Information Retrieval Techniques

Case Study – Sentimental Analysis

Topic: Tweets on Climate Change

Abstract

Sentiment analysis (or opinion mining) uses NLP to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. Sentiment analysis is the use of natural language processing (NLP), machine learning, and other data analysis techniques to analyze and derive objective quantitative results from raw text.

In our dataset, contributors evaluated tweets for belief in the existence of global warming or climate change. The possible answers were "1" if the tweet suggests global warming is occurring, "0" if the tweet suggests global warming is not occurring. Dataset has a total of 4224 instances. Our aim is to build classification model and predict positive and negative tweets. And finding the accuracy of the model. Also sample outputs will be tested to predict whether they will be positive or negative.

Attributes:

1. Tweets

2. Existence (0, 1)

Total instances – 4224

Positive tweets – 3111

Negative tweets – 1113

Introduction

Global warming is the phenomenon of a gradual increase in the temperature near the earth's surface. This phenomenon has been observed over the past one or two centuries. This change has disturbed the climatic pattern of the earth. However, the concept of global warming is quite controversial but the scientists have provided relevant data in support of the fact that the temperature of the earth is rising constantly.

There are several causes of global warming, which have a negative effect on humans, plants and animals. These causes may be natural or might be the outcome of human activities. In order to curb the issues, it is very important to understand the negative impacts of global warming.

Some major effects of Global Warming:

- Melting of glaciers
- Climate change
- Droughts
- Diseases
- Rise in sea level
- Heat waves
- Wild fires

Method of Analysis Performed

Machine learning algorithms work only with fixed-length vector of numbers rather than raw text, the input (in this case text data) need to be parsed. The method for transforming the texts into features is called the **Bag of words model** of text, which is a commonly used method of feature extraction. The approach works by creating different bags of words that occur in the training

data set where each word is associated with a unique number. This number shows the occurrence of each word in the document. The model is called a bag of words because the position of the words is in the document discarded.

Texts generated by humans in social media sites contain lots of noise that can significantly affect the results of the sentiment classification process. Moreover, depending on the features generation approach, every new term seems to add at least one new dimension to the feature space. That makes the feature space more sparse and high-dimensional. Consequently, the task of the classifier has become more complex. To prepare messages, such text preprocessing techniques as replacing URLs and usernames with keywords, removing punctuation marks and converting to lowercase were used in this program.

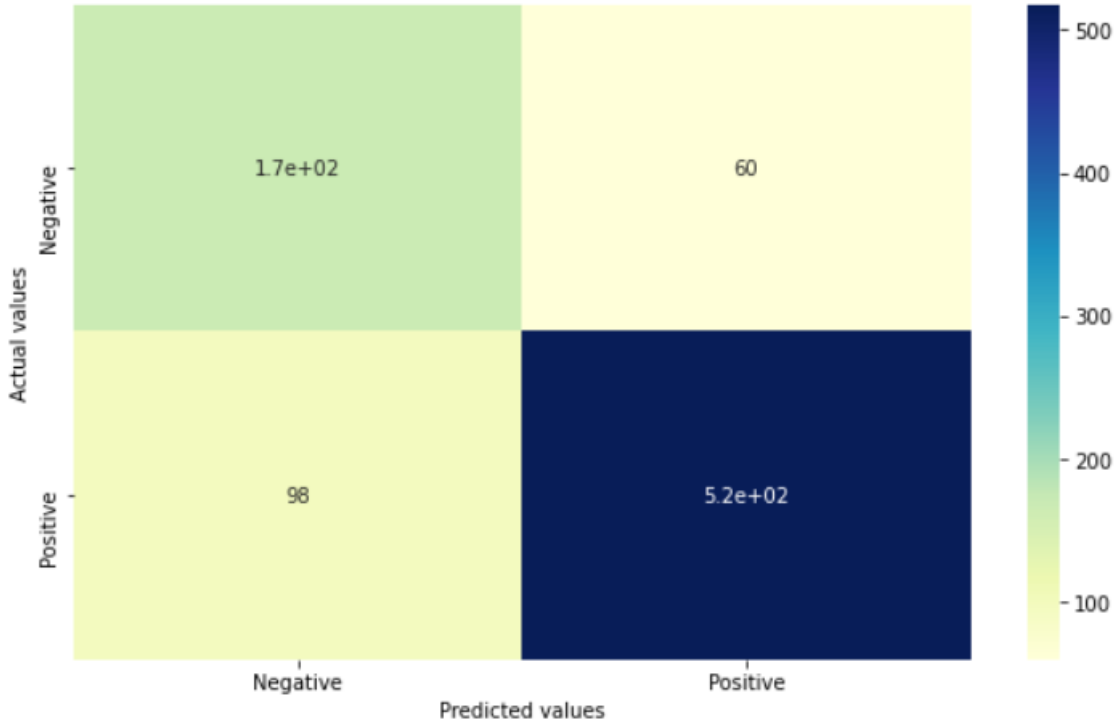
Naïve Bayes Classification model:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. When used for textual data analysis, such as Natural Language Processing, the Naive Bayes classification yields good results. Simple Bayes or independent Bayes models are other names for nave Bayes models. All of these terms refer to the classifier's decision rule using Bayes' theorem. In practice, the Bayes theorem is applied by the Naive Bayes classifier. The power of Bayes' theorem is brought to machine learning with this classifier. Multinomial Naive Bayes classification algorithm tends to be a baseline solution for sentiment analysis task. The basic idea of Naive Bayes technique is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes.

Analysis and discussion of the result

To carry out the experiments, each classifier algorithm needs to be trained before being tested. In order to train and use the classifiers, the data was divided into two data sets as training and testing data sets. After using SentimentIntensityAnalyzer()

Method we were able to see that, about 616 has positive reviews, 229 has negative reviews.



169	60
98	518

Conclusion

In this paper, we proposed machine learning NLP techniques for classification of restaurant reviews. We removed stop words from the given dataset and apply stemming for efficiency.

The data was divided into two data sets as training and testing data sets. We have fitted Naive Bayes to the training set and predicted the test results. Then we have plotted a confusion matrix for the test. We got an accuracy of 81.3 %, Precision of 0.9 and recall of 0.84.

```
---- Scores ----
Accuracy score is: 81.3%
Precision score is: 0.9
Recall score is: 0.84
```

Sample tweet prediction:

Sample = “global warming is occuring”

```
This is a POSITIVE review.
```

Sample = “global warming is not occuring”

```
This is a NEGATIVE review!
```

References

https://byjus.com/biology/global-warming/

https://monkeylearn.com/sentiment-analysis/

https://towardsdatascience.com/sentiment-analysis-introduction-to-naive-bayes-algorithm-96831d77ac91

https://www.conserve-energy-future.com/globalwarmingeffects.php