Climate Change Misinformation Detection

Harshal Harish Shah
Student Id: 1020849
Masters of Data Science
School of Mathematics and Statistics
The University of Melbourne
harshal.shah@student.unimelb.edu.au

Abstract

The world is going through a very bad phase of climate change and as per the official authorities the reason behind these problems is the influence of human activities. With the advent of digital media, the amount of data generated with respect to climate change and its skeptics that can be processed has skyrocketed. This report presents various techniques used to classify the text as a climate change fact or a skeptic using an exhaustive training data set and some machine learning techniques

Keywords: *Climate change, Logistic Regression, Naive Bayes, Feed Forward Neural Network, features, vectorizers, LSTM, etc.*

1 Introduction

You can't be distracted by the noise of misinformation - James Daly

Misinformation is the information that is presented as the truth to the world but found as fake or false in the future after proper analysis. In recent times, climate change is one of the major challenges that is putting the world into danger. However, because of all the human activities, the snow has started melting down, the temperatures are increasing drastically, the wildfires are occurring more frequently, and there are many other life-threatening problems due to climate change. As per the Intergovernmental Panel on Climate Change (IPCC), the main reason behind the increase of global warming between 1951 and 2010 was the human influence (Climate Change Australia, 2020). Approximately, 90 percent of the temperature increase during that time frame was because of the humans and their environmentally irresponsible activities. The main sources of getting the data related to climate change are the official government website, the news channels, and the environmental organizations across the world. The overwhelming sharing of fake news amid climate changes raised concerns amongst the people and the organizations that are responsible for maintaining accurate information about climate change. In this case, the information which can only be trusted was the data from the official sources. How to identify whether the information is accurate or it is a skeptic? Understanding climate change denial is one of the most important aspects of the economies across the globe. For example, the information is always in the contradictory or negative theme (global warming isn't happening because of humans).

Example of Climate Change Information: Australia's climate has warmed by 0.9 degree Celsius since 1910, and the pace of warming has accelerated since 1950. Over the past 10 years, the temperatures over land have been the equal-highest on record, whereas those for the sea surface have been the warmest ever.

Example of Climate Change Misinformation: Why climate change seems to have faded from the news in USA Francis Menton, in the US journal 'Manhattan Contrarian' explains why climate change seems to have faded by showing data from the easily-available UAH global lower troposphere record, derived from satellite sensors. That record exists from 1979 to present, shown in the latest chart from UAH going through the end of June 2018.

2 Dataset

To construct the final data set, the textual data for the the climate misinformation was provided by the teaching team for Natural Language Processing (COMP90042). The information which is not related to the climate change was scraped from many

Type of Articles	Target
Climate change Misinformation	1
Climate Change Facts	0
Information not related to Climate Change	0

Table 1: Dataset Info.

different sources. These sources include the news headlines and articles, the official government websites, the information provided by the organisations that deal with climate change.

- The extra information which is not related to climate change was scraped from inshort.com, which is a news website from India.
- The articles scraped from the news websites were from the English speaking countries which include the USA, UK and Australia. These relevant news websites are The Guardian, the Economist, Fox News, Reuters, Betoota Advocate, and the official organisations such as CSIRO, NASA and Climate Change Australia.
- The number of characters in the text is greater than 300 and there is no upper limit.
- Three different types of textual information available in the data set are specified in Table 1 along with their labels.
- The number of sentences related to climate information skeptics are 1168. The total number of sentences that account for factual information related to climate change are 1724 and the ones that are not in the context of climate change nor in the context of misinformation are 175.

3 Methodology

To build a classifier that detects the climate change misinformation, there are a series of steps which were followed. Amongst those steps, the first one is preprocessing of the data. The main reason behind tidying up the data is to make it more sensible. Data cleaning and preprocessing has a significant impact on the performance of the machine learning models.The various options considered in the preprocessing of the data is:

1. Remove special characters
2. Remove numbers
3. Stripping off extra white spaces and any other irrelevant spaces
4. Word Tokenisation
5. Removing single characters (length is 1)
6. Removal of stop words
7. Stemming or lemmatization (Lemmatization worked better)
8. Joining the processed words to create sequences
9. Encoding the target labels

Feature engineering is a very important step in any machine learning process. The numeric machine learning models used in the project use vector representations as their input. Converting the documents into its vector representations give us the ability to generate all the meaningful insights from the data. Each element of the vector representation is a feature and depicts a unique property about the data set. In this project, there were various features using which the models predicted the final output. Some of them are the Count Vectorizers, Hash Vectorizers, TF-IDF Vectorizers and the TF-IDF Vectorizers on N-Gram and Character level.

The most popular feature is the count vectorizer or in other terms the bag of words. It helps us store the count of all the words present in the text and use that as the feature while training the machine learning model. For example, ”Climate change is the most most important change in today’s world”. Here the world change appears 2 time, most appears 2 times and the other words appear once. A dictionary with a key and value is used as the feature. The performance of count vectorizers degrades when the vector is sparse i.e. we have a very big size of vocabulary and that is the reason Hash Vectorizer was also implemented. Hashing doesn’t use up all the space in memory and hence efficient in terms of memory and time (O’Reilly,

2014).

If there is a spell mistake or the data is vague, the count and the hash vectorizer won't ignore that. For example, the words "climate" and "climte" will be given the same importance but TF-IDF (both character and N-gram level) does not work like that. More importance is given to the word "climate" as it is more useful. TF-IDF is also based on the bag of words, but it calculates the scores based on the frequency count in the entire document rather than a single row (O'Reilly, 2014).

All the preprocessed data was further sent to the machine learning models for training and predicting the final output. The various machine learning algorithms helpful in this context are elaborated below. They are: Logistic Regression, Multinomial Naive Bayesian, Random Forest Classifier, Feed Forward Neural Network and Long-Short Term Memory(LSTM).

All the models are provided by python's Scikit Learn and Keras libraries. In the training phase, the hyper parameters of the models are tuned to get the best possible combination of the parameters and the best optimal solution is then applied on the test set. The feature vectors are supplied as the input to all of the above models and the final predictions are generated. The three basic machine learning models are initially used and the based on their output, we need to understand that whether there is a requirement of deep learning models in this project or not. The feed forward network and the LSTM model are based on 4 layers and it uses optimizer as 'SGD' and binary cross entropy as the loss, since this is a binary classification problem. One is the input layer with 'sigmoid' activation function, 2 hidden layers with 'relu' activation function and the final output layer with 'sigmoid' activation function. The input dimensions are the vocabulary size generated from the training data set and the final embedding for output as 10.

4 Analysis and Results

The analysis of the climate change misinformation detection system is based on the development set. The first step was to decide the best feature vector for the basic machine learning models and after training the model and evaluating its performance on the development set, we get the following results.

Based on the results in the Figure 1, the best feature is the Term Frequency - Inverse document frequency on N-gram level. The values for the accuracy were obtained after tuning the hyper parameters of the models. The best optimal values of the parameters for each of the model are specified in the Table 2.

Model	Parameter	Value
Logistic Regression	C	1.0
Naive Bayes	Alpha	5.4
Random Forest	Estimators	900

Table 2: Hyperparameter Values

Considering the advanced machine learning models, the results we get for feed forward neural networks and the LSTM model using bag of words and sequences as their input features are provided in the Figure 2. As per the results obtained, the best model to achieve better results for the development set is he feed forward neural network with 2 hidden layers, optimizer as the Stochastic Gradient Descent and loss as the binary cross entropy where the input feature is the bag of words. After comparing the accuracy and the results achieved from the basic and the advanced machine learning models, for the available data and the limited number of features, Logistic Regression gives us the best predictions. Although, the results are not very accurate but as compared to rest of the models, the accuracy and the F1-score is pretty high. Error Analysis for all the above models was performed but in this report, the error analysis for the development set mentioned is based on the Logistic Regression model with its best optimal value for the hyperparameter. Figure 3 displays the metrics for the Logistic Regression model.

After inspecting the predictions made by the model manually on the development set, it was observed that out of 100, 15 values very predicted incorrectly. The statistics for those values are mentioned in the table 3. The number of '1' that are predicted incorrectly i.e they are predicted to as '0' are 3 whereas the number of '0' that are predicted as '1' are 12. There are several reasons that the model predicts the value incorrectly. there could be various reasons behind this. Some of them are:

1. Insufficient Dataset
2. Removing stop words may change the mean-

	Accuracy of the Model based on the Develpment Set			
Models/Features	Count Vectorizers	TF-IDF	TF-IDF N-gram	TF-IDF character
Random Forest	0.51	0.53	0.69	0.49
Logistic Regression	0.71	0.84	0.85	0.6
Multinomial Naive Bayesian	0.5	0.6	0.79	0.57

Figure 1: Accuracy for Basic Machine learning models

	Accuracy of the Model based on the Develpment Set	
Models/Feature	Bag of Words	Sequences
Feed Foward Neural Network	0.69	0.58
LSTM	-	0.5

Figure 2: Accuracy for Deep learning models

	precision	recall	f1-score	support
0	0.93	0.76	0.84	50
1	0.80	0.94	0.86	50
accuracy			0.85	100
macro avg	0.86	0.85	0.85	100
weighted avg	0.86	0.85	0.85	100

Figure 3: Logistic Regression Classification Report

ing of the sentence.

3. Use of contractions is making the process repetitive. (we're and we are are two different strings for the machine but the meaning is same)

4. Also, the feature engineering could be an issue here. Term Frequency-Inverse Document frequency does not consider semantic similarity between words and thus if the vocabulary is big, then the errors are bound to come. (weather and climate are synonyms, but are considered as different words)

Figure 4 shows the learning curve for the logistic regression model with 10 fold cross validation. As we can see that the accuracy of the model increases as the number of training instances increases. But towards the end both the curves are about to converge. This means that adding more data will not help in this case. As the number of training data increases, the gap between the two curves decrease thereby indicating that the variance is low. In this case, on the development set, the accuracy is high and therefore the model is a good fit but when it comes to a bigger dataset (test data) then this is not the case. The accuracy is low and the error rate is high but the curves will follow a similar trend for the model. From this, we deduce that more work is needed to find the data with a very good quality and performing better feature engineering (Machine Learning Mastery, 2019).

5 Leaderboard Score Analysis

The F1 score after the evaluation of the Logistic Regression model on the final test dataset is 0.52 with precision as 0.36 and recall as 0.95. With these results, it is visible that the system is returning almost all answers but most of them are predicted incorrectly. In this system, as per the results obtained from the development set, the labels '0' are predicted as '1' and thereby returning wrong answers which has a severe effect on precision. As per the ranking on the leaderbaord, there has been a slight improvement since the start but it can be improved more if proper enhancements are made.

6 Conclusion and Further Enhancements

The goal of the project was to design a system that detects the climate change misinformation and that was done successfully with the help of vectorizers and basic machine learning models. The system designed is not up to the mark and it needs a lot of improvement. In future, to improve the performance of the model the first step should be to gather the quality data and then cleaning it thoroughly by removing noise, less popular words,etc. For feature engineering, there are various pre-trained models which can be used. Also, there are specific cost based measures, distance (similarity) based measures which can be implemented. Normalizing the corpus and keeping it domain oriented helps a lot as far as the accuracy of the models is considered. The key thing to improve the model's accuracy after data cleaning and feature engineering is to tune the hyperparameters of the model and thereby making the predictions on the optimal value of the parameters for that particular model which in turn will give us the best results.

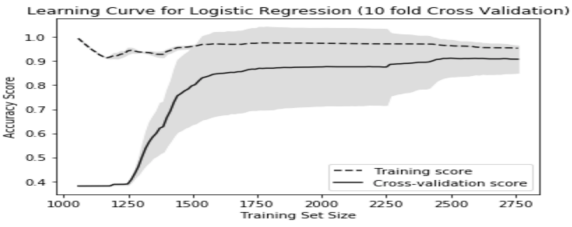

Figure 4: Learning Curve for Logistic Regression model

References

Climate Change Australia (2020). *Causes of Past and Recent Global Climate Changes*. Retrieved from : https://www.climatechangeinaustralia.gov.au/en/climate-campus/global-climate-change/causes-of-change/

O'Reilly (2014). *Chapter 4. Text Vectorization and Transformation Pipelines*. Retrived from: https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/ch04.html

Machine Learning Mastery (2019). *How to use Learning Curves to Diagnose Machine Learning Model Performance*. Retrieved from: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-\learning-model-performance/