**CLIMATE CHANGE TWEET CLASSIFICATION**

**USING TWITTER DATA**

**BY**

**CAMERON PIKE**

**1.INTRODUCTION**

This presentation will provide insight into the factors and opinions surrounding climate change .

I created a Machine Learning model that is able to classify whether or not a person believes in climate change, based on their tweet data.

The evaluation metric will be the [Mean F1-Score](https://en.wikipedia.org/wiki/F-score). The F1 score, uses the outcome of precision and recall.

Precision is the ratio of true positives to all predicted positives. Recall is the ratio of true positives to all actual positives.

## **Dataset Description (provided)**

**The files provided**

* train.csv
* test.csv
* SampleSubmission.csv

**Features of train.csv**

* sentiment: Which class a tweet belongs in (refer to Class Description above)
* message: Tweet body
* tweetid: Twitter unique id

**Class Description**

* 2 - News: the tweet links to factual news about climate change
* 1 - Pro: the tweet supports the belief of man-made climate change
* 0 - Neutral: the tweet neither supports nor refutes the belief of man-made climate change
* -1 - Anti: the tweet does not believe in man-made climate change Variable definitions

**2.Approach**

**2.1 Data Capture**

I imported the training, test and sample files into dataframes. A dataframe is similar to an excel spreadsheet

**2.2 Exploratory Data Analysis**

I viewed the first five rows of the data files to get a feel of the data. I dropped the last column of the sample data as I would replace this with my predicted data.

Below is the training data

sentiment message tweetid

0 1 PolySciMajor EPA chief doesn't think carbon di... 625221

1 1 It's not like we lack evidence of anthropogeni... 126103

2 2 RT @RawStory: Researchers say we have three ye... 698562

3 1 #TodayinMaker# WIRED : 2016 was a pivotal year... 573736

4 1 RT @SoyNovioDeTodas: It's 2016, and a racist, ... 466954

Below is the test data

message tweetid

0 Europe will now be looking to China to make su... 169760

1 Combine this with the polling of staffers re c... 35326

2 The scary, unimpeachable evidence that climate... 224985

3 @Karoli @morgfair @OsborneInk @dailykos \nPuti... 476263

4 RT @FakeWillMoore: 'Female orgasms cause globa... 872928

I then viewed the columns, shape of the dataframe and the different data types in the data

These are the columns in the training data

Data columns (total 3 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 sentiment 15819 non-null int64

1 message 15819 non-null object

2 tweetid 15819 non-null int64

Count of rows in the data is: 15819

Unique sentiment values in train

[ 1 2 0 -1]

These are the columns in the test data – to note there is no sentiment column

Data columns (total 2 columns):

# Column Non-Null Count Dtype

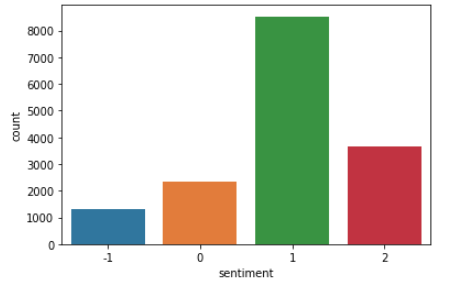
--- ------ -------------- -----

0 message 10546 non-null object

1 tweetid 10546 non-null int64

Count of rows in the data is: 10546

**Below is the count of sentiment values:**



**3.Clean the data**

3.1 Converted the data to lower case

3.1 Removed punctuation eg ?!

3.2 Removed repeating characters

3.3 Removed URL’s

3.4 Removed repeating numbers

3.5 Removed hashtags

3.6 Removed mentions

Below is an output after the cleaning:

after clean numbers test data

10541 rt brittanybohrer brb writing a poem about cli...

10542 the year climate change came home during the ...

10543 rt loopvanuatu pacific countries positive abou...

10544 rt xanria you’re so hot you must be the cause ...

10545 rt chloebalaoing climate change is a global is...

Name: message, dtype: object

after clean hastags

15814 rt ezlusztig they took down the material on gl...

15815 rt washingtonpost how climate change could be ...

15816 notiven rt nytimesworld what does trump actual...

15817 rt sarasmiles hey liberals the climate change ...

15818 rt chetcannon kurteichenwalds climate change e...

Name: message, dtype: object

after clean hastags test data

10541 rt brittanybohrer brb writing a poem about cli...

10542 the year climate change came home during the ...

10543 rt loopvanuatu pacific countries positive abou...

10544 rt xanria you’re so hot you must be the cause ...

10545 rt chloebalaoing climate change is a global is...

Name: message, dtype: object

after clean mentions

15814 rt ezlusztig they took down the material on gl...

15815 rt washingtonpost how climate change could be ...

15816 notiven rt nytimesworld what does trump actual...

15817 rt sarasmiles hey liberals the climate change ...

15818 rt chetcannon kurteichenwalds climate change e...

Name: message, dtype: object

3.7 Tokenization

Word tokenization is the process of splitting a large sample of text into words. This is a requirement in natural language processing tasks where each word needs to be captured and subjected to further analysis like classifying and counting them for a particular sentiment etc.

we use the word\_tokenize method to split the paragraph into individual words.

after tokenise

0 [[, 'polyscimajor, ', ,, 'epa, ', ,, 'chief, '...

1 [[, 'its, ', ,, 'not, ', ,, 'like, ', ,, 'we, ...

2 [[, 'rt, ', ,, 'rawstory, ', ,, 'researchers, ...

3 [[, 'todayinmaker, ', ,, 'wired, ', ,, 'was, '...

4 [[, 'rt, ', ,, 'soynoviodetodas, ', ,, 'its, '…

3.8 Remove stopwords

Data Cleaning plays important role in NLP to remove noise from data.

Stopwords : refers to the most common words in a language (such as “the”, “a”, “an”, “in”)

which helps in formation of sentence to make sense, but these words does not provide

any significance in language processing so remove it .

3.9 PorterStemmer

Stemming is the process of producing variants of a root/base word.

Stemming programs are commonly referred to as stemming algorithms or stemmers.

A stemming algorithm reduces the words “chocolates”, “chocolatey”, and “choco” to the root word, “chocolate”

3.10 Lemmatization

Major drawback of stemming is it produces Intermediate representation of word. Stemmer may or may not return meaningful word.

To overcome this problem Lemmatization comes into picture.

Stemming algorithm works by cutting suffix or prefix from the word.On the contrary Lemmatization consider morphological analysis of the words and returns meaningful word in proper form.

Hence,lemmatization is preferred.

**4.Modeling**

4.1 TF-IDF stands for Term Frequency-Inverse Document Frequency

“Term frequency–inverse document frequency, is a numerical statistic that is intended to

reflect how important a word is to a document in a collection or corpus.”

Term Frequency: is a scoring of the frequency of the word in the current document.

Inverse Document Frequency: is a scoring of how rare the word is across documents.

TF-IDF model contains information on the more important words and the less important ones as well.

**X\_train before vectoriser.transform X\_train**

8249 ['[', "'on", "'", ',', "'climate", "'", ',', "...

4822 ['[', "'rt", "'", ',', "'stevesgoddard", "'", ...

13862 ['[', "'rt", "'", ',', "'cookiebo", "'", ',', ...

7759 ['[', "'rt", "'", ',', "'wickedbeaute", "'", '...

8953 ['[', "'rt", "'", ',', "'sensanders", "'", ','...

...

13418 ['[', "'rt", "'", ',', "'safeagai", "'", ',', ...

5390 ['[', "'rt", "'", ',', "'businessinsider", "'"...

860 ['[', "'rt", "'", ',', "'bramnessellen", "'", ...

15795 ['[', "'china", "'", ',', "'", '’', "'", ',', ...

7270 ['[', "'rt", "'", ',', "'climatecentral", "'",...

Name: message, Length: 15028, dtype: object

**X\_train after vectoriser transform X\_train**

(0, 133697) 0.2567699012185734

(0, 133662) 0.15239565557871654

(0, 128604) 0.20642513448110705

(0, 128603) 0.20642513448110705

(0, 119019) 0.14817950555839499

4.2 Train models

The following training algorithms were used to train and evaluate the data

BernoulliNB

LinearSVC

LogisticRegression

RandomForestClassifier

**3.Findings**

The best performing algorithm was LinearSVC

precision recall f1-score support

Below is the scoring with an average weighted f1-score of 0.77

precision recall f1-score support

-1 0.80 0.57 0.67 61

0 0.71 0.41 0.52 123

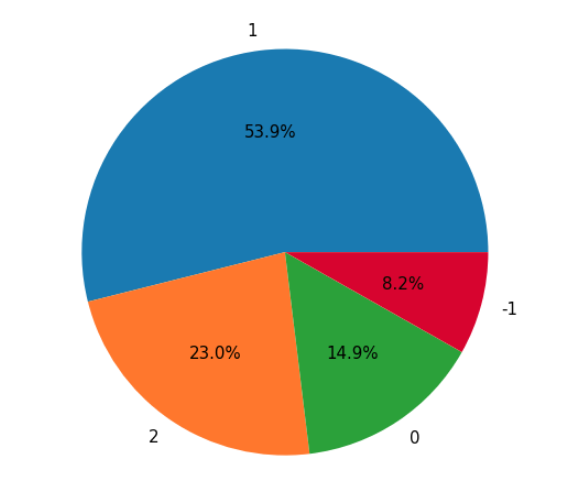
1 0.79 0.91 0.85 441

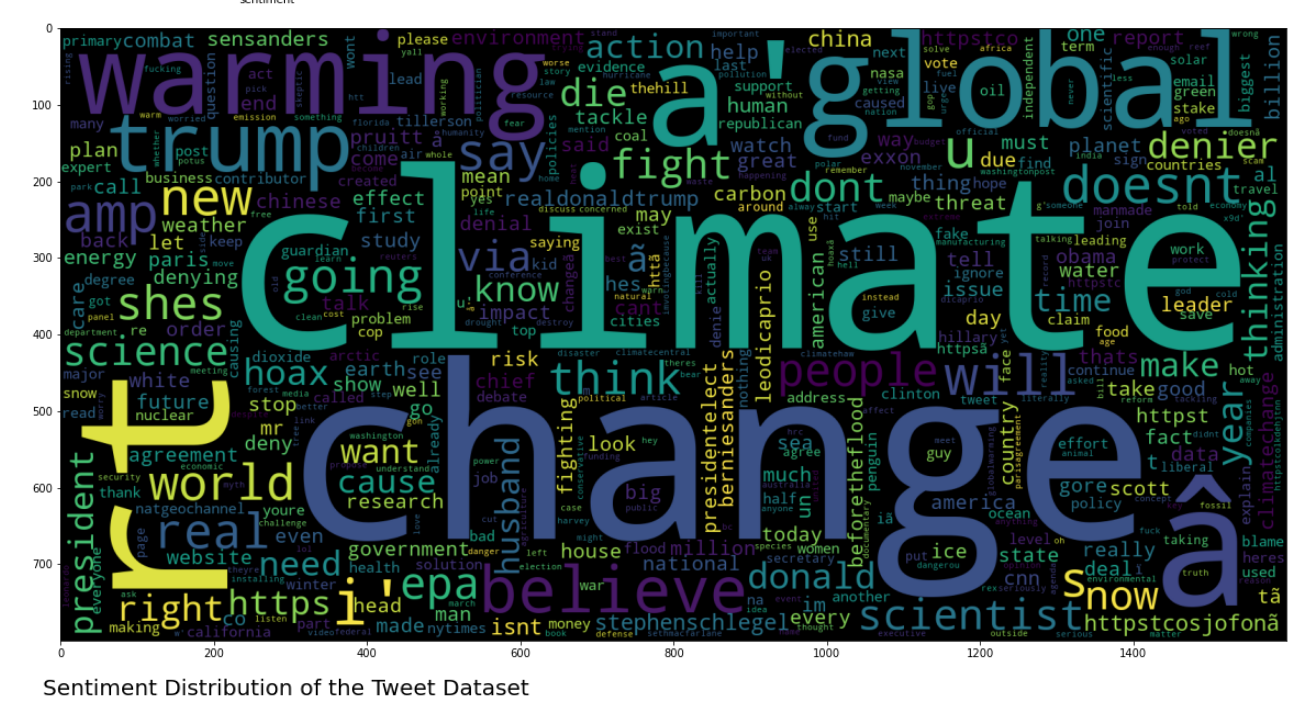
2 0.80 0.83 0.82 166

accuracy 0.79 791

macro avg 0.78 0.68 0.71 791

weighted avg 0.78 0.79 0.77 791

Above is the distribution of the sentiment. 53% is 1 therefore the majority of tweets are pro climate change. 23% is factual news and only 8.2% is anti climate change

Above is a word cloud. The larger the word the more important it is.