Forecasting the Future using Diverse Social Media Sources

Social media signals have been effectively used to predict real-world events, identify public opinions, and forecast influenza dynamics. However, the majority of existing social media analytics focuses on predictive and not forecasting capabilities; relies on one specific type of signal e.g., text; and is primarily executed on English text streams. We propose to evaluate the predictive power of mixed social media signals e.g., images and text jointly in combination with recently emerged deep learning models for forecasting the future weather.

Our experiments rely on 170M tweets and 862,937 images collected for 25 geo-locations in the U.S. and 6 international locations as described in (Volkova et al., 2017), global summary of the day(GSOD) gathered from NOAA.gov[1], and Flickr weather images[2]. Two main tasks are used to determine the predictability of weather based on the five GSOD values gathered: a regression task that consists of forecasting the next day dew point temperature in Celsius, air temperature in Celsius, wind speed in Knots, and visibility in miles, and a classification task that consists of predicting if there will be precipitation. We evaluate the model's generalizability across these tasks and 31 geo-locations.

Our approach uses multiple data streams of images and text, deep learning architectures, and multilingual analysis to predict each weather target. The models developed rely on the joint image and text signals to learn the dynamically changing discourse and graphical content in social media by using the past 12 days of data.

The five main models were developed by using (1) previous Ys, (2) tweet text only, (3) image content only, (4) a merged model of text and images and (5) an ensemble approach of text and images to predict future Ys. For each of these experiments, we applied the state-of-the-art machine learning model for timeseries prediction – AdaBoost (Santillana et al., 2015), and recurrent neural network architecture that relies on Long Short-Term Memory (LSTM) implemented in scikit-learn and keras, respectively. The text and image models contained two LSTM layers and a dense layer. The merged model combined the text and image LSTM layers through a merge layer then one dense layer. The ensemble method had the text and image models predict their values and take the average.

Flickr data was used to develop a new classification model to classify images as cloud, storm, snow and clear weather. We used ResNet's ImageNet weights and added extra dense layers to then train the model to predict the four categories. Once trained, we used the model to extract weather-related feature vectors for the Twitter images that were used for the weather prediction.

We found that performance varies across all geo-locations and AdaBoost outperforms LSTM. For max daily temperature forecast, the LSTM model resulted with a Pearson range from .23 to .86 across geo-locations, R^2 range from -.002 to .50, and RMSE ranged from 1.1 to 9.7. The AdaBoost model resulted with a Pearson range from .42 to .93, R^2 range from -.15 to .83, and RMSE ranged from .88 to 9.4. The LSTM previous Y model had upper bounds of .95 of Pearson, .86 of R^2, and a RMSE of .88. For many locations, the text only model outperformed images and image merged model. For example, location L14 with text only outperformed images only and images merged models by .3 and .06 in Pearson, respectively. However, the images ensemble model had .04 less than text only. We also discovered that it was easier to predict air temperature and wind speed which had an average root mean squared error of 5.811 and 5.87 compared to dew point temperature and visible miles which had an average error of 376 and 77.

Future work includes building unified models capable of forecasting the weather from multilingual social media streams by relying on character and byte embeddings, combined with image representations. In addition, we will rely on Google's Tensorflow object detection tool[3] to extract objects from Twitter images and learn joint representations between tweet text and extracted objects.

1. Volkova, Svitlana, et al. "Uncovering the relationships between military community health and affects expressed in social media." *EPJ Data Science* 6.1 (2017): 9.
2. Santillana, Mauricio, et al. "Combining search, social media, and traditional data sources to improve influenza surveillance." PLoS computational biology 11.10 (2015): e1004513.

[1] NOAA GSOD: https://data.noaa.gov/dataset/global-surface-summary-of-the-day-gsod

[2] Project Weather Flickr: https://www.flickr.com/groups/projectweather/

[3] Google's Tensorflow object detection tool: https://github.com/tensorflow/models/tree/master/object_detection