Sentiment Analysis of Flipkart Review Data

Input feature – Summary

Target feature- Sentiment

Steps:

1. Checking data uniformity across the classes

The dataset was highly unbalanced across the sentiment class. Used under sampling technique to make the data uniform across sentiment. This will help to reduce bias of the machine learning model. As the total number of records for minority class was 10239 which was significant, hence decided to go with random under sampling

1. Pre-processing the data

Removed HTML tags, stopwords and used stemming to convert the words into root form

1. Featurization technique

Experimented with Tf-Idf and Word2Vec to convert the words to vectors. Got better performance with Tf-Idf.

Tf-Idf takes into account the presence of a word in a document with respect to presence of the same word in other document, measuring its importance. It does not preserve the contextual relationship of words. Whereas Word2Vec maps the words to vectors in n dimensional space preserving its context with respect to the words appearing in its neighbours in the documents

1. Machine Learning model

Used Naïve Bayes, Random Forest and Xgboost to train the data (vectors).

Below are the results of the ML models used:

For TF-IDF featurization technique:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-score |
| Naïve Bayes | 76.72% | 77% | 77% | 76% |
| Random Forest | 78.19% | 78% | 78% | 78% |
| XgBoost | 78.46% | 78% | 78% | 78% |

For Word2Vec featurization technique:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-score |
| Random Forest | 70.8% | 71% | 71% | 71% |
| XgBoost | 71.01% | 71% | 71% | 71% |

The performance went down after using Word2Vec technique for getting features

As the accuracy along with other metrics were below 79%, went ahead with BERT (base-uncased) to improve the performance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-score |
| BERT (base-uncased) | 80% | 80% | 80% | 80% |

Because of GPU constraint, BERT was trained for 4 epochs only. But if trained for a greater number of epochs, it can certainly increase the precision, recall and F1-score.

BERT uses bidirectional context understanding where the relationship of each word in a sequence is calculated with other words using attention weights. This mechanism helps BERT to achieve context rich vectors along with capturing long range dependencies. Hence BERT significantly increase the performance when using for downstream tasks like text classification and sentiment analysis.