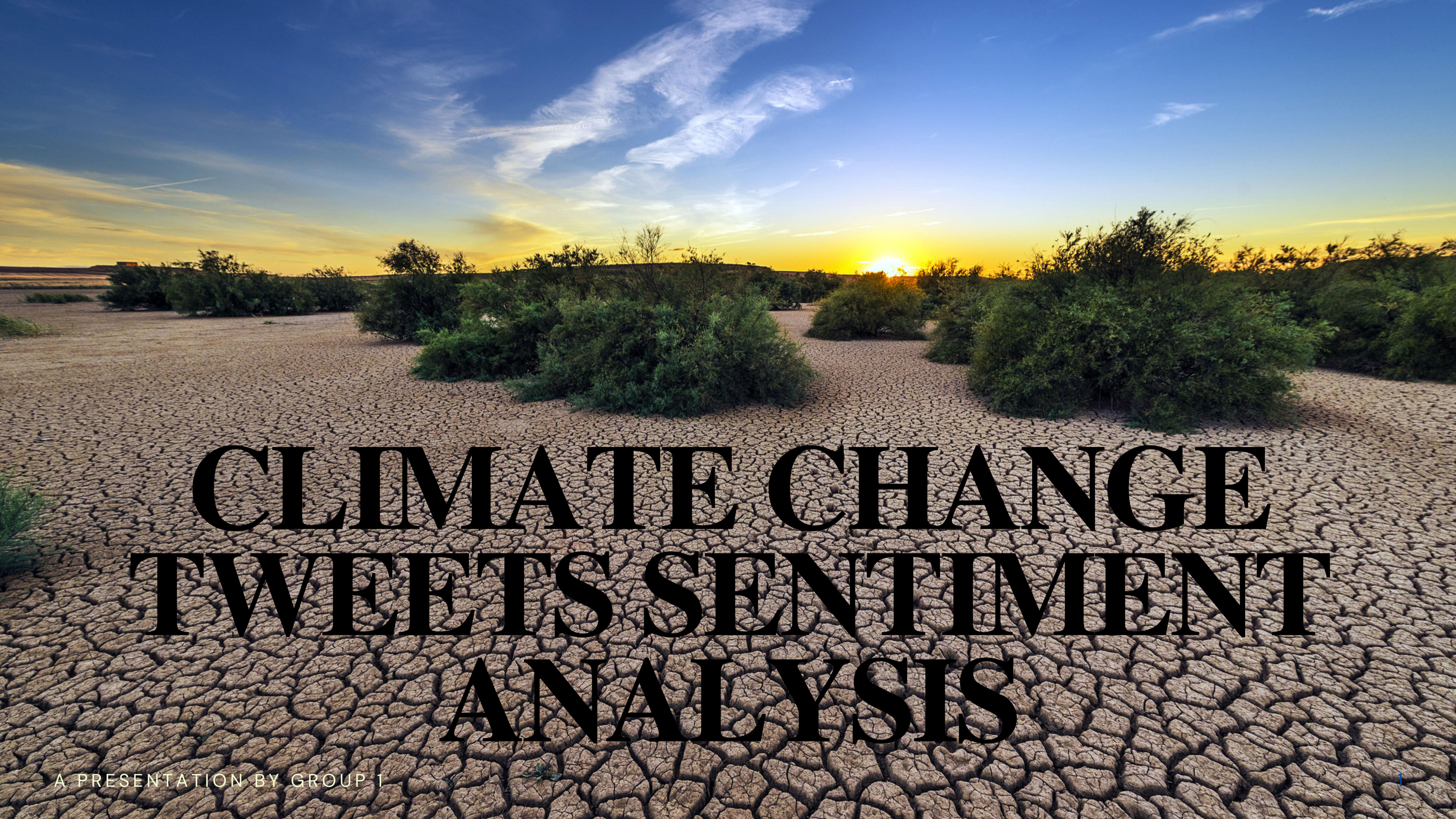


CLIMATE CHANGE TWEETS SENTIMENT ANALYSIS

A PRESENTATION BY GROUP 1

Definition

CLIMATE CHANGE:

Climate change refers to long-term shifts in temperatures and weather patterns. These shifts may be natural, such as through variations in the solar cycle. But since the 1800s, human activities have been the main driver of climate change, primarily due to burning fossil fuels like coal, oil and gas.

Problem statement

- Twitter being a social media app where people can freely air their opinions and drive social trends plays a key role in climate change conversations. Using hashtags(#), pro-climate change organizations have inspired the climate change talks to the public while concurrently studying the tweets and responses and consequently driving the environmental conservation and climate change awareness to the masses. Twitter therefore qualifies as a key driver of the climate change circa.
- In this project the aim is to build a model to classify tweets as either tweets that do not believe in man-made climate change or neutral tweets on climate change or tweets that support the belief in climate change or tweets that are linked to factual news about climate change.

Main objective

Build a climate change tweets classifier using avarious machine learning algorithms and decide the best method based on various success metrics.

Applications

- Can be used to detect and flag tweets that propagate false information about climate change.
- Can be used as an indicator to show level of public knowledge on climate change and this can show myths that need to be debunked and this can be done on the same social media platforms through targeted ads etc.

NLP

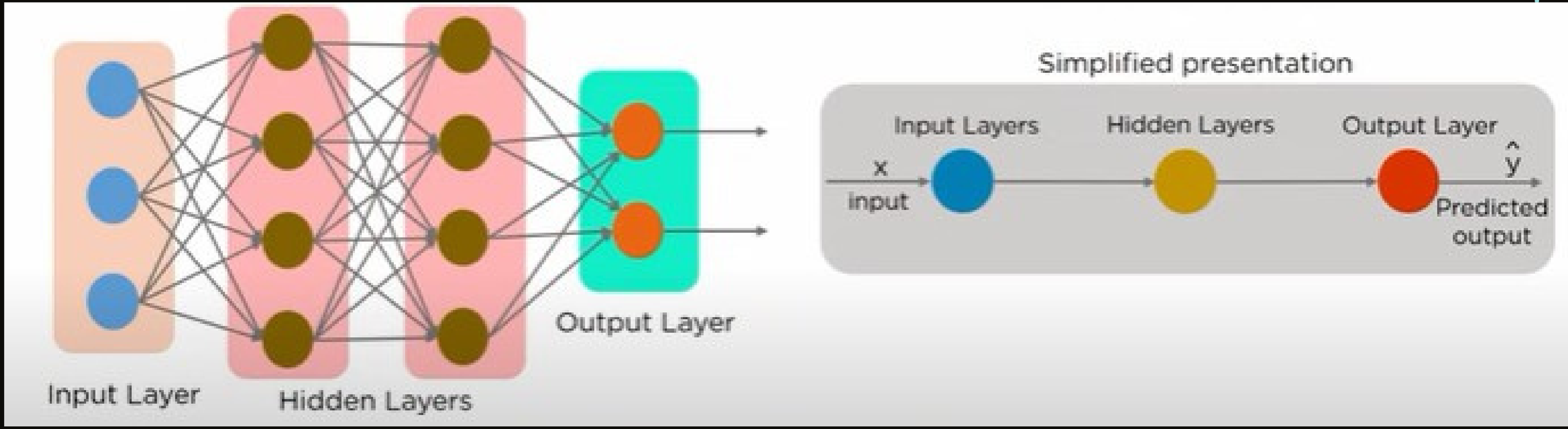
- Branch of Artificial Intelligence that gives machines the ability to understand text and spoken word and derive meaning like humans do, combines field of linguistics and computer science. NLP has many applications like autocorrect, autofill etc. The aspect of NLP we will be looking at is sentiment analysis specifically using twitter data.

- Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service, or idea.

- Various machine learning algorithms such as Support Vector Machine (SVM), Recurrent Neural Network (RNN), Random Forest, Naïve Bayes, and Long Short-Term Memory (LSTM) and many more can be used for sentiment analysis.

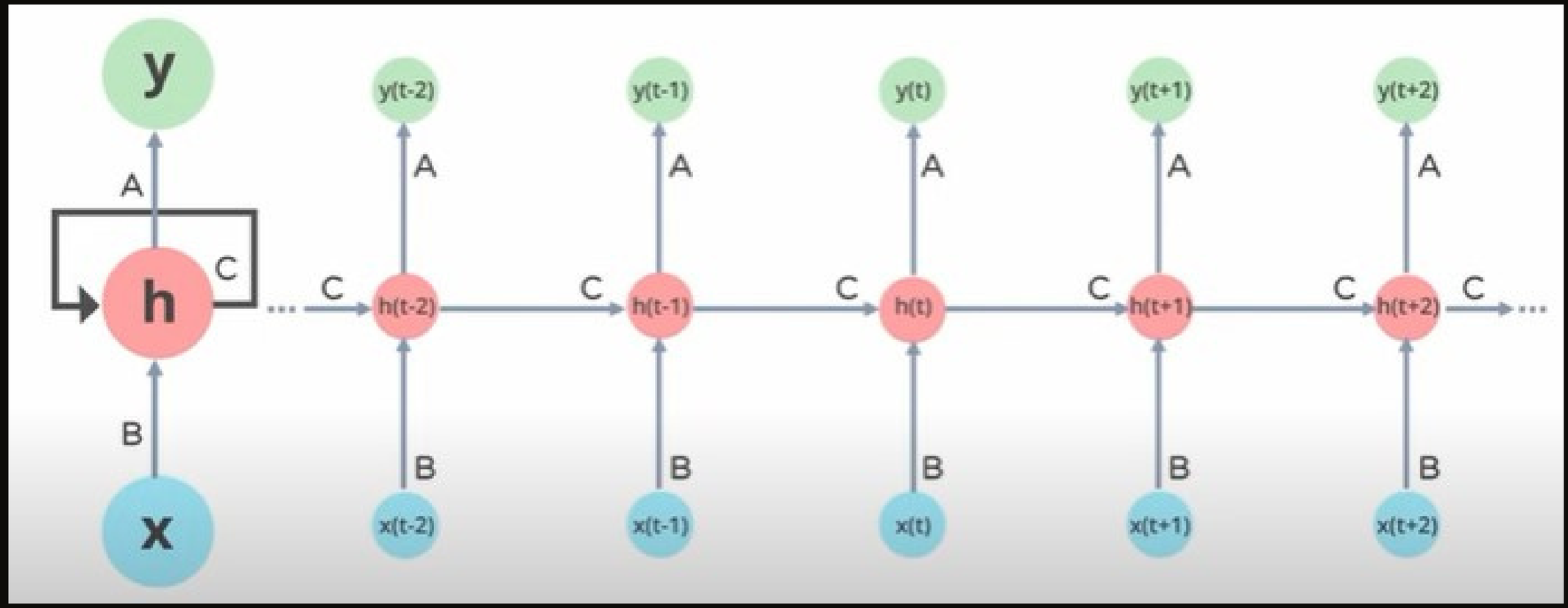
RNN

- The simplest artificial neural network which is the feed forward neural network has information flowing only in the forward direction. Decisions are made based on current input and there is no memory of the past hence the need for RNN.

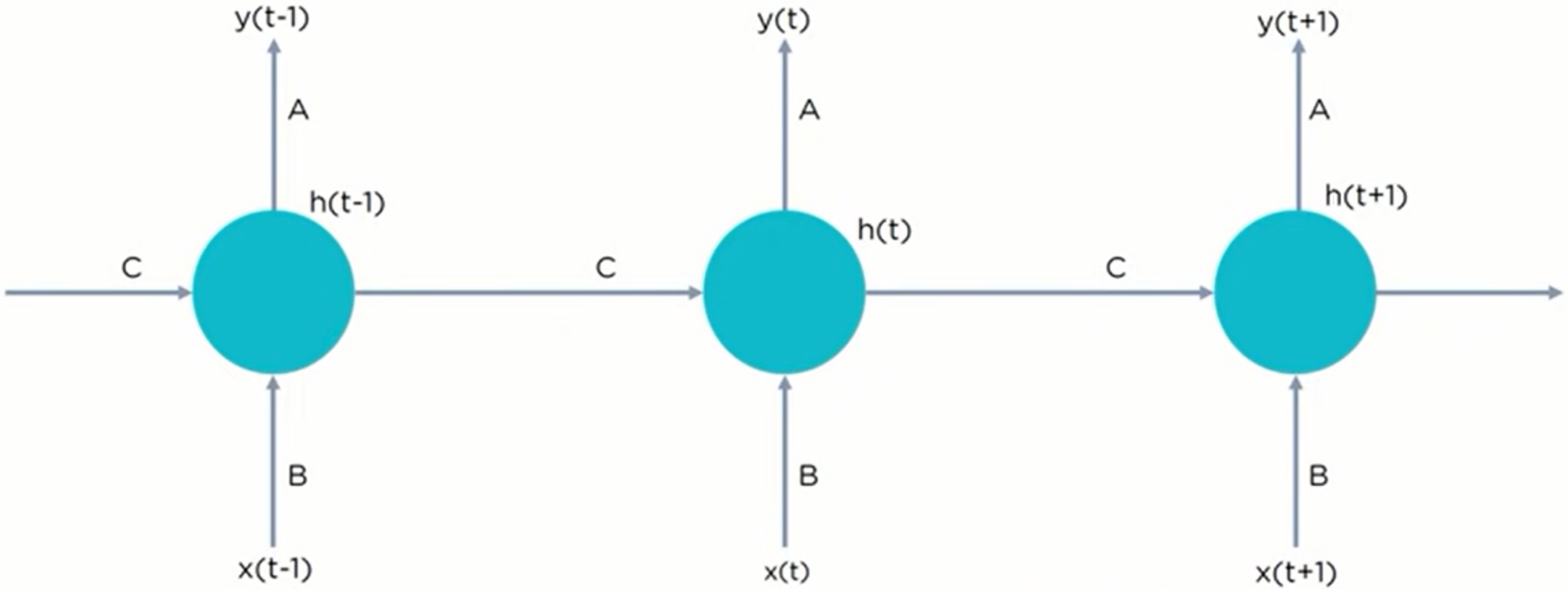


How RNN works

- RNN works by the principle of saving the output of a layer and feeding it back into the output in order to predict the output of the layer.



How RNN Works



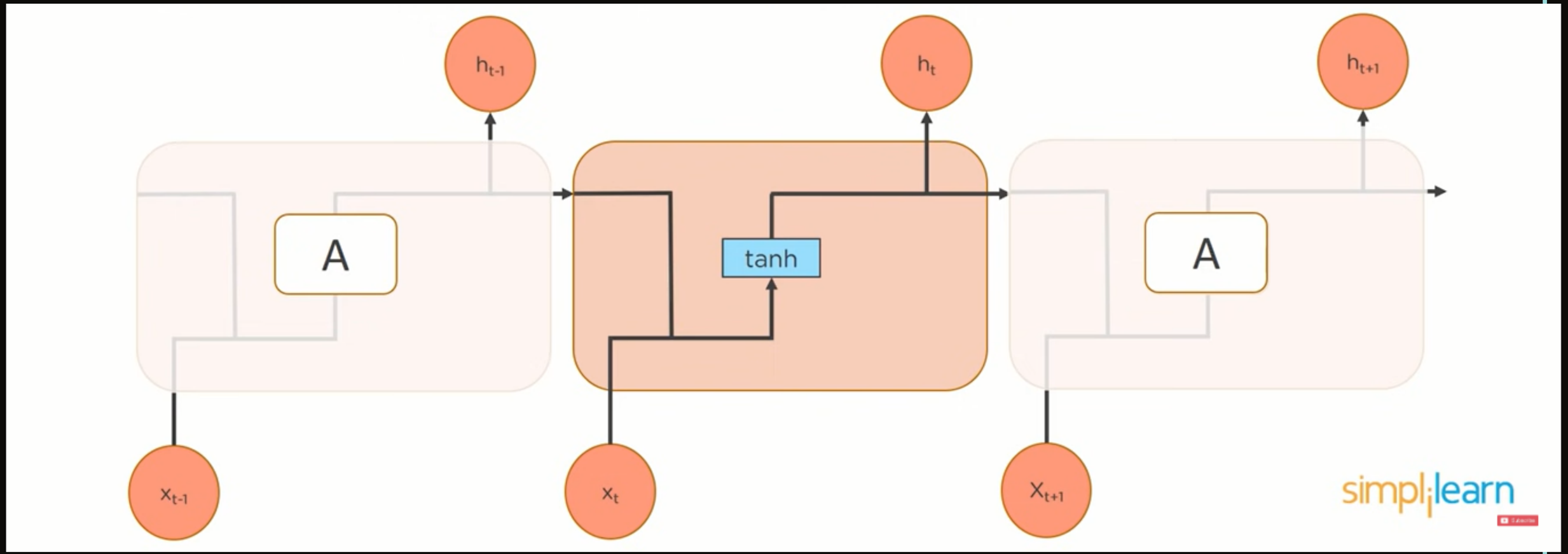
$h(t) = f_c(h(t-1), x(t))$

h(t) = new state
f_c = function with parameter c
h(t-1) = old state
x(t) = input vector at time step t

simplilearn

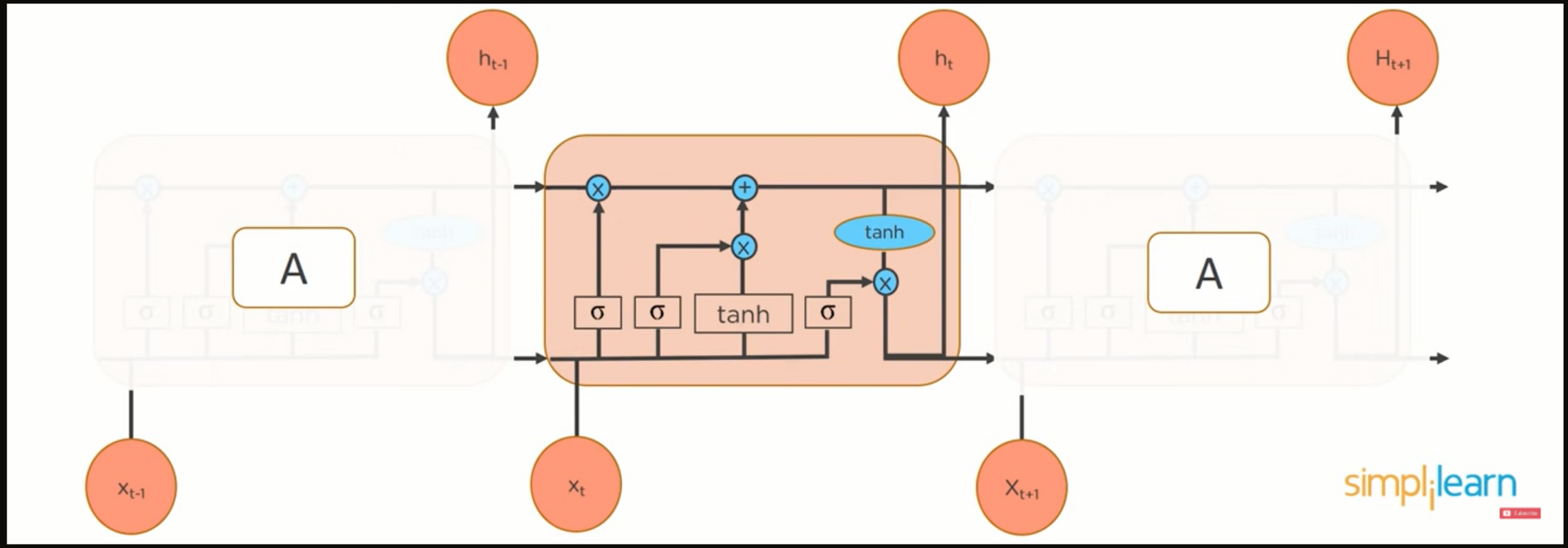
How RNN Works

- RNN structue is a chain of repeating modules of Neural networks and in an RNN this will have a very simple structure such as a single tanh layer.



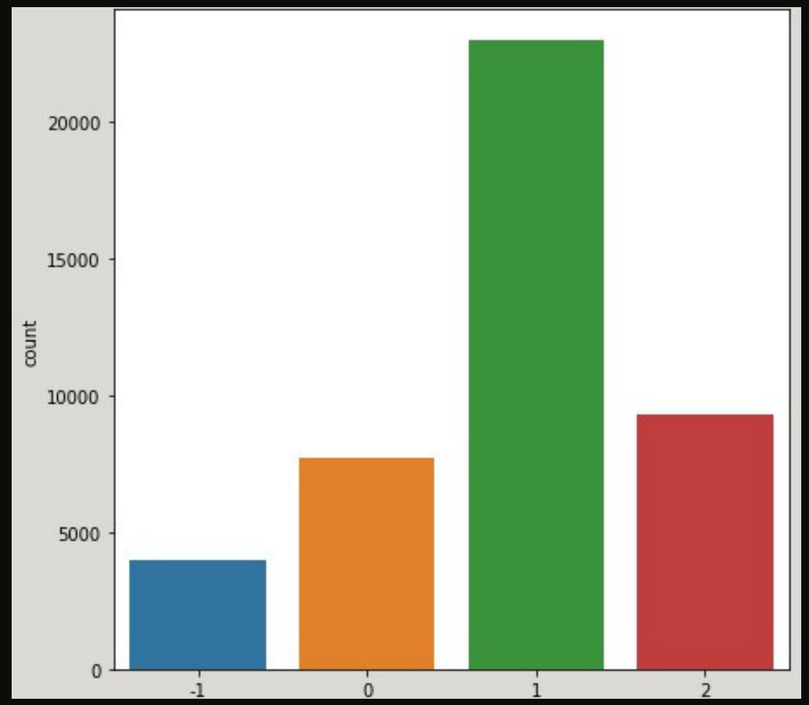
LSTM

- LSTMs are a type of RNN that are supposed to deal with issues of vanishing and exploding gradient experienced in RNN. LSTMs have a structure that can maintain information in memory for longer which has 4 interacting layers.

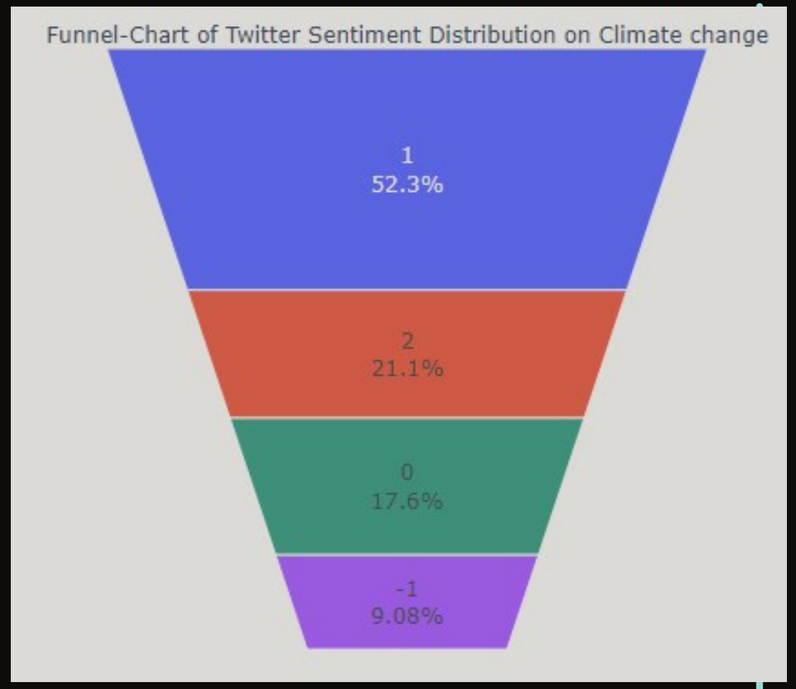


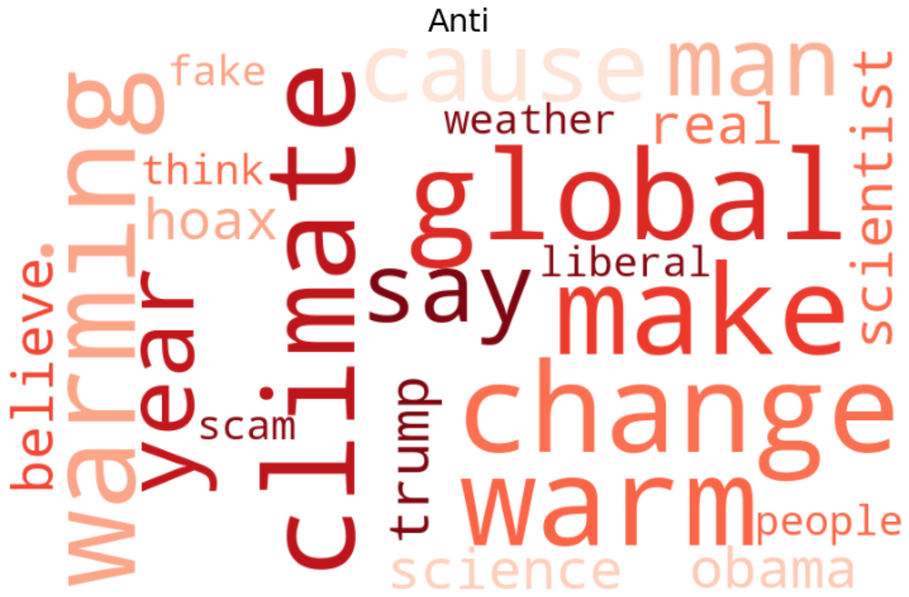
The data

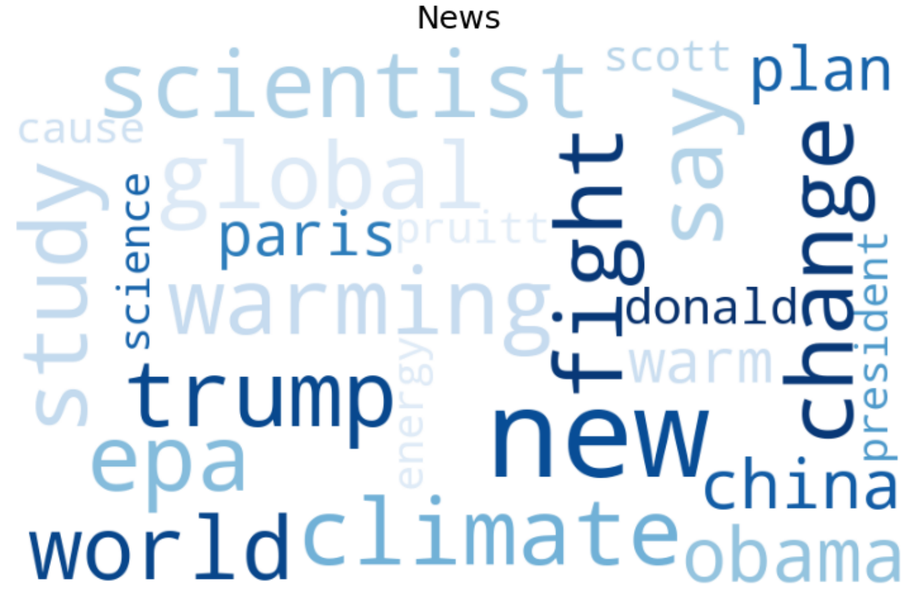
- The data available was sentiments from twitter divided into 4 categories: -1 for tweets that do not believe in man-made climate change, 0 for neutral tweets, 1 for tweets that support the belief in climate change and 2 for tweets that are linked to factual news about climate change. From the EDA majority of the tweets were those that support the belief in climate change then those from factual news outlets then the neutral and lastly those that do not believe in man-made climate

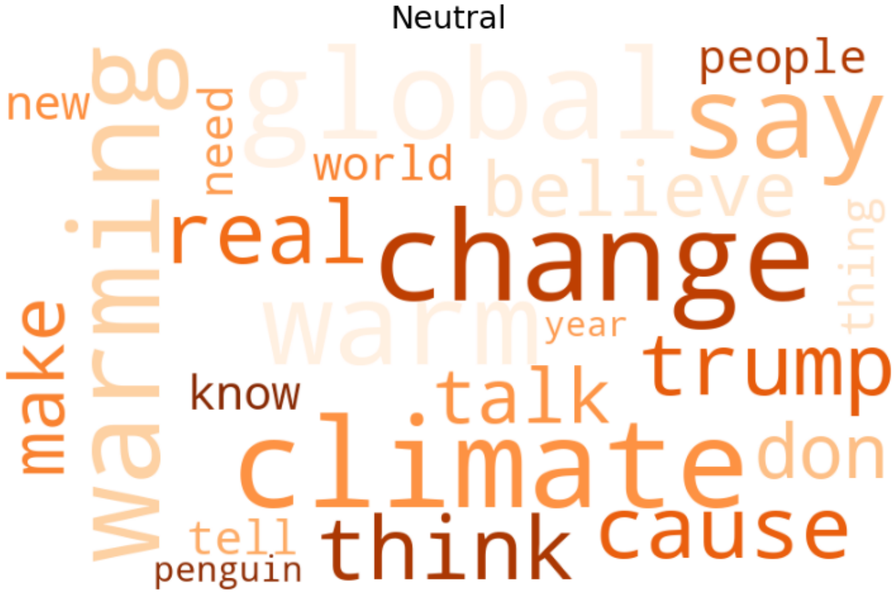


	sentiment	message
2	1	22962
3	2	9276
1	0	7715
0	-1	3990









Word clouds were used to identify some of the major buzzword in the data. To expound further, it is visible that tweets which are pro climate change have words such as change, need, believe, global, people. The anti sentiments has words such as say, fake, think, scam, hoax while the news have scientists, epa, world, climate among others.

Data cleaning and pre-processing

- Data was cleaned by:

1. Importing the libraries and loading the data.
2. Checking features of the data such as the shape.
3. Checking for missing values which the data did not have.
4. Dropping the tweet id column
5. Checing for duplicates which were 2902 and dropping them.

- Data preprocessing was done by creating a function that was for:

1. Lowering the case.
2. Removing urls.
3. Removing RT
4. Removing @
5. Removing punctuations

- The function was then applied to the data then the data was split.

Naive Bayes, decision trees and K-Nearest Neighbours models

- 4 models were employed to the data which are: Decision trees, Naïve Bayes, LSTM and K-Nearest Neighbours.
- For Naïve Bayes, decision tree and K-Nearest Neighbours after splitting the data count vectorizer is implemented with a parameter to remove stop words then it is fitted on the train data then both the train and test data are transformed to document-term matrix. In short they are being transformed to tokens(words) then a matrix of integers. The word data has been transformed to a language the model can understand.
- After this step both the Naive Bayes, decision tree and K-Nearest forest models are built.
- The Naive Bayes has an accuracy of 67% , the K-Nearest Forest has an accuracy of 44% and the decision trees has an accuracy of 57%
- The decision tree model did perform better than KNN but still didn't reach the Naive Bayes score.

LSTM model

- For LSTM after the data is split it needs to be tokenized and padded. The summary of the process is converting the strings into tokens then tokens into array of integers representing the individual words or tokens/sequences. These sequences of both the train and test data need to be the same length so they are padded which is adding zeros until they are the same length.
- The LSTM model is then been built and fitted on the data and it has an accuracy of 87%.
- Hence the LSTM model will be used to classify the climate change tweets.

THANKYOU