

Figure 1: Project DAG in Airflow

To acquire the tweet data we make use of the Twitter API through the 'tweepy' module, which allows us to stream incoming tweets for analysis. We limit the data to tweets sent from the city of New York and using the English language. This allows us to limit the variables concerning the correlation we're trying to find. To further process the tweet text we remove any usernames mentioned in the tweet, as they are not relevant to sentiment classification. Tweets with fewer than 4 words are also removed entirely from the dataset, as classification on them is prone to being incorrect. We chose to keep hashtags, as they can be a valuable indication of sentiment, as opposed to usernames, which are immutable. Along with the text of the tweet, we also collect information about the exact time of creation and geographical information. The time of creation will be used to associate the tweet with weather conditions at that time. As our project will be limited to the city of New York, geographical information taken will not be used, but it is collected for possible future expansion. Figure 2 shows the output table from this step.

1	Date,Text,Box
2	2021-11-30-19-51,Completely got my ass kicked today but the stomach flu. Beating me and Luca up. Yay!,"[[-75.563587, 38.788657], [-75.563587, 41.357424], [-73.88506, 4
3	2021-11-30-19-51,1st lmfaooooo,"[[-73.962582, 40.541722], [-73.962582, 40.800037], [-73.699793, 40.800037], [-73.699793, 40.541722]]"
4	2021-11-30-19-51,"Seeing Cam Thomas in high school, then LSU, now the NBA makes it all worth it","[[-73.933612, 40.785365], [-73.933612, 40.91533], [-73.765271, 40.915
5	2021-11-30-19-52,Incident on #2Line Both directions at Flatbush Av - Brooklyn College https://t.co/DhHHOUl1by,"[[-74.041878, 40.570842], [-74.041878, 40.739434], [-73.8
6	2021-11-30-19-52,Incident on #5Line at Flatbush Av - Brooklyn College https://t.co/HWSxvRiNHf,"[[-74.041878, 40.570842], [-74.041878, 40.739434], [-73.855673, 40.7394
7	2021-11-30-19-52,"ðŸ‚ðŸŽ¶ðŸðŸŽ¶ðŸ‚ ðŸŽ¤ðŸµðŸ–¤ðŸ§”ðŸ»ðŸŽ¹ ðŸ«€ðŸŽžðŸ‘ðŸ¥¨ðŸŒ»ðŸª²ðŸ¦–ðŸ¦‹ðŸ«ðŸ‘¾ðŸ‘©ðŸ»â€ðŸŽ¤ðŸŒ¸ðŸ¦„ LOVE YOU!!!ðŸ–¤ðŸ’ @ Central Park
8	2021-11-30-19-52,i found a beer Advent Calendar @ConsumersBeer and thought I'd share each day with you https://t.co/ZusXWJ0LbJ,"[[-79.76259, 40.477383], [-79.76259,
9	2021-11-30-19-52,Updated: Incident on #ELine at Queens Plaza https://t.co/XtbYhoh9zc,"[[-73.962582, 40.541722], [-73.962582, 40.800037], [-73.699793, 40.800037], [-73
10	2021-11-30-19-52,Updated: Incident on #FLine at Queens Plaza https://t.co/Q1UOHFXk0V,"[[-73.962582, 40.541722], [-73.962582, 40.800037], [-73.699793, 40.800037], [-
11	2021-11-30-19-52,Updated: Incident on #MLine at Queens Plaza https://t.co/Rw4QiXmZ0S,"[[-73.962582, 40.541722], [-73.962582, 40.800037], [-73.699793, 40.800037], [-
12	2021-11-30-19-52,IOHK publicizes Cardano (ADA) block measurement improve as it strikes to â€˜scalingâ€™ part: Cardano (ADA) is heading iâ€¦ https://t.co/ip7eitmOZD,"[[-
13	2021-11-30-19-52,Dream start for #NYCFC courtesy of a Santi Rodriguez goal. This could be a barnburner #MLSCupPlayoffs,"[[-73.157725, 40.725255], [-73.157725, 40.7550
14	2021-11-3 comes across Mariska ðŸ˜- heâ€™s out there living my dreams of not only seeing her but heâ€¦ https://t.co/5D1sHs0tOo,"[[-79.76259, 40.477383], [-79.76259, 4
15	2021-11-30-19-52,"Emma Wade-Smith, Uk Consul General in NY and Her Majestyâ€™s Trade Commissioner to North America. Welcome!! https://t.co/xNUsd2MDD4","[[-74
16	2021-11-30-19-52,Cautiously optimistic,"[[-74.026675, 40.683935], [-74.026675, 40.877483], [-73.910408, 40.877483], [-73.910408, 40.683935]]"
17	2021-11-30-19-52,itâ€™s only wavy kuz thereâ€™s an on/off switch,"[[-74.041878, 40.570842], [-74.041878, 40.739434], [-73.855673, 40.739434], [-73.855673, 40.570842]
18	2021-11-30-19-52,Guns and Roses Use Your Illusion is nowhere near as good as Beggars Banquet Let it Bleed Sticky Fâ€¦ https://t.co/ygS7ujKjsn,"[[-74.026675, 40.683935],
19	2021-11-30-19-52,Some of the highest rain totals ever observed have been in Hawaii. This will mostly be a good thing since some partâ€¦ https://t.co/52DAW5pChm,"[[-73.
20	2021-11-30-19-52,Uh oh for someone alright,"[[-73.962582, 40.541722], [-73.962582, 40.800037], [-73.699793, 40.800037], [-73.699793, 40.541722]]"
21	2021-11-30-19-52,is the biggest racist on Twitter. How tf is he still allowed on this platform??? @TwitterSupport,"[[-74.010799, 40.9354], [-74.010799, 40.955124], [-73.975
22	2021-11-30-19-52,Guess that friend https://t.co/cNp3rpP263,"[[-74.026675, 40.683935], [-74.026675, 40.877483], [-73.910408, 40.877483], [-73.910408, 40.683935]]"
23	2021-11-30-19-52,En eff wha ??,"[[-74.026675, 40.683935], [-74.026675, 40.877483], [-73.910408, 40.877483], [-73.910408, 40.683935]]"
24	2021-11-30-19-52,His name is Santiago Rodriguez!!! #NYCFC,"[[-74.026675, 40.683935], [-74.026675, 40.877483], [-73.910408, 40.877483], [-73.910408, 40.683935]]"
25	2021-11-30-19-52,ok. I wonâ€™t might going to Camden for a game or 2 now.,"[[-73.473176, 40.734618], [-73.473176, 40.782376], [-73.435171, 40.782376], [-73.435171, 4
26	2021-11-30-19-52,"Shot by @candicekellyxo, wearing @hanifaofficial â€" November, 2020 Walking the #HanifaDreamInDC runway â€" November,â€¦ https://t.co/GaP3XH
27	2021-11-30-19-53,Wale x The Breakup Song,"[[-79.76259, 40.477383], [-79.76259, 45.015851], [-71.777492, 45.015851], [-71.777492, 40.477383]]"
28	2021-11-30-19-53,"I have three of these on Criterion! I don't have New World. Tree of Lifs to me is perfect, I still haven't watchedâ€¦ https://t.co/PYlbPBn7A4","[[-74.0266

Figure 2: Table with tweet information

At the same time the Twitter data is being acquired and processed, weather information is also being streamed through the OpenWeather API. Two different parameters are collected to build the regression model: temperature in Celsius and weather status, the ladder being a description of the weather condition, e.g., clear, clouds, rain, etc. Those two parameters were selected because we made an assumption a priori that they would be the most relevant to produce results. The weather data is collected every minute to account for even slight changes in temperature, thus ensuring that any given temperature will have enough tweets associated with it to build a meaningful model. We also collect information about the exact time the data was collected, as it will be used to link it with the twitter data, which is being collected in parallel. Figure 3 showcases a table with weather information collected for 13 minutes in one of the runs.

1	Time,Status,Temperature
2	2021-11-30-19-51,Clouds,3.91
3	2021-11-30-19-52,Clouds,3.91
4	2021-11-30-19-53,Clouds,3.91
5	2021-11-30-19-54,Clouds,3.91
6	2021-11-30-19-55,Clouds,3.91
7	2021-11-30-19-56,Clouds,3.91
8	2021-11-30-19-57,Clouds,3.91
9	2021-11-30-19-58,Clouds,3.91
10	2021-11-30-19-59,Clouds,3.91
11	2021-11-30-20-0,Clouds,3.91
12	2021-11-30-20-1,Clouds,3.9
13	2021-11-30-20-2,Clouds,3.9
14	2021-11-30-20-3,Clouds,3.9

Figure 3: Table with weather information

The next step is to combine the two types of data collected and classify the tweets based on their expressed sentiment. As the data was streamed and processed it was also stored in csv files, thus we can easily access it. For each tweet collected the time of creation is checked. That time is then searched for in the weather information, and when the corresponding match is found, the weather data from that time is assigned to the tweet. To classify the sentiment of the tweet two modules are used: textblob and vader. They are both modules which can take a string of text and assign a value to it based on calculated sentiment. We use a binary classification instead, classifying tweets as either positive or negative, and we make use of both modules so that only tweets classified the same way for both are taken into consideration, the rest are deleted. This is done to increase the accuracy of the prediction model, as some tweets can be incorrectly classified. The resulting table at the end of this step, shown below, contains all the information collected about tweets, weather, and the sentiment classification.

	A	B	C	D	E	F
1	Date,Text,Box,Status,Temperature,Sentiment					
2	2021-11-27-15-20,I would never understand the logic behind sitting righ					
3	2021-11-27-15-20,thanks playing Music ðŸŽ¶ By @Madonna this aftern					
4	2021-11-27-15-20,It's always gonna be a debate on everything because					
5	2021-11-27-15-20,She looks pretty Period,"[[-73.560512, 40.743142], [					
6	2021-11-2 some great peopleâ€¦ https://t.co/3fpG1Zx8NV","[[-74.02					
7	2021-11-27-15-20,The secondary owes the Dline dinner Good lord,"[[-					
8	2021-11-27-15-20,"25 looks good, wondering what 26 has in store for					
9	2021-11-27-15-21,"Can you recommend anyone for this Security Offic					
10	2021-11-27-15-21,Been taking the train since Iâ€™ve been 12ish. Ngl t					
11	2021-11-27-15-21,I could do more than handle that sexy,"[[-74.02667					
12	2021-11-27-15-21,How can you not love college football?,"[[-73.9625					
13	2021-11-27-15-21,Good Morning ! @SushovanBaner15 @narendramo					
14	2021-11-27-15-21,"Rep. Eric Swalwell tells Joy Reid, â€˜Here we saw D					
15	2021-11-27-15-22,YALL GOTTA DO BETTER COACHES! DONT UNDERST					
16	2021-11-27-15-22,WoW. Their coffee is too strong for me. My stomac					
17	2021-11-27-15-22,Yâ€™all that Italian man wasnâ€™t real!!!!!! Hahah					
18	2021-11-27-15-22,Thanks!! ðŸ’–ðŸ’–ðŸ’–,"[[-74.026675, 40.683935], [-					
19	2021-11-27-15-22,I am far from rich but I know that I am extremely pri					
20	2021-11-27-15-22,"This job is now open at Booz Allen in New York, NY					
21	2021-11-27-15-22,Your supposed apology was insincere and you shoul					
22	2021-11-27-15-22,U work hard,"[[-74.026675, 40.683935], [-74.02667					

Figure 4: Table with tweet and weather information

Finally we can use the table made in the previous step to build and test various regression models. The models will take as training input the weather status and temperature. The training output is a ratio of positive tweets to total tweets made under those weather conditions, and our goal is to be able to find a meaningful relationship between the weather and Twitter public sentiment. Different regression models will be built and tested to find the one which can most accurately predict that relationship based on real world testing. The models are: linear regression, ridge regression with different parameters choices, and SVR. With the data we have collected so far we have already built those regression models, however we have found that the data is insufficient as of now, as the weather conditions have not varied significantly enough over the course of the training to have a good model. Below is an example of a prediction of positive tweets to total tweets with an SVR model.

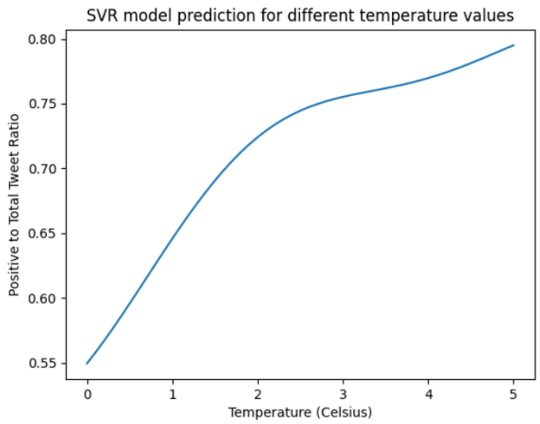

Figure 5: Twitter positivity ratio by temperature

4. As mentioned above, the model had insufficient data to work with, as temperatures below a certain threshold have not been collected yet. This effect shows in the graph by a huge drop in positivity ratio below the temperature of 2 Celsius, which is unlikely to be accurate.

It is important to note that the graph showcased in Figure 5 shows the sentiment prediction for temperature under the weather condition of clear skies. Other weather condition, such as snow, rain, or cloudy, are likely to result in different graphs, as they may greatly interfere with public sentiment. However, for display purposes in this paper only the aforementioned condition is shows. On the webpage users will be able to select which condition they choose to view, so that they will have access to the entirety of the information from the regression model.

As for the front end, so far we were able to build a dashboard using Django and display the weather csv file as a table. The preview is as follows:

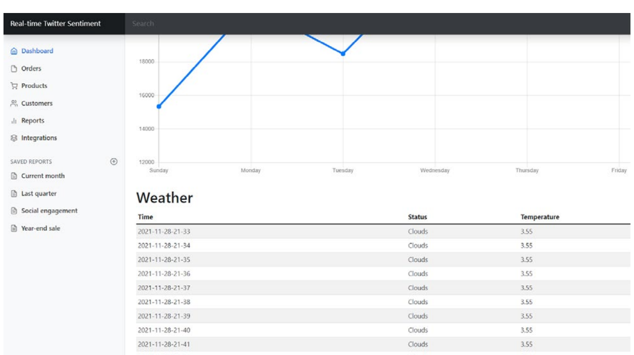

Figure 6: Snapshot of the dashboard

In the future we plan on showcasing the regression models used and their predictions for positivity ratio along with their accuracies in real time. This way expected public sentiment can be shown for the time of viewing. The dashboard will showcase graphs outlining the results for each model used, taking into account the current weather status. The user will be able to change the condition of the status manually, so that they can see how different weather conditions affect the expected sentiment withouth having to wait for those conditions to happen in real time.

5. Planned experiments and figures

The planned experiments of our project can be broken down into the following parts:

Firstly, to mitigate the issue that our current model is lacking data, we will continue running the DAG to collect more tweets using Google Cloud Computing. After we have collected enough tweets, we will conduct the experiment by training different regression models on the text corpus. Because of the high volume of the database, the models will be robust.

At the same time as data is being collected, we will run manual experiments on the sentiment classification and determine if the current classifiers are correctly modeling the tweets as positive or negative. We will look at 1000 tweets in total and manually assign their sentiment to then compare against the results given by the classifier. We can then use the resulting accuracy to determine whether we're content with the classifier or if it should be further improved upon by use of more robust techniques.

Next, we will evaluate our trained regression models by comparing the prediction results between different models. Specifically, each trained model will predict a sentiment positivity ratio based on the current weather. We can stream real-time weather data and twitter data, and compare the predicted results with actual sentiment. A closer match will be an indication of good modeling. Based on the evaluation, we will choose the best model and improve upon our implementation.

Lastly, to better visualize our work, we will continue polishing our front end webpage. On the dashboard, we will stream weather data and twitter data for each predefined time interval, which will be displayed against the predictions made from the models. Users entering our page can easily access the current weather condition and our real-time predictions of tweets sentiment.

Some possible tables and figures to be included in our report are:

1. Tables showing the combined database of tweets and weather conditions.
2. Figures of sentiment predictions made by different classifiers after enough data is gathered.
3. A snapshot of our dashboard showcasing the ease of user interactions and a real-time display of prediction results.
4. A comprehensive evaluation metrics including

statistics and graphs.

6. References

[1] Baylis, P., Obradovich, N., Kryvasheyeu, Y., Chen, H., Coviello, L., Moro, E., Cebrian, M., & Fowler, J. H. (2018). Weather impacts expressed sentiment. PLOS ONE, 13(4). https://doi.org/10.1371/journal.pone.0195750

[2] Tweetin' in the Rain: Exploring Societal-scale Effects of Weather on Mood - ccs.neu.edu. (n.d.). Retrieved November 19, 2021, from https://www.ccs.neu.edu/~amislove/publications/Weather-ICWSM.pdf.

[3] Twitter sentiment analysis introduction - stanford university. (n.d.). Retrieved November 19, 2021, from https://nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf

[4] Baylis, Patrick, et al. "Weather Impacts Expressed Sentiment." PLOS ONE, vol. 13, no. 4, 2018, https://doi.org/10.1371/journal.pone.0195750.