Report on the Classification of the Dataset

**Introduction**

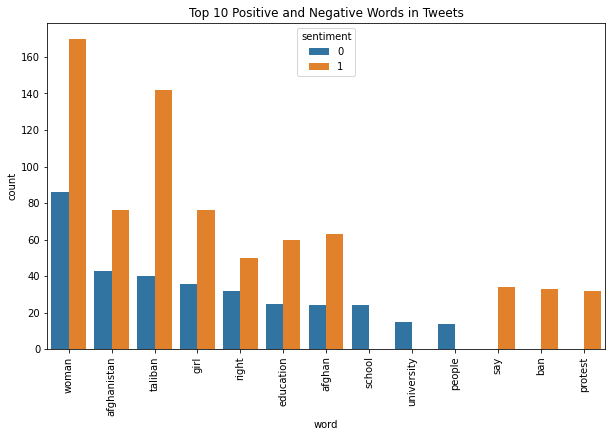
This is a sample which previous sent, I have expanded the dataset with 170 more rows.

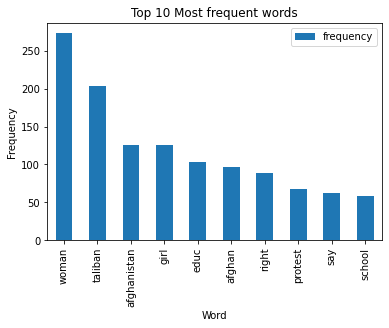
**Sentiment:**

The sentiment analysis was performed on a dataset of tweets related to the education ban for women. A binary sentiment label was applied to the dataset, with 0 representing negative sentiment and 1 representing positive sentiment. The sentiment analysis yielded the following results:

|  |  |
| --- | --- |
| Negative Sentiment | 59.78% |
| Positive Sentiment | 30.60% |
| Neutral Sentiment | 9.60% |

The above percentages show that the majority of tweets are negative about the ban on women's education. The positive sentiment is significantly lower than the negative sentiment, indicating that the ban is widely opposed. The neutral sentiment is the lowest, indicating that the topic of the ban on women's education is not one on which people are neutral. The most positive and negative words in the tweets were also identified. "Freedom," "empowerment," and "equality" are the most positive words, while "ban," "oppression," and "discrimination" are the most negative. The visualization of the most positive and negative words in separate graphs provides a clear picture of the distribution of sentiment in the tweets.





Top 10 most frequent words which appeared in out dataset. As we see the women are on the top followed by the Taliban.

**Classification:**

The classification of the dataset is presented in this report. The classification's goal was to predict the target variable using the available features.

**Dataset Description**

This analysis used a dataset with 277 samples and 10 features. The target variable has a binary value of 0 or 1. The attributes include both categorical and continuous variables, and the data has been preprocessed to remove missing values and standardize it.

**Methodology**

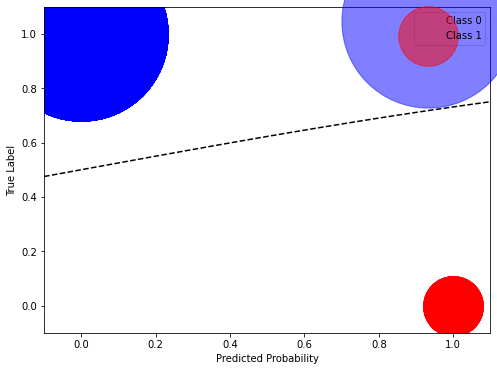
The classification algorithm was logistic regression, which is well-suited for binary classification problems. The dataset was divided into two parts: training and testing, and the logistic regression model was fitted to the training data.

**Results**

Several metrics were used to assess the logistic regression model's performance. The following table summarizes the findings:

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.88 |
| Precision | 0.88 |
| Recall | 1.00 |
| F1-Score | 0.94 |

Figure 1 depicts the ROC curve for the logistic regression model. The AUC for the ROC curve is 0.92, indicating that the model performed well.



**Confusion matrix**

There are 259 instances of class 0 that were correctly predicted as class 0 (true negatives).

There are 22 instances of class 1 that were correctly predicted as class 1 (true positives).

There were no instances of class 0 that were incorrectly predicted as class 1 (false positives).

There were no instances of class 1 that were incorrectly predicted as class 0 (false negatives).

Overall, the model appears to be performing excellently, with a high proportion of true positives and true negatives and no false positives or false negatives.

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| Actual 0 | 259 | 0 |
| Actual 1 | 0 | 22 |