SENTIMENT ANALYSIS OF GLOBAL WARMING

Introduction

Global Warming has become a major environmental concern in recent years. Rapid and massive industrial growth, urbanization, overall increase in the number of vehicles and appliances, increase in the usage of fossil fuels etc. have posed some major threats to the planet Earth and ‘Global warming’ is one of such hazardous environmental threats, the effects of which can be evidently observed though the changing weather patterns and overall increase in average temperature all around the world.

With the increase in problems pertaining to the ecological balance of planet Earth, We have also observed an increase in the general public awareness regarding these problems. People are speaking up, expressing their concern and trying to spread awareness regarding the dire need to protect our environment and this is how ignorance is gradually being eliminated from society. When it comes to speaking up and expressing your opinion and to make it reach billions of people , there is no better way than putting it up on social media. Among all the social media platforms, Twitter has been the first choice of people to put up their opinions and it definitely has the potential to be a strong digital forum for discussions over several issues and so is the case with global warming. Over the years people have been tweeting about how global warming is a serious issue that we've been facing, what can be the consequences if the ignorance persists and what are the possible solutions to curb it or atleast to slow down its rate.

Since Twitter is a free medium which can be accessed by all, there are various opinions regarding this issue which necessarily do not match with each other. So in this project we are going to perform the sentiment analysis for global warming and will try to find if the sentiment is towards or against the existence of global warming i.e. whether people believe that global warming is occurring or not and for the same we have obtained a dataset that contains tweets which represent people’s sentiments over global warming.

Related Work

A lot of research work has been done regarding global warming and climate change to find out the reasons behind it, and to look for the possible solutions. Although the majority of the work that has been published does not include any statistical and/or mathematical studies or procedures, they are majorly theoretical studies.

Sentiment analysis has become one of the most popular fields of study among researchers in recent years. It's being brought under use in so many domains such as to gauge the public sentiment over a newly launched product, a newly released movie/show etc. People's sentiment towards a government and its policies can also be studied using sentiment analysis and it can also be used to know the general opinion about various issues such as the one that we'll be implementing in our project.

Some of the related works are mentioned below.

◉ Study of Twitter sentiment analysis using machine learning algorithms on Python

The above mentioned paper is based on the sentiment analysis of twitter data using machine learning techniques and sentiment classifiers like Bayesian logistic regression, Naive Bayes, Support Vector Machine algorithm, Maximum entropy classifier, Case Base reasoning, Artificial neural network, ensemble classifier etc.

◉ Statistical analysis using machine learning for business intelligence

The given paper focuses on how sentiment analysis can be an effective and efficient tool to evaluate and analyze opinions of users, their reviews, feedbacks and suggestions available over the web resources and how it can improve the decision making process of a particular business domain.

◉ Scientific text analysis using machine learning techniques

The given paper is based on sentiment analysis in the scientific domain. Several algorithms such as SVM, decision trees, random forest classifiers etc. have been used to analyze the sentiments of citation sentences extracted from different research papers and it also focuses on the linguistic differences in the domain.

◉ **A comparative study of sentiment analysis using NLP and other machine learning techniques for US Airline Twitter data**

This objective of this research is find out the people's opinion about the services offered by US airlines and to gauge the level of overall customer satisfaction and for the same the researcher has used the data from 2007-2011 about US airline's quality service and customer satisfaction and has used techniques of NLP like Bag of Words (BOW) and ML methods like logistic regression , multilevel Naive bayes etc.

◉ **Sentiment Analysis of English text with multilevel features**

The above mentioned paper studies sentiment analysis on English text based on multiple features including emotions as well. It proposes an analysis method that integrates multiple features and construct 3 features which are sentiment value feature , emotion feature and improved semantic feature and then combines these 3 to build a text sentiment classification model.

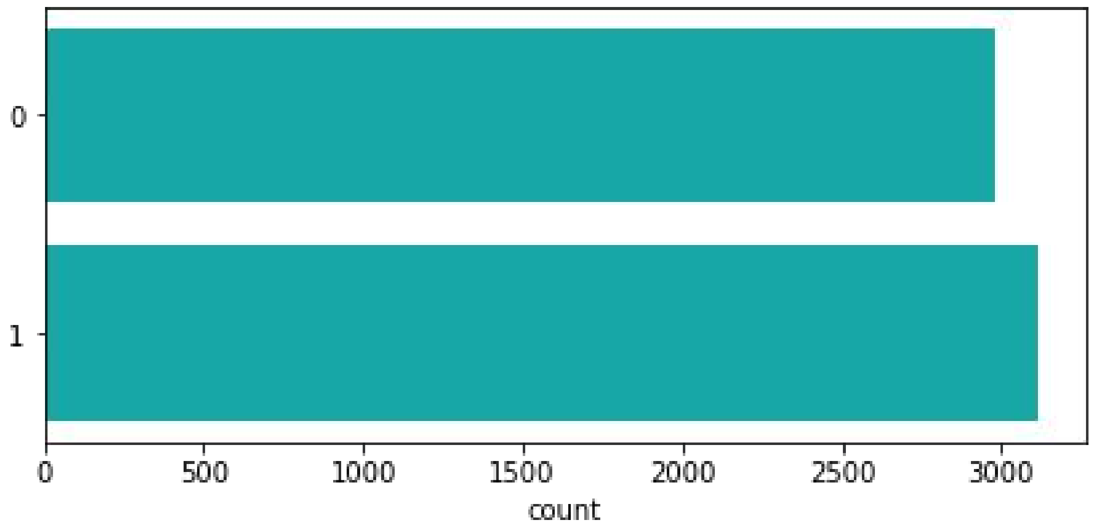
Methedology

Data Collection

Our dataset contains 6027 different tweets regarding global warming, some of them indicate the existence of global warming while others negate its existence. The dataset has 6027 rows and 3 columns namely 'tweet' , 'existence' and 'existence_confidence'.

Tweet : This column contains all the tweets that have been made by different users expressing their opinions about global warming. The data type of this column is string.

Existence: This column is a string type column and contains values in the form of Yes/No. If the tweet indicates the existence of global warming, it would be 'Yes' and if the tweet negates the existence of global warming, it would be 'No' and if the tweet is unrelated to global warming the column value would be N/A



Data Preparation

Text Cleaning is necessary for preparing the data for NLP so that machine can understand human language. The text cleaning is done in five steps which includes : normalizing the text, removing the unicode characters, removing stopwords, lemmatization and tf-idf vectorization.

Normalizing the text means removing the capitalization that would confuse a computer model i.e 'Hey' becomes 'hey'. We noticed that there is fair bit of noise since URL, emojis is converted into unicode thus making them unhelpful for analysis. Thus we need to eliminate the unicode characters. The same concept applies to punctuations aswell. Removing a stopwords is also plays important in model accuracy. A number of stopwords list for english and other language exists and can be easily applied for our problem. Lemmatization groups the word based on the root defination so that they can be identified by the word's lemma. TF-IDF stands for “Term Frequency - Inverse Document Frequency”. This is a technique to quantify words in the set of documents. The score for each word is computed that signify its importance in document and corpus.

Modelling

Three classification models were taken into consideration for the study i.e Gaussian Naive Bayes, MLP Classifier and SGD Classifier. The performance of each model is observed using the standard metrics namely accuracy, precision, recall and F1-score. The metrics are utilized to compare the machine learning algorithms.

1. Gaussian Naive Bayes

Gaussian Naive Bayes is a variant of normal Naive Bayes Algorithm that supports continuous data. It is a simple model but effective probabilistic classification model in machine learning that draws influence from Bayes Theorem.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

2. MLP Classifier

MLP Classifier stands for Multi-Layer Perceptron Classifier which is same as a neural network. Unlike other classification algorithms MLP classifier relies on neural network like architecture to perform the classification task. Its multiple layer and non-linear activation functions helps it to distinguish data that are not linearly seperable.

3. SGD Classifier

SGD Classifier is a simple and very efficient approach for discriminative learning. Given the data is sparse, the classifier is able to easily scale according to the problem. The SGD classifier implements plain stochastic gradient descent which supports different penalties i.e loss function for classification.

Model Evaluation

To evaluate the model following performance parameters are used:

1) Accuracy

This measurement determines the proportion of actual cases out of a total number of observations examined.

2) Precision

It is the proportion of predicted positives that is truly positive. It models how correct we are in identifying the global warming out of all sentence in actual global warming category.

3) Recall

It tells what proportion of predicted positives are actually positive. It models how accurately our classifier was able to identify the relevant data

4) F1-Score

It is the harmonic mean of precision and recall and ranges between 0 and 1. It maintains a balance between the precision and recall of a classifier.

Results and Conclusion

To predict the 6027 different tweets regarding global warming, which indicates the existence of global global warming while other negate its existence. The examination was performed in multiple seed utilizing calculation for 2 classes titled as YES (1) and NO(0). The YES (1) means the the text belongs to the category that tells about existence of global warming and NO(0) negates its existence.

The accuracy score for MLP Classifier, SGD Classifier and Gaussian Naive Bayes algorithms are 0.72,0.74, 0.68 respectively. On the other hand, precision are 0.72, 0.76 and 0.81 respectively and recall is 0.73, 0.75 and 0.66. The F1- score of three models are 0.73, 0.76 and 0.73 respectively.

Evaluating the algorithms as per the Accuracy metric is the most reliable method as our data is balanced in nature. We see that SGD Classifier has the highest accuracy of incorrectly classified instances Hence, the best algorithm to predict sentiment of existence of global warming is SGD Classifier.