# Predicting Sentiment on Climate Change Across the Globe using Deep Learning

Sarah Alamdari, Wesley Beckner, Neal Dawson-Elli, Yusong Liu
Chemical Engineering, University of Washington

## NLP

Natural Language Processing, or NLP, is a branch of Machine Learning focused on creating models that process and interpret human language. NLP and social media allow systematic analysis of large populations. One example of a trending NLP application is in virtual assistants, which must divine intent from colloquial or shorthand spoken language.

Other applications include:

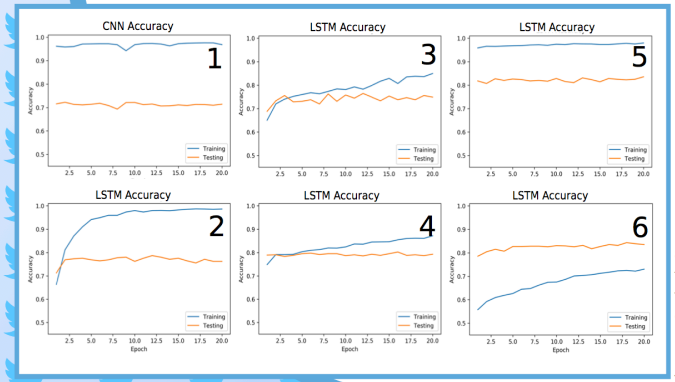
- Text Classification
- Semantic Parsing
- Question Answering
- Text to Speech Conversion

## Model Development

**1 Text Representation.** In separate trials, Google's pre-trained Word2Vec embeddings, non pre-trained embeddings (i.e. these were trained on our twitter corpus), and bag of words methods, were used to represent words. The subsequent, vectorized tweets were then fed into either a convolutional or recurrent neural network. Consistently, google's Word2Vec pre-trained library was a top performer. The data shown here are confined to the Word2Vec results.

| | CNN | BD-LSTM W2V Embedding | | | | |
|---|---|---|---|---|---|---|
| | Don't Balance Don't Train Embedding | Balance Train Embedding | Balance Don't Train Embedding | Don't Balance Don't Train Embedding | Class Weighting Don't Train Embedding | Class Weighting Don't Train Embedding Keep Nan |
| Accuracy | 0.71 | 0.76 | 0.75 | 0.79 | 0.84 | 0.84 (w/o Nans) |
| Precision | 0.62 | 0.80 | 0.79 | 0.85 | 0.88 | 0.87 |
| Recall | 0.67 | 0.73 | 0.71 | 0.88 | 0.91 | 0.92 |
| Accuracy (Training) | 0.97 | 0.99 | 0.85 | 0.88 | 0.98 | 0.73 (w/ Nans) |
| Image ID | 01 | 02 | 03 | 04 | 05 | 06 |

**2 Model Architectures.** Convolutional and recurrent neural networks (CNN and RNN), two related methods that have the ability to spatially/temporally contextualize data, were used to create the final models. The CNN, with its fixed filter size, underperformed against the bi-directional *long-short term memory* (BD-LSTM) RNN. In addition to having the ability to decide when to forget or recall certain features, the *bi-directional* LSTM learned from inverse-fed tweets, allowing it to learn certain word ordering redundancy that occurs in English (e.g. Jane ran to the barn is the same as to the barn, Jane ran).




**3 Data Preprocessing.** Apart from text representation, two additional data preprocessing tricks were used: balancing and class weighting. Despite our unbalanced classes (2:1), balancing the dataset severely data-starved the model, and resulted in poor performance. Class weighting resulted in higher performance in all three key metrics (accuracy, precision, and recall).

**4 Inclusion of Nans.** In a final attempt to better handle the nature of our live twitter data, where certain tweets wouldn't have anything to do with climate sentiment, we provided our model with 'Nan' labeled tweets. In this final scheme, we didn't one-hot encode the 'Nan' labels, but rather incorporated them into the existing one-hot labeling scheme with the 'Yes' and 'No' labels by assigning them targets of [0, 0]. The idea here is that by denying them a unique class, we still expose 'Nan' tweets to our model but without the loss in yes/no accuracy associated with adding a third class. This becomes clear when observing the difference between testing and training accuracy ( 'Nan' was removed from the test set)

## Training Data

Data was taken from figure eight. Tweets were hand labeled and evaluated for belief in the existence of global warming or climate change. Possible answers were:

- "Yes", suggesting climate change is occuring
- "No", suggeting it is not occuring
- "I can't tell", if the tweet is ambiguous or unrelated

| tweet | existence | existence.confidence |
|---|---|---|
| Global warming report urges governments to act... | Yes | 1.0000 |
| Fighting poverty and global warming in Africa ... | Yes | 1.0000 |
| Carbon offsets: How a Vatican forest failed to... | Yes | 0.8786 |
| Carbon offsets: How a Vatican forest failed to... | Yes | 1.0000 |
| URUGUAY: Tools Needed for Those Most Vulnerabl... | Yes | 0.8087 |

### Challenges

NLP problems can be challenging because it is difficult to identify and interpret patterns in language. Tweets specifically can be very ambiguous. With the training dataset, the following features were found to be limiting in creating a comprehensive model.

- Small dataset (~6,000 labeled tweets)
- A quarter of the dataset was comprised of ambiguous tweets
- Tweets do not contain much context (limited to 250char)
- Many locations are self reported


With this in mind, we set out to tackle these problems by including features like pre-trained libraries, and special considerations in our models.

## Twitter API

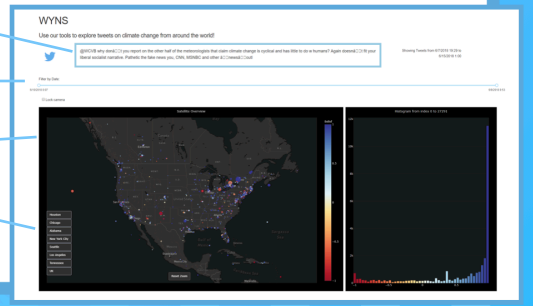
Mined tweets met the following constraints:

- Shared location data (self-reported OR with location services turned on)
- Contained the phrase "climate change" OR "global warming"

Live tweets were mined using Tweepy, a python wrapper for the twitter API

## User Interface

A user interface (UI) was built using Dash by Plotly, a python-based framework for building web-applications. The app is currently hosted on Google Cloud. To highlight some features, the UI includes:




- Tweets displayed by location
- A timeline slider
- Interactive histogram displaying breakdown of tweets
- Pop up of the tweet when hovered

## Acknowledgements

We would like to thank our Capstone Sponsors, KPMG for their mentorship throughout this project. Thanks to Dave Beck for his valuble input throughout the course, and Kelly Thorton for organizing the DIRECT events