Isaac Penny
CS 229
Term Project
Final Report

1. Introduction

For the term project, I applied machine learning to text classification in ancient documents. In particular, I used a machine learning algorithm, trained on the Pauline epistles of the Bible's New Testament, to determine the probability that Paul also authored the Epistle to the Hebrews. Of the twenty seven books in the New Testament, only the Epistle to the Hebrews does not contain an explicit claim of authorship. However, tradition and the writings of several early church leaders indicate Paul as the book's author.

2. Approach

A Support Vector Machine, was chosen for the logistic classification process. An SVM was chosen for its "off-the-shelf" ease of use and its wide acceptance within the field of text classification[1]. The support vector machine further utilizes a simplified version of Sequential Minimization and Optimization algorithm[2]. See the referenced papers for more algorithm details.

2.1 The Text

All books were evaluated in Greek to avoid the affects of translation. The Greek text used is the Stephanus edition of the *Textus Receptus*, compiled in 1550 A.D. The text itself is in the public domain, however the project utilizes a proprietary version obtained from Hermeneutika Software, under an academic license[3]. The Hermeneutika version of the text also provides the Greek root word and morphology (part of speech, number, person, tense, mood, and voice) for each word in the text. An excerpt from the text showing John 3:16 ("For God so loved the world…") is shown below:

> John 3:16 Οὕτως οὕτω bo γὰρ γάρ c ἠγάπησεν ἀγαπάω viaa3s ὁ ὁ dnms θεὸς θεός nnms τὸν ὁ dams κόσμον κόσμος nams ὥστε ὥστε c τὸν ο῾ dams υἱὸν υἱός nams αὐτοῦ αὐτός rpgms τὸν ο῾ dams μονογενῆ μονογενής aamsn ἔδωκεν δίδωμι viaa3s ἵνα ἵνα c πᾶς πᾶς anmsn ὁ ὁ dnms πιστεύων πιστεύω vppanms εἰς εἰς p αὐτὸν αὐτός rpams μὴ μή xo ἀπόληται ἀπόλλυμι vsam3s ἀλλ ἀλλά c ἔχῃ ἔχω vspa3s ζωὴν ζωή nafs αἰώνιον αἰώνιος aafsn

Each word above appears in triplet. The first word is the original Greek word. The second word is the Greek root word. The third word in each triplet is the morphology of the Greek word (ex: vsam3s means a **v**erb with **s**ubjunctive mood, **a**orist tense, **m**iddle voice, which is **3**rd person and **s**ingular in number).

2.2 Training Examples

Each book in the New Testament is divided up into chapter and verse divisions, by scholars to aid in easy referencing. Individual verses from each book were used as training examples. Positive training examples were provided by the 13 Pauline books in the New Testament. Negative training examples were provided by the 13 non-Pauline books in the New Testament.

2.3 Parsing the Text

The data needed to be extracted from its plain text format and stored in a useful data structure, before it could be used for classification. The data was extracted using a simple text string search, where spaces were treated as delimiters between words. The data was then parsed into a five dimensional cell array with the following dimensions:

1. Book number (1 = Matthew, 2 = Mark, etc.)
2. Chapter number
3. Verse number
4. Word number within the verse
5. String type (1 = inflected word, 2 = uninflected word, 3 = morphology)

This structure maintains all of the original relationships between the data, while making it easy to extract the desired from of a particular word from the text.

2.4 Feature Selection

N-gram frequencies were used as input features for each training example. The density of the data in the feature space of n-grams of size two and higher was deemed too sparse to be useful, thus only unigrams were used.

Only root (uninflected) unigrams were used for classification. This approach results in a smaller number of features, than if all of the inflected forms of a given word were used. The smaller number of features results in a less sparse set of training data. The more dense training set helps the classifier generalize better to test sets where the test data set has a predominantly different morphology than the training set. Ignoring the morphology in determining authorship assumes that the choice of root word (ex: play versus compete) is a more significant indicator of authorship than is the choice of morphology (ex: played versus have been playing).

2.5 Creating Dictionaries

A dictionary of all uninflected unigrams was created by scanning the five dimensional datastructure mentioned above. The frequency of occurance for each unigram was also recorded.

One of the main goals of the project was to quantify the effect of feature space size on classification. Thus dictionaries of various sizes were created. d_{100} is a dictionary composed of the one hundred unigrams that occur most frequently in the New Testament. Choosing the most frequent unigrams has two benefits. First, the feature space will be less sparse and therefore more useful for classification purposes. Also, frequency of use with common unigrams (ὁ **the**, *καί* **and**, etc.) is only slightly affected by a work's content. As such they are commonly used indicators of authorship[4]. Equation 1 shows an example of the aforementioned dictionary:

$$d_{100} = \begin{bmatrix} \text{ὁ} & 20392 \\ \text{καί} & 9267 \\ \dots & \dots \\ \text{θυμόω} & 1 \end{bmatrix} \in \mathbb{R}^{100\times 2} \tag{1}$$

2.6 Cross-validation

K-fold cross-validation was used to explore the effect of feature space size on classification error. In *K*-fold cross-validation, the original training data set is divided into *K* subsets. Of the *K* subsets, a single subset is retained as test data, while the remaining $K - 1$ subsets are used as training data. The cross-validation process is repeated *K* times, with each of the *K* subsets being used exactly once as the test data set. Cross-validation error is then the mean classification error among the K repetitions[5].

3. Results

3.1 Effect of Feature Space Size on Cross-validation

The training data set was divided into ten subsets for cross-validation purposes. Figure 3.1 shows the effect of feature space size on cross-validation error.

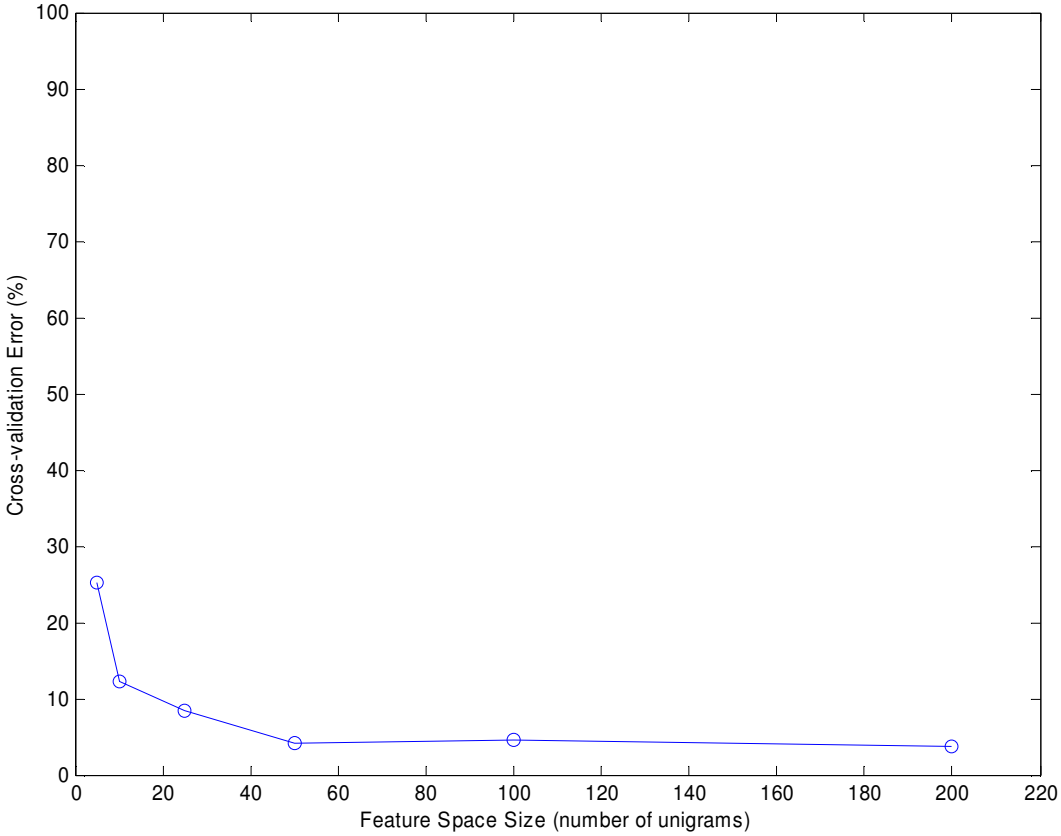

Figure 3.1. Effect of feature space size on cross-validation error.

As seen in the figure, using a dictionary size of more than fifty unigrams does not significantly reduce classification error.

3.2 Percent Verses Classification

Each verse in the book of Hebrews was individually classified as pauline or nonpauline. This process was repeated using each dictionary. The process also repeated for several other books believed to be either pauline or nonpauline. Figure 3.2 shows the resulting percentage of verses classified as Pauline.

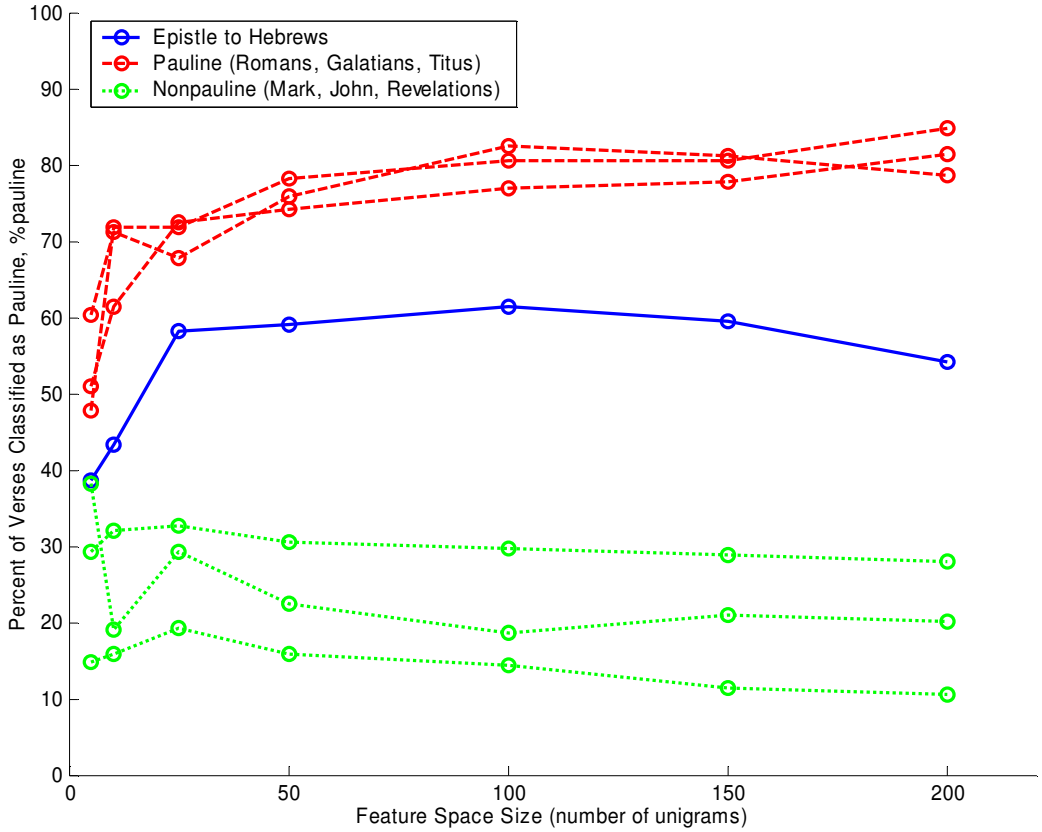

Figure 3.2. Percentage of verses in Hebrews and several other NT books classified as Pauline.

As expected given the cross-validation study, a feature space size (dictionary size) of greater than fifty unigrams does not further separate the various book categories. Also, it should be noted that the variance in the nonpauline category is significantly larger than that in the Pauline category. This is to be expected, as there is a lurking variable of multiple authors within the nonpauline category. The resulting analysis shows that for dictionary sizes over fifty unigrams, the mean percent of verses in Hebrews classified as Pauline is 58%.

3.3 Statistical Significance of Results

The classification of verses from Hebrews certainly appears closer to that of the pauline books than that of nonpauline books. The question becomes whether this difference is statistically significant. Using the central limit theorem, we hypothesize that the percentage of verses classified as pauline for a book will itself be distributed normally about the mean value for that book's category. Thus we can use a standard normal z-test to calculate the probability that Hebrews is in the Pauline and non Pauline categories. For a feature space of size fifty, the pauline and nonpauline z-scores were calculated. A z-score measures the distance of a data point from a category's mean in units of the category's standard deviation. The standard normal distribution can then be used to find the probability that the variation within each category can explain the datapoint's departure from the category mean. This calculated probability is the probability that the data point belongs to the category. The z-scores and probabilities are summarized in Table 3.3.

Table 3.3. Results of Z significance test using d_{50}.

Category	Z-score $= \frac{\%_{pauline} - \mu_{category}}{\sigma_{category}}$	$P(\text{Hebrews in Category})$
Pauline	5.4	3.8×10^{-8}
Nonpauline	6.7	1.1×10^{-11}

4. Conclusions

The z-test indicates that Hebrews as a whole is more likely to be in the pauline category than it is to be in the nonpauline category. The low probability that Hebrews belongs to either the pauline or nonpauline categories might also suggest that Hebrews was written by a mystery author whose writings are not otherwise included in the New Testament. However, since we do not have training data for the mystery author, one cannot evaluate such a hypothesis using the current approach.

5. Future Work

This conclusion is premature. Multiple authors were lumped into the nonpauline category. Thus it could be that the variation within the writings of an *individual* nonpauline author is high enough to account for Hebrews' deviation from that author's mean classification score. However, training individual classifiers for each of the New Testament authors has the downside of a lack-of-training-data problem. This approach is suggested for future work.

This project only considers the authorship of Hebrews as a complete unit. Future work might statistically analyze the distribution of Pauline classified verses within Hebrews to determine if certain sections of the book are more or less likely to have been written by Paul.

6. Bibliography

1. Platt, John. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods – Support Vector Learning*. MIT Press: Boston.

2. Boser, Guyon, Vapnik. 1992. A training algorithm for optimal margin classifiers. *5th Annual ACM Workshop on COLT*. ACM Press: Pittsburgh. pp144-152.

3. Erasmus, Desiderius. 1550. Textus Receptus. Public Domain. Made available by Hermeneutika Software. www.bibleworks.com Accessed November 20, 2007.

4. Various. 2007. Context-free grammar. *Wikipedia Online Encyclopedia*. Accessed at http://en.wikipedia.org/wiki/Context-free_grammar on December 4, 2007.

5. Various. 2007. Cross-Validation. *Wikipedia Online Encyclopedia*. Accessed at http://en.wikipedia.org/wiki/Cross_validation on December 5, 2007.