Open the Black Box

Explanation of the Use of Naive Bayes Model for NLP to Classify Reddit Posts

Kai Zhao

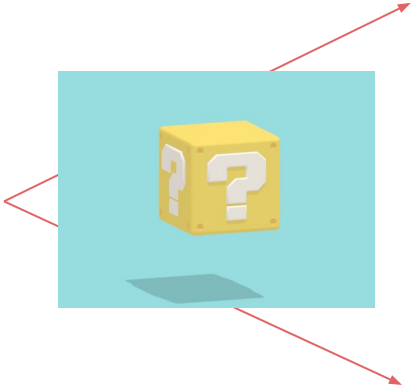
Agenda

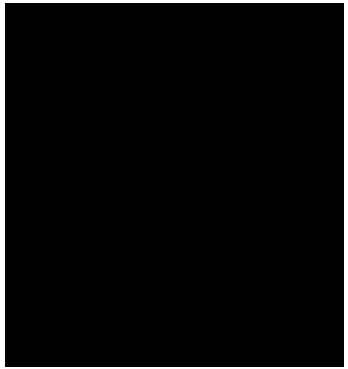
1. **Part 1: Reddit Post Classification Using NLP**
    a. State the Problem
    b. Data
    c. Methodology
    d. Results
    e. Conclusion & Recommendation
    f. Next Step
2. **Part 2: A Closer Look at Multinomial Naive Bayes Model**
    a. Reproduce MNB Model Results from Part 1
    b. Bonus: Scattertext Visualization

PART I:

Reddit Post Classification Using NLP

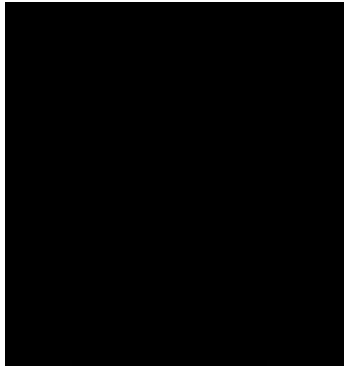
State the Problem





r/airpollution

904 Posts



r/breastcancer

983 Posts

Total 1887 Posts

Methodology

1. EDA & Data Cleaning
2. Tokenization, Lemmatization, and Vectorization
3. Train/Test Split
4. Fit and Run Model Using Pipeline and GridSearchCV
    a. **Primary Model**: Multinomial Naive Bayes
    b. **Alternative Model**: Logistic Regression (for comparison)

Results Summary - Accuracy

Naive Bayes

Metric	Baseline	CountVectorizer	TFIDFVectorizer
Accuracy Train	0.52	0.996	0.999
Accuracy Test	-	0.996	0.996
MisClassification Test	-	2	2

Logistic Regression

Best Model

Metric	Baseline	CountVectorizer	TFIDFVectorizer
Accuracy Train	0.52	1.0	0.993
Accuracy Test	-	1.0	0.994
MisClassification Test	-	0	3

Results Summary - Parameter

Naive Bayes

Metric	CountVectorizer	TFIDFVectorizer
Tokenizer	default	default
Processer	Lemmatization	Lemmatization
min_df	2	2
max_df	0.9	0.9
max_features	1000	1000
ngram_range	(1, 1)	(1, 1)
stop_words	english	english

Logistic Regression

Metric	CountVectorizer	TFIDFVectorizer
Tokenizer	default	default
Processer	Lemmatization	Lemmatization
Regulization	Lasso	Lasso
min_df	2	3
max_df	0.98	0.9
max_features	1000	500
ngram_range	(1, 1)	(1, 1)
stop_words	english	english

Conclusion & Recommendation

Conclusion:

1. For the selected two subreddits, the **Logistic Regression** model with **Lasso** regularization (i.e., 100% accuracy) performs marginally better than the **Multinomial Naive Bayes** model (i.e., 99.9% accuracy)
2. The misclassified documents would be correctly classified if reviewed by human.
3. The concern of the naive assumption (i.e., all features are independent) has little impact to the model's classification capability for this project.

Recommendation:

1. Naive Bayes model is generally a good model for text classification and should be considered as a standard choice along with other classification models.
2. Frequency of the feature matters. Setting a reasonable range of max feature options (e.g, 100, 500, 1000, 2000.. etc.) for GridSearchCV is recommended for best model results.

Next Step

1. Further evaluate the model performance by pepeating model multiple times using more similar subreddit posts.

2. Investigate the impact of number of features and MNB model's smoothing mechanism.

3. Find a case when MNB fails to understand the limitation of the model.

PART II

A Closer Look at Naive Bayes Model

Reproduce the Results from Part I

	index	post_title	bc_prob	aq_prob	class_pred	class
0	1514	With a peak over 300 AQI EPA this weekend in B...	1.693144e-02	0.983069	1	1
1	1642	A Beijing artist wore a face mask wedding dres...	3.970811e-06	0.999996	1	1
2	476	Have you been diagnosed with cancer? We need y...	9.999961e-01	0.000004	0	0
3	1007	Is my home air making me sick?	2.009316e-05	0.999980	1	1
4	639	29/M lump in breast	9.999954e-01	0.000005	0	0
5	1088	Question about Air Quality Index	1.549139e-05	0.999985	1	1
6	1622	Fewer children visited ER for asthma problems ...	1.892275e-05	0.999981	1	1
7	1100	China Renewable Energy Growth Soars & Coal...	9.155739e-06	0.999991	1	1
8	1200	A Chinese company is offering free training fo...	2.364122e-06	0.999998	1	1
9	397	Rare phyllodes tumour	8.827349e-01	0.117265	0	0
10	1458	Fire Continues to smoulder at Parkersburg, WVa...	1.070009e-04	0.999893	1	1
11	1265	Sydney under blanket of smoke as hazard reduct...	2.777599e-09	1.000000	1	1
12	1150	High CO2 Levels	1.011805e-03	0.998988	1	1

Results based on model using X_test data

Review of Bayes' Theorem

Bayes' Theorem (Bayes' law or Bayes' rule):

The probability of an event, based on prior knowledge of conditions that might be related to the event

$$P(A|B) * P(B) = P(A \cap B) \qquad P(A \cap B) = P(B|A) * P(A)$$

$$P(A|B) * P(B) = P(B|A) * P(A)$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Equation Translation

12	1150	High CO2 Levels	1.011805e-03	0.998988	1	1

$$P(BC|\text{high co2 levels}) = \frac{P(\text{high co2 levels}|BC)\ P(BC)}{P(\text{high co2 levels})}$$

Test Data | Train Data

$$P(AQ|\text{high co2 levels}) = \frac{P(\text{high co2 levels}|AQ)\ P(AQ)}{P(\text{high co2 levels})}$$

Equation Translation

$$P(\text{high co2 levels}|BC) * P(BC) = P(\text{'high'}|BC) * P(\text{'co2'}|BC) * P(\text{'levels'}|BC) * P(BC)$$

Naive!!!

BAM!

$$P(\text{high co2 levels}|AQ) * P(AQ) = P(\text{'high'}|AQ) * P(\text{'co2'}|AQ) * P(\text{'levels'}|AQ) * P(AQ)$$

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC)

X_train	high	co2	levels
bc_prob	0.001292	0.000258	0.000258
aq_prob	0.005821	0.006269	0.002537

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ)

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC)

X_train	high	co2	levels
bc_prob	0.001292	0.000258	0.000258
aq_prob	0.005821	0.006269	0.002537

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ)

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC)

X_train	high	co2	levels
bc_prob	0.001292	0.000258	0.000258
aq_prob	0.005821	0.006269	0.002537

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ)

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC)

X_train	high	co2	levels
bc_prob	0.001292	0.000258	0.000258
aq_prob	0.005821	0.006269	0.002537

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ)

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * **P(BC)**

X_train	high	co2	levels
bc_prob	0.001292	0.000258	0.000258
aq_prob	0.005821	0.006269	0.002537

X_train	count	prio_prob
bc	737.0	0.520848
aq	678.0	0.479152

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * **P(AQ)**

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC) = 4.493121 x 10^-11

X_train	high	co2	levels
bc_prob	0.001292	0.000258	0.000258
aq_prob	0.005821	0.006269	0.002537

X_train	count	prio_prob
bc	737.0	0.520848
aq	678.0	0.479152

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ) = 4.435206 x 10^-8

Revisit Equation

12	1150		High CO2 Levels	1.011805e-03	0.998988	1	1

$$P(BC|\text{high co2 levels}) = \frac{P(\text{high co2 levels}|BC)\ P(BC)}{P(\text{high co2 levels})}$$

Test Data | Train Data

$$P(AQ|\text{high co2 levels}) = \frac{P(\text{high co2 levels}|AQ)\ P(AQ)}{P(\text{high co2 levels})}$$

Revisit Equation

12	1150	High CO2 Levels	1.011805e-03	0.998988	1	1

$$P(BC|\text{high co2 levels}) = \frac{4.493121 \times 10^{-11}}{P(\text{high co2 levels})}$$

Test Data | Train Data

$$P(AQ|\text{high co2 levels}) = \frac{4.435206 \times 10^{-8}}{P(\text{high co2 levels})}$$

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC) = 4.493121 x 10^-11

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ) = 4.435206 x 10^-8

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC) = 4.493121 x 10^-11

P('high co2 levels'|BC) * P(BC)

P('high co2 levels'|AQ) * P(AQ)

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ) = 4.435206 x 10^-8

Simply Plug In Numbers

P('high'|BC) * P('co2'|BC) * P('levels'|BC) * P(BC) = 4.493121 x 10^-11

P('high co2 levels'|BC) * P(BC)

P('high co2 levels' ∩ BC)

P('high co2 levels') = 4.439699 x 10^-8

P('high co2 levels' ∩ AQ)

P('high co2 levels'|AQ) * P(AQ)

P('high'|AQ) * P('co2'|AQ) * P('levels'|AQ) * P(AQ) = 4.435206 x 10^-8

DOUBLE BAM!!

Revisit Equation

12	1150	High CO2 Levels	1.011805e-03	0.998988	1	1

$$P(BC|\text{high co2 levels}) = \frac{4.493121 \times 10^{-11}}{4.439699 \times 10^{-8}}$$

Test Data | Train Data

$$P(AQ|\text{high co2 levels}) = \frac{4.435206 \times 10^{-8}}{4.439699 \times 10^{-8}}$$

Final Verification

12	1150	High CO2 Levels	1.011805e-03	0.998988	1	1

$$P(BC|\text{high co2 levels}) = \frac{4.493121 \times 10^{-11}}{4.439699 \times 10^{-8}} = 1.011805 \times 10^{-3}$$

TRIPLE

BAM!!!

$$P(AQ|\text{high co2 levels}) = \frac{4.435206 \times 10^{-8}}{4.439699 \times 10^{-8}} = 0.998988$$

BONUS?

Visualization Using Scattertext

Quick Review of Two Concepts

1. Harmonic Mean

- arithmetic mean of **a** and **b** = (a + b)/2
- geometric mean of **a** and **b** = **sqrt**(a*b)
- harmonic mean of **a** and **b** = 2/(1/a + 1/b) = 2ab/(a+b)

"the reciprocal of the arithmetic mean of the reciprocals of the given set of observations"

Quick Review of Two Concepts

2. F-Score: the metric measures the usefulness

	Predict Breast Cancer	Predict Not Breast Cancer
Actual Breast Cancer	10	8
Actual Not Breast Cancer	2	99,980

Accuracy = (10 + 99,980) / 100,000 = 99.99% - amazingly accurate but useless

Precision = TP/(TP + FP) = 10/(10+2) = 83.3%

Recall = TP/(TP+FN) = 10/(10 + 8) = 55.6%

F-Score = 2*Precision*Recall/(Precision + Recall) = 2TP / (2TP + FN + FP) Range from 0 to 1
F-Score = 20/(20+8+2) = 66.7% much more representative. Especially for unbalanced data.

Visualization with Scattertext

Juypter Notbook Presentation
(link here)