Naive Bayes Text Classification for "Stressed" vs "Not Stressed"

To determine whether a text is 'stressed' or 'not stressed' based on probability, we can use a classification algorithm like Naive Bayes. Naive Bayes is particularly well-suited for text classification problems because it works well with large feature spaces (words in the text) and makes the simplifying assumption that features are independent given the class.

Step-by-Step Process

Step 1: Calculate Prior Probabilities

Prior probabilities are the probabilities of each class (stressed or not stressed) in the training dataset.

P(Stressed) = Number of stressed texts / Total number of texts

P(Not Stressed) = Number of not stressed texts / Total number of texts

Step 2: Calculate Likelihoods

The likelihood is the probability of each word appearing in a text given the class. For each word w in the vocabulary, calculate:

P(w|Stressed) = (Number of times w appears in stressed texts + 1) / (Total number of words in stressed texts + Vocabulary size)

P(w|Not Stressed) = (Number of times w appears in not stressed texts + 1) / (Total number of words in not stressed texts + Vocabulary size)

Adding 1 is for Laplace smoothing to handle words that might not appear in the training dataset.

Step 3: Calculate Posterior Probabilities

For a given text, the posterior probability is the probability that the text belongs to a class (stressed or not stressed) given the words in the text.

For a text T = {w1, w2, ..., wn}:

$P(\text{Stressed}|T) = P(\text{Stressed}) \times \Pi\, P(w_i|\text{Stressed})$

$P(\text{Not Stressed}|T) = P(\text{Not Stressed}) \times \Pi\, P(w_i|\text{Not Stressed})$

Step 4: Make a Prediction

Compare the posterior probabilities:

If P(Stressed|T) > P(Not Stressed|T) then classify as Stressed

Otherwise, classify as Not Stressed

Example Calculation

Given a small example dataset with the following texts and labels:

Text: 'I am feeling very stressed', Label: Stressed

Text: 'I am so relaxed today', Label: Not Stressed

Text: 'This is a stressful situation', Label: Stressed

Text: 'I feel calm and happy', Label: Not Stressed

Step 1: Calculate Prior Probabilities

P(Stressed) = 2/4 = 0.5

P(Not Stressed) = 2/4 = 0.5

Step 2: Calculate Likelihoods

Assuming the vocabulary is: {'I', 'am', 'feeling', 'very', 'stressed', 'so', 'relaxed', 'today', 'This', 'is', 'a', 'stressful', 'situation', 'feel', 'calm', 'and', 'happy'}

Using Laplace smoothing (adding 1 to each count):

P(stressed|Stressed) = (2+1) / (5+17) = 3/22 ≈ 0.136

P(stressed|Not Stressed) = (0+1) / (5+17) = 1/22 ≈ 0.045

Step 3: Calculate Posterior Probabilities

For a new text 'I am stressed':

P(Stressed|'I am stressed') = P(Stressed) × P(I|Stressed) × P(am|Stressed) × P(stressed|Stressed)

P(Not Stressed|'I am stressed') = P(Not Stressed) × P(I|Not Stressed) × P(am|Not Stressed) × P(stressed|Not Stressed)

Step 4: Make a Prediction

Compare the posterior probabilities:

If P(Stressed|'I am stressed') > P(Not Stressed|'I am stressed'), classify as Stressed.

Otherwise, classify as Not Stressed.

Conclusion

This method allows you to classify texts as 'stressed' or 'not stressed' based on the probabilities calculated from your dataset. By using Naive Bayes, you leverage the word occurrences in your dataset to make probabilistic predictions about new, unseen texts.