CIS530 HW2: Submission Example

2.1 All-complex (4 points)

Results

- Train Set:
  - Precision: 0.43275
  - Recall: 1.0
  - F-score: 0.604083057058105
- Dev Set:
  - Precision: 0.418
  - Recall: 1.0
  - F-score: 0.5895627644569816

2.2 Word-length baseline (6 points)

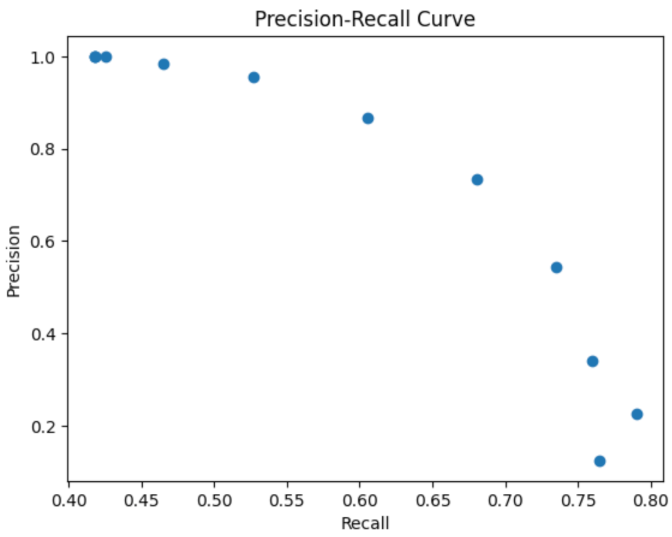
Results

- Train Set:
  - Precision: 0.6007401315789473
  - Recall: 0.8440207972270364
  - F-score: 0.7018976699495555
- Dev Set:
  - Precision: 0.6053511705685619
  - Recall: 0.8660287081339713
  - F-score: 0.7125984251968505

Threshold Analysis

- Range of thresholds tested: from 1 to 12
- Best threshold: 7

Precision-Recall Curve



P-R Curve Analysis

The value for precision and recall is always between 0 and 1. From the curve, we can see there's a tradeoff between precision and recall. When precision goes high, recall goes low, or vice versa. Based on the f-score, the largest f-score is not from the largest precision or largest recall, they are in the middle.

2.3 Word-frequency baseline (6 points)

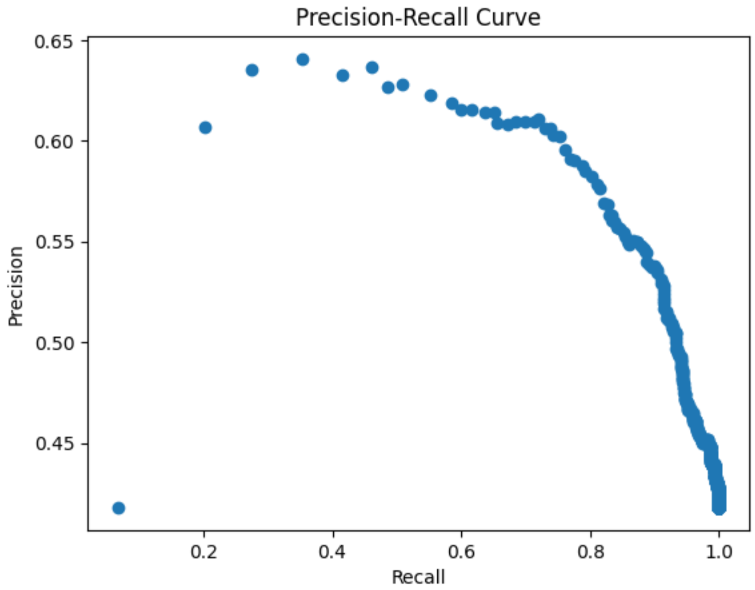
Results

- Train Set:
  - Precision: 0.5816102067751869
  - Recall: 0.7637203928365107
  - F-score: 0.6603396603396602
- Dev Set:
  - Precision: 0.5784982935153583
  - Recall: 0.8110047846889952
  - F-score: 0.6752988047808764

Threshold Analysis

- Range of thresholds tested: From 1 to 47,376,829,651. It was tested by 500,000 step-length with total tested 94,754 thresholds
- Best threshold: 15,100,000

Precision-Recall Curve



P-R Curve Analysis

The range of precision and recall are always between 0 and 1. From the graph, as we can see, the recall could be very close to 1, but when the recall is close to 1 (identified all positive cases, where the complex words are all captured) and the precision is close to 0. In this way, the fscore value is small. The largest f-score comes from precision as 0.5784982935153583 and recall as 0.8110047846889952. The

Discussion

I used the logic that when the counts in ngram_counts is bigger than the threshold, it will be treated as simple words. If the counts are smaller than the threshold, it will be treated as complex words. The assumption is that simple words are more popular than complex words.

3.1 Naive Bayes (4 points)

Results

- Train Set:
  - Precision: 0.6840796019900498
  - Recall: 0.635470826112074
  - F-score: 0.6588799041629231
- Dev Set:
  - Precision: 0.7346278317152104
  - Recall: 0.5430622009569378
  - F-score: 0.624484181568088

3.2 Logistic Regression (4 points)

Results

- Train Set:
  - Precision: 0.6128888888888889
  - Recall: 0.7966493356441363
  - F-score: 0.6927907560914344
- Dev Set:
  - Precision: 0.6272727272727273
  - Recall: 0.8253588516746412
  - F-score: 0.712809917355372

3.3 Discussion on NB and LR (6 points)

Model Comparison

Overall, the logistic regression is better than the Naïve Bayes method. Because for precision, recall and f-score, the logistic regression is better than Navis Bayes.

Performance Analysis

Based on f-score, Naïve Bayes model with 2 features performs worse when just use one feature. This might because there are correlations between these two features.

4.1 Own model's performance (6 points)

Dev Set Performance

- Precision: 0.7053140096618358
- Recall: 0.6985645933014354
- F1-score: 0.701923076923077

4.2 Own model's analysis (15 points)

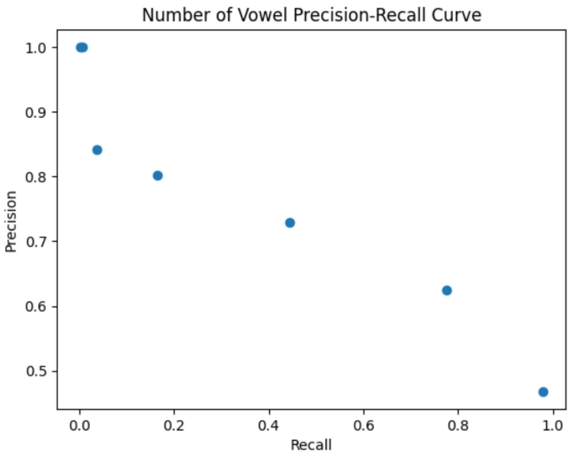
Model Description

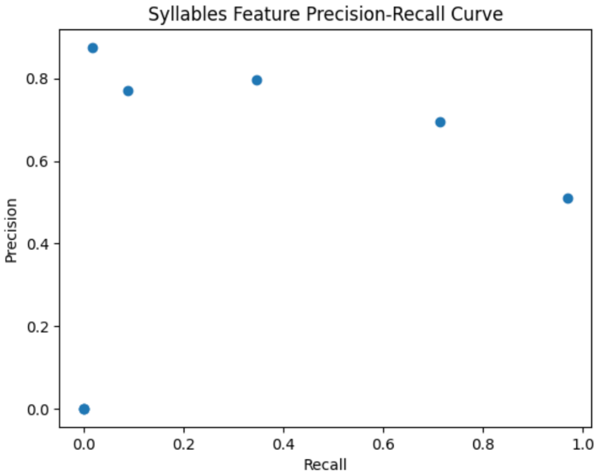
[Describe the model(s) you experimented with, and which one provided the best performance]

I used SVM(Support Vector Machine) regression model, which has 2 features: 1st feature is the number of vowels in the word. The 2nd feature is the number of syllables of the word.

Features Description

I select 2 features:

1. The number of vowels in the word: the best value to use is 2 which offers largest f-score for this feature only. Means if the word has more than 2 vowels, this word is a complex word.
    a. 

2. The number of syllables of the word: the best value to use is 2, which offers largest f-score for this feature.
    a. 

Feature Selection Justification

Feature 1 number of vowels: The reason to choose this feature because simple words usually have less vowels. For example: gate, poor are simple words, beautiful and magnificent are complex words. Based on the test, turns out the best threshold is 2.

Feature 2 the number of syllables: The reason is the complex words usually are longer and can have more meanings, thus it can have more syllables. Based on the test, turns out the best value is 2.

Performance Examples

- Examples of words the model classified correctly:
  - True Complex: photographer, shorebirds, hibernate, relating
  - True Simple: warnings, hands, bread, mixed
- Examples of words the model classified incorrectly:
  - False Complex: academy, continues, overthrown
  - False Simple: coup, lilting, steamy

Error Analysis

- Category 1: The words have short length and less than 2 vowels, but these words are complex. The words are complex, but they are not true for these 2 features. Basically, the feature cannot cover these cases.
  - Examples: coup, lilting, toughest
  - Potential reason: The features cannot correctly identify these words.
- Category 2: The 2 features, one is true and the other is false
  - Examples: academy, friction, furious
  - Potential reason: When the features are not consistent with each other, the model cannot perform correctly.