"Apple announced a new MacBook Pro with faster processors." classify the statement using naive bayes algorithm

Step 1: Preprocessing

First, we need to preprocess the statement:

Tokenization: Split the statement into individual words or tokens.

Lowercasing: Convert all tokens to lowercase to ensure consistency.

Remove Stopwords: Remove common words that do not contribute to the meaning (like "a", "the", "with", etc.).

After preprocessing, the statement might look like:

Step 2: Feature Extraction

Next, represent the preprocessed statement as a vector of word counts (bag-of-words representation):

Define a vocabulary based on all unique words encountered in the training data.

Convert the statement into a vector where each element represents the count of a word in the vocabulary as given below

Vocabulary: [apple, announced, new, macbook, pro, faster, processors]

Statement Vector: [1, 1, 1, 1, 1, 1, 1]

Step 3: Naive Bayes Classification

Now, apply Bayes' theorem to calculate the probability of the statement belonging to each category (in this case, let's consider two hypothetical categories: Technology News and Other):

Prior Probabilities (from training data):

P(Technology News)

P(Other)

Conditional Probabilities (word likelihoods from training data):

For each word wiwi in the statement vector:

P(wi | Technology News)

P(wi | Other)

Calculate the posterior probabilities P(category|statement) for each category using Bayes' theorem:

$$P(\text{category}|\text{statement}) \propto P(\text{category}) \cdot \prod_i P(\text{word}_i|\text{category})$$

Simple calculations

Assume we have the following hypothetical probabilities (these would be derived from the training data):

Prior Probabilities:

P(Technology News)=0.6

P(Other)=0.4

Conditional Probabilities (for illustration purposes):

P(apple|Technology News)=0.3

P(announced|Technology News)=0.2

P(new|Technology News)=0.4

P(macbook|Technology News)=0.5

P(pro|Technology News)=0.3

P(faster|Technology News)=0.1

P(processors|Technology News)=0.2

Similarly, you would have conditional probabilities for the category Other.

Calculate $P(\text{Technology News}|\text{statement})$ and $P(\text{Other}|\text{statement})$ using the formula:

$$P(\text{category}|\text{statement}) \propto P(\text{category}) \cdot \prod_i P(\text{word}_i|\text{category})$$

For example, assuming the actual probabilities are computed:

$$P(\text{Technology News}|\text{statement}) \propto 0.6 \cdot (0.3 \cdot 0.2 \cdot 0.4 \cdot 0.5 \cdot 0.3 \cdot 0.1 \cdot 0.2)$$

P(Other|statement)∝0.4

(product of conditional probabilities for Other category)

Conclusion:

Compare P(Technology News|statement) and P(Other|statement). The category with the higher posterior probability is the predicted category for the statement "Apple announced a new MacBook Pro with faster processors." according to the Naive Bayes algorithm.