Caleb L'Italien

CSC-483 Project 2

02/02/2023

Table of Contents

I. ABSTRACT

This project focused on text classification, specifically on categorizing words as ‘simple’ or ‘complex’. Five different classification models were used on various input files, and their performance compared to classification done by human annotators. The project aims to show the trade-offs in evaluation metrics and the complexity of text classification. The models past the baseline classifiers are implemented with the sklearn library. The frequency distribution of various words are taken from the Google ngram corpus.

II. BASELINE TEXT CLASSIFIERS

The first classifier classified every word it considered to be complex. It performed on the training and development data and produced the measurements in Figure 1.

	Training Data (%)	Development Data (%)
Accuracy	42	44
Precision	42	44
Recall	100	100
F-score	59	61

Figure 1

The word length baseline classified words with a length equal to or greater to a certain threshold to be complex. After experimenting with thresholds between 0 and 15, the classifier produced the results in Figure 2.

	7 Characters (%)	8 Characters (%)
Training Set		
Accuracy	68	72
Precision	59	66
Recall	84	70
F-Score	69	68
Development Set		
Accuracy	70	76

Precision	61	72
Recall	86	74
F-Score	71	73

Figure 2

Either of these thresholds could serve as the best threshold. While each threshold has similar F-scores, the 7 character limit performed much higher in recall, while the 8 character limit performed slightly higher in accuracy and precision.

The word frequency baseline classified words with a frequency in the training and development sets equal to or greater than a certain threshold to be complex. After experimenting with thresholds between 0 and 100,000, the best threshold was found to be between 1883 - 1948 instances. The measurements for 1883 instances are shown in Figure 3.

	1883 Instances (%)
Training Set	
Accuracy	44
Precision	43
Recall	96
F-score	59
Development Set	
Accuracy	43
Precision	43
Recall	94
F-score	59

Figure 3

III. NAIVE BAYES AND LOGISTIC REGRESSION CLASSIFIERS

Next, Naive Bayes and Logistic Regression classifiers were used on the training and development sets. The classifiers used word length and word frequency as their features. Measurements for each classifier are shown in Figure 4.

	Naive Bayes (%)	Logistic Regression (%)
Training Set		
Accuracy	55	74
Precision	48	71
Recall	97	64
F-score	65	68
Development Set		
Accuracy	55	77
Precision	49	76
Recall	98	69
F-score	65	72

Figure 4

IV. DECISION TREE AND RANDOM FOREST CLASSIFIERS

Finally, Decision Tree and Random Forest classifiers were trained on the training sets. The classifiers used word length, word frequency, number of synonyms, number of syllables, number of common characters, and number of uncommon characters as their features. Common letters and uncommon letters are listed in Figure 5. The measurements for the development set are shown in Figure 6.

Common Characters	Uncommon Characters
e, a, r, i, o, t, n, s, l, c	j, q, x, z, -

Figure 5

	Decision Tree (%)	Random Forest (%)
Development Set		
Accuracy	70	76
Precision	67	74
Recall	62	71
F-score	64	72

Figure 6

Figure 7 shows a sample of words the Random Forest classifier incorrectly and correctly classified from the development set.

Incorrectly Classified	Correctly Classified
ate	shooter
seventh-grader	intense
argued	diving
triangle	pans
sun-splattered	hills

Figure 7

As can be noticed in Figure 7, the classifier consistently incorrectly classified words with hyphens. Furthermore, it incorrectly classified complex words with small character limits and common letters.