Training Set		Stop Words	
Category/Label	**Documents**	**With**	**Removed**
entertaining	the actor gives a convincing, charismatic performance as the multifaceted	10	6
	Spielberg gives us a visually spicy and historically accurate real life story	12	10
	His innovative mind entertains us now and will continue to entertain generations to come	14	9
boring	Unfortunately, the film has two major flaws, one in the disastrous ending	12	8
	If director actually thought this movie was worth anything	9	7
	His efforts seem fruitless, creates drama where drama shouldn't be	11	8
Test Set			
??	film is a innovative drama, entertains, but disastrous ending		

Prior from training

$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$ $\quad P(ent) = 3/6 = 1/2$

$P(bor) = 3/6 = 1/2$

WITH STOP WORDS

1. Summary of training and test set:
   Using word_tokenize and FreqDist functions, training set has 68 words with 58 vocabularies. 36 words categorized as "entertaining" and 32 words categorized as "boring".

2. Drop "is" and "but"

3. add-1 smoothing is used since not all words in test set are appearing in both categories.

Word	Entertaining (ent)	Boring (bor)
film	-	1
a	2	-
innovative	1	-
drama	-	2
entertains	1	-
disastrous	-	1
ending	-	1

Likelihoods from training:

$$P(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \epsilon V} count(w, c)) + |V|}$$

Calculations for each word:

P("film"|"ent") $= \frac{0+1}{36+58} = \frac{1}{94}$	P("film"|"bor") $= \frac{1+1}{32+58} = \frac{2}{90}$

P("a"|"ent") = $\frac{2+1}{36+58} = \frac{3}{94}$	P("a"|"bor") = $\frac{0+1}{32+58} = \frac{1}{90}$
P("innovative"|"ent") = $\frac{1+1}{36+58} = \frac{2}{94}$	P("innovative"|"bor") = $\frac{0+1}{32+58} = \frac{1}{90}$
P("drama"|"ent") = $\frac{0+1}{36+58} = \frac{1}{94}$	P("drama"|"bor") = $\frac{2+1}{32+58} = \frac{3}{90}$
P("entertains"|"ent") = $\frac{1+1}{36+58} = \frac{2}{94}$	P("entertains"|"bor") = $\frac{0+1}{32+58} = \frac{1}{90}$
P("disastrous"|"ent") = $\frac{0+1}{36+58} = \frac{1}{94}$	P("disastrous"|"bor") = $\frac{1+1}{32+58} = \frac{2}{90}$
P("ending"|"ent") = $\frac{0+1}{36+58} = \frac{1}{94}$	P("ending"|"bor") = $\frac{1+1}{32+58} = \frac{2}{90}$

4. Scoring the test:
   P("ent")P(S|"ent") $= \frac{1}{2} \times \frac{1 \times 3 \times 2 \times 1 \times 2 \times 1 \times 1}{94^7} = 9.25\text{e-}14$
   P("bor")P(S|"bor") $= \frac{1}{2} \times \frac{2 \times 1 \times 1 \times 3 \times 1 \times 2 \times 2}{90^7} = 2.51\text{e-}13$

STOP WORDS REMOVED

1. Summary of training and test set:
   Using word_tokenize, stopwords, and FreqDist functions, the training set has 48 words with 44 vocabularies. 25 words categorized as "entertaining" and 23 words categorized as "boring".

2. Drop "is", "a" and "but"

3. add-1 smoothing is used since not all words in test set are appearing in both categories.

Word	Entertaining (ent)	Boring (bor)
film	-	1
innovative	1	-
drama	-	2
entertains	1	-
disastrous	-	1
ending	-	1

Likelihoods from training:

$$P(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|}$$

Calculations for each word:

P("film"|"ent") = $\frac{0+1}{25+44} = \frac{1}{69}$	P("film"|"bor") = $\frac{1+1}{23+44} = \frac{2}{67}$

P(”innovative”|”ent”) $= \frac{1+1}{25+44} = \frac{2}{69}$	P(”innovative”|”bor”) $= \frac{0+1}{23+44} = \frac{1}{67}$
P(”drama”|”ent”) $= \frac{0+1}{25+44} = \frac{1}{69}$	P(”drama”|”bor”) $= \frac{2+1}{23+44} = \frac{3}{67}$
P(”entertains”|”ent”) $= \frac{1+1}{25+44} = \frac{2}{69}$	P(”entertains”|”bor”) $= \frac{0+1}{23+44} = \frac{1}{67}$
P(”disastrous”|”ent”) $= \frac{0+1}{25+44} = \frac{1}{69}$	P(”disastrous”|”bor”) $= \frac{1+1}{23+44} = \frac{2}{67}$
P(”ending”|”ent”) $= \frac{0+1}{25+44} = \frac{1}{69}$	P(”ending”|”bor”) $= \frac{1+1}{23+44} = \frac{2}{67}$

4. Scoring the test:

P(“ent”)P(S|”ent”) $= \frac{1}{2} \; x \; \frac{1 \; x \; 2 \; x \; 1 \; x \; 2 \; x \; 1 \; x \; 1}{69^6} = 1.85\text{e-}11$

P(“bor”)P(S|”bor”) $= \frac{1}{2} \; x \; \frac{2 \; x \; 1 \; x \; 3 \; x \; 1 \; x \; 2 \; x \; 2}{67^6} = 1.33\text{e-}10$