Detecting Hate Speech Using Naive Bayes

A Step-by-Step Probabilistic Text Classification Exercise

Andreas Kollias

The goal is to **"train" a Naive Bayes classifier** to determine whether a new message shows signs of hate speech and test it on a new test message.

Test Message: "threat law hate"

Training

The training set consists of a small collection of labeled sentences designed to simulate a simple hate speech detection task. Each sentence is manually classified into one of two categories:

- **Hate** – messages containing harmful or threatening language
- **Not Hate** – neutral or non-threatening messages

Purpose of the Training Set

We have 4 short messages labeled as either "Hate" or "Not Hate".

Message	Message	Label
1	I **hate** those people—they bring nothing but **crime** and **violence**	Hate
2	freedom justice peace	Not Hate
3	attack threat destroy	Hate
4	rights equality law	Not Hate

Use stopwords to remove words that have no importance in hate speech detection.

Word Counts per Class		
Word	Count in Hate	Count in Not Hate

hate	1	0
crime	1	0
violence	1	0
attack	1	0
threat	1	0
destroy	1	0
freedom	0	1
justice	0	1
peace	0	1
rights	0	1
equality	0	1
law	0	1
TOTAL	**6**	**6**

Vocabulary size V=6+6=**12**.

Class priors: The prior probabilities for each class in the Naive Bayes model

$$P(\text{Class}) = \frac{\text{Number of messages in class}}{\text{Total number of messages}}$$

Class	Count	Prior
Hate	2	0.5
Not Hate	2	0.5

These values represent the probability of a message belonging to each class before seeing any words—that is, based purely on the distribution of classes in the training data.

Even if the words in a message are more associated with one class, a very high prior for another class can still influence the outcome.

If we had 100 Hate messages and only 5 Not Hate messages, what should the model expect *before* seeing any words?

Class	Count	Prior
Hate	100	0.952381

Not Hate	5	0.047619

Step-by-Step Probabilities

Test Message: "threat law hate"

$$\log P(C \mid \text{message}) = \log P(C) + \sum_{i=1}^{n} \log P(\text{word}_i \mid C)$$

$$P(\text{word} \mid \text{class}) = \frac{\text{count}(\text{word in class}) + 1}{\text{total words in class} + V}$$

For class = **Hate**

Word	Count	Formula	Smoothed Prob	Log Prob
threat	1	(1+1)/(6+12) = 2/18	**0.1111**	$\ln(0.1111) \approx -2.197$
law	0	(0+1)/(6+12) = 1/18	**0.0556**	$\ln(0.0556) \approx -2.890$
hate	1	(1+1)/(6+12) = 2/18	**0.1111**	$\ln(0.1111) \approx -2.197$
Prior	—	—	**0.5**	$\ln(0.5) = -0.693$

Total Log Probability for Hate:

$$\log P(\text{Hate} \mid \text{message}) = -0.693 + (-2.197) + (-2.890) + (-2.197) = \boxed{-7.977}$$

For class = **Not Hate**

Word	Count	Formula	Smoothed Prob	Log Prob
threat	0	(0+1)/(6+12) = 1/18	**0.0556**	$\ln(0.0556) \approx -2.890$
law	1	(1+1)/(6+12) = 2/18	**0.1111**	$\ln(0.1111) \approx -2.197$
hate	0	(0+1)/(6+12) = 1/18	**0.0556**	$\ln(0.0556) \approx -2.890$
Prior	—	—	**0.5**	$\ln(0.5) = -0.693$

Total Log Probability for Not Hate:

$$\log P(\text{Not Hate} \mid \text{message}) = -0.693 + (-2.890) + (-2.197) + (-2.890) = \boxed{-8.670}$$

Final Prediction:

Hate: -7.977

Not Hate: -8.670

The higher log-probability is for **Hate**, so the predicted class is: **Hate**