Feature Selection and Representation for Textual Language Recognition

Christopher Jeschke cjeschk2@jhu.edu
Engineering for Professionals
Johns Hopkins University
Elkridge, MD 20175, USA

Editor: Christopher Jeschke

Abstract

This research project report describes a series of text classification experiments using Naïve Bayes and Support Vector Machine classifiers to identify the language of an article based on its salient textual features - namely words and word n-grams. Paragraphs of text in Germanic and Romance languages were extracted from a corpus of Wikipedia articles as a training and test data set, against which various feature selection and representation techniques were applied to develop intuition into modeling text classification problems and understanding the difficulties in separating lexically similar languages.

Keywords: Feature Selection, Language Classification, SVM, Naïve Bayes

1. Introduction

The automated extraction of information from text and its categorical classification is a well studied area within the fields of Information Retrieval and Machine Learning (ML). The ability to process unstructured, so called "free" text for useful information powers several tools and information platforms we generally take for granted: key-word based search engines (e.g. Google), content filters, and even text-to-speech applications.

A specific application of interest is the identification of the language that a segment of free text was written in. Human languages have evolved over time to share portions of syntax (e.g. rules regarding word order), semantics and large portions of vocabulary, prompting the question of what textual components of a language allow it to differentiated from others. Complicating this are lexically similar languages - such as French and Portuguese - or languages with considerable vocabulary overlap, like English and German.

2. Statement of Problem

How do we identify a written language? Intuitively, literate individuals can identify languages in which they are fluent by recognition of the structure, words and other syntactical signatures present in the text. Translating this to a problem of machine learning introduces interesting challenges in feature identification and representation. The structure of text is at odds with the expectations of most classification algorithms. A sample of free text is often of variable length and largely non-numeric, hardly comparable to the fixed length vectors of floats or integers consumable as training samples in most ML classifiers. The

naive approach of assigning unique numeric identifiers to each individual word in the text as an index into a vector is plausible, but then we must consider what to use for the value at that index and the impact the length of the text has on the resulting feature space. Simply observing the sentence proceeding this one, we can identify 42 distinct word features for what amounts to roughly 4 lines of text. Learning a classifier from a large corpus of text could entail hundreds of thousands of candidate word features, testing the scalability of a classification technique.

This research project develops understanding of feature identification and representation challenges in text classification. Specifically, the questions this project aim to answer are two fold:

1. What textual features are useful or relevant for a textual classification problem such as identifying a language?

2. How should those features be represented for a classification technique? And related to this, which classification algorithms drawn from Machine Learning are preferred?

Our choice of textual features to extract can vary considerably. We can assume a single feature per unique word or limit ourselves to words meeting certain criteria, such as minimum counts or frequency. The extracted features need not be single words either. Word order may have implications. A statement such as "is a" might have more discriminatory value than "is" and "a" by themselves. The field of Computational Linguistics introduced the concept of *n-grams* for describing a sequence of n letters, words, symbols and other component decompositions of a language as a means to extract additional contextual information from text or audio, opening up even more feature identification and extraction opportunities.

Representation of features for classification also requires attention. Binary indicators are one option, but convey no information beyond the presence or absence of the feature. Other options are counting the number of occurrences of the feature within the sample, or calculating its frequency with respect to the rest of the words present in the sample. Term-frequency inverse-document-frequency (TF-IDF) is a popular numerical statistic to represent the importance of a word in a corpus for problems in information retrieval, scoring higher those words that occur frequently within a limited number of documents but infrequently within the rest of the corpus, thereby filtering out common words such as "a", "the" or "and" when used with English text. The simpler metric term-frequency (TF) is evaluated as part this study. Common words are actually useful in language recognition, ensuring some reasonable subset of features is identifiable in each training and test sample.

Language classification introduces some interesting nuances. This study discusses lexically similar languages, with similarity defined as the degree of overlap in vocabulary between languages (Ethnologue). As English is the language this report is authored in and is presumably understood by the intended reader, it has been chosen as one of the language categories, along with those languages from which the largest portions of its vocabulary are drawn - German (60%) and French (27%) - as identified by Williams (1975) (re-published in Wikipedia: English Language). Williams also identified a size-able portion of the English vocabulary came from Latin (29%), so this study has also chosen the modern Romance languages of Spanish, Portuguese and Italian - in addition to French - for their high degree

Table 1: Lexical Similarities

	English	French	German	Italian	Portuguese	Spanish
English	1	.27	.60	-	-	-
French	.27	1	.29	.89	.75	.75
German	.60	.29	1	-	-	-
Italian	-	.89	-	1	-	.82
Portuguese	-	.75	-	-	1	.89
Spanish	-	.75	-	.82	.89	1

of lexical similarity and historical relationship with Latin. Table 1 shows a summary of lexical similarity figures as supplied by Ethnologue.

3. Approach

The experimental approach is an empirical study using a multi-lingual corpus constructed from Wikipedia articles written in the following languages: English, German, French, Spanish, Portuguese and Italian. A raw corpus of Wikipedia articles was extracted from a set of comparable corpora made publicly available by LinguaTools (Linguatools), as compiled from Wikipedia articles extracted in 2014 and made available under the Creative Commons Attribution-ShareAlike license. A series of n-gram based feature extraction methods were applied to training data assembled from combinations of languages in the corpus, after which Support Vector Machine (SVM) and Naïve Bayes classifiers were trained for each language and then evaluated for Accuracy and area under the ROC curve (AUC). The n-grams were typically unigrams (1 word) and bigrams (2 words) unless otherwise stated, and the terms feature and n-gram may be used interchangeably throughout this report. The platform of choice throughout this effort was Python 2.7.12, using the packages BeautifulSoup 4 and Scikit-Learn 0.19.1, running within Jupyter notebook 4.2.1 instance.

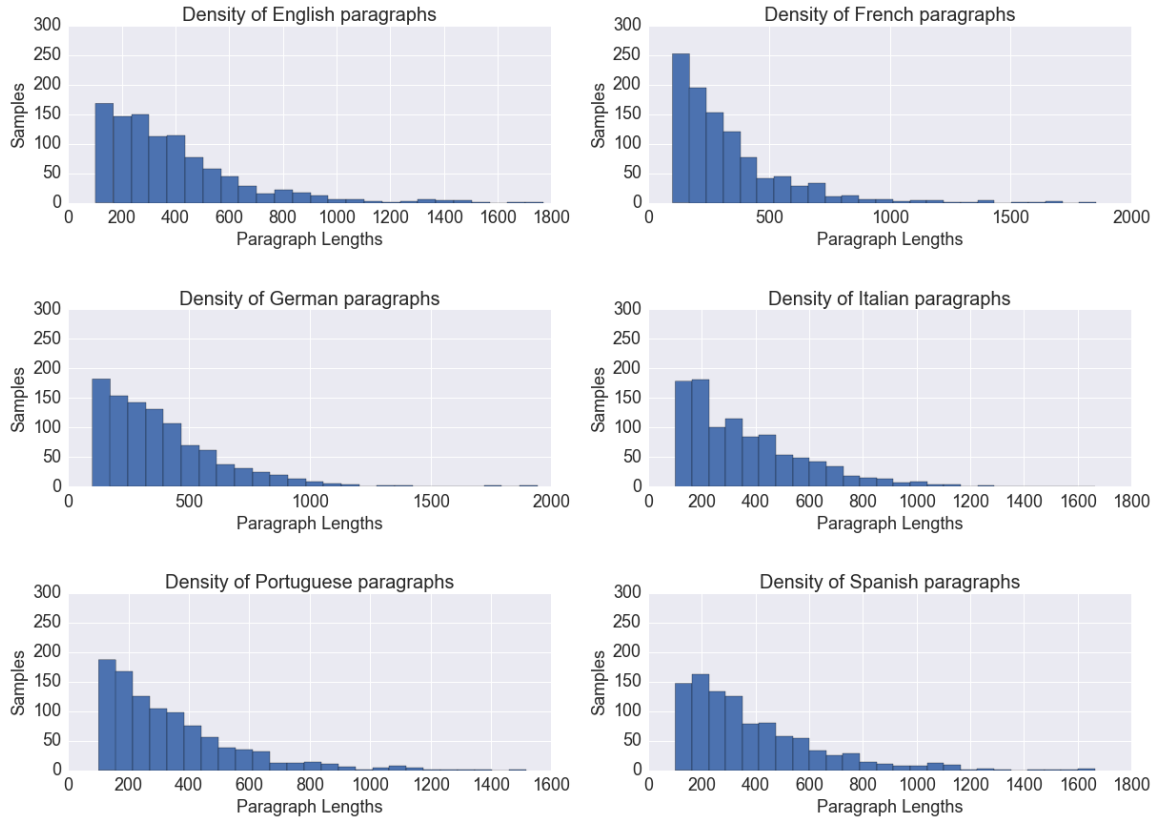
3.1 Assembling the Corpus

The aforementioned Wikipedia articles were serialized as XML documents that readily identified the language of the article and its content. For each language of interest, the content section was parsed for paragraphs identified by the opening and closing HTML tags: `<p>` and `</p>`. Paragraphs were stripped of any embedded HTML tags used for formatting, though anchor tags were given special treatment. The link text present between the opening `<a href>` and closing `<a>` tags was often an integral part of the sentence and was therefore retained. For example, the paragraph `<p>` **The cat in the** `<a href=http://hats.com>` **hat.** `</a>` would be parsed to **The cat in the hat.**.

A total of 1000 paragraphs were extracted for each language, ranging in length from 100 to 2000 characters. The distributions of sample length by language are displayed in Figure 1 for reference.

Admittedly, shorter French messages are over-expressed compared to the other languages, but for purposes of this study the distribution in sample size was felt sufficiently similar to not require further re-balancing each corpus. The lexical similarity between each

Figure 1: Language Sample Sizes



language in the corpus recorded in Table 2 to determine if the figures from Ethnologue held true.

Table 2: Unique Words & Lexical Similarity (Observed)

Language	# Unique Words	English	French	German	Italian	Portuguese	Spanish
English	9,951	-	.08	.05	.05	.05	.06
French	10,947	.08	-	.03	.04	.04	.03
German	13,607	.05	.03	-	.03	.03	.05
Italian	12,363	.05	.04	.03	-	.07	.06
Portuguese	11,556	.05	.04	.03	.07	-	.11
Spanish	12,283	.06	.04	.05	.06	.11	-

Table 2 clearly shows the degree of similarity is significantly lower than described in Table 1, though this does not tell us the frequency with which words in appearing in multiple vocabularies are actually expressed in the corpus and is therefore not a fair measure of word feature overlap. We are also not necessarily seeing the entire vocabulary for each language. There is some confirmation of higher similarity in certain Romance language pairs, such as (Spanish,Portuguese) at .11 and (Portuguese,Italian) at .07. French and German are quite dissimilar at .03.

3.2 Features drawn from combined corpus

An initial experiment was run to provide a baseline with unigram and bigram word features drawn from a combined corpus of English, French, German and Spanish using the same restriction imposed by the lexical similarity comparison: a minimum single alphabetical character separated by white space or punctuation. This allows us to contrast identifying features from a language specific corpus, as in an applied setting a classifier for a specific language may not have access to corpora covering all other languages.

A feature sampling corpus of 10% of the paragraphs was drawn uniformly and at random from each language category. The restriction to 10% was to avoid over-fitting the feature space to that of the entire corpus, and is held constant throughout the remaining studies. The top-n, $n \in \{1, 2, ..., 24, 25\}$ word ngrams were chosen from the feature corpus by their sample frequency. SVM and Naïve Bayes classifiers were trained and evaluated for each language using their default hyper-parameters from the scikit-learn package. The SVM used a regularization parameter of $C = 1$ and *radial basis function* for its kernel with $\gamma = \frac{1}{\#\ of\ features}$. The Naïve Bayes classifier used *Laplace* smoothing to account for features absent from the language in question or the *other* category into which the remaining languages fell. Classifier training and evaluation occurred for each language as follows:

1. Construct a training/testing corpus with two categories: the language being classified and *other*. Example: English versus Other (German, French, Spanish).

2. For a set of features in n, apply a feature transformation to the corpus samples to convert each sample to a vector of feature *counts*. Each index in the vector maps to a selected word feature, and the value at that index is the number of times that word appeared in the sample.

3. Execute a 10-fold Cross Validation run for the SVM and Naïve Bayes classifiers against the training/testing corpus, recording the mean accuracy and ROC AUC when evaluating the *test* fold, along with their 95% confidence interval bounds.

4. Repeat the above for all n and for each language.

3.3 Language Specific Features and Feature Representation

The methodology laid out above was then used for evaluating language specific feature extraction and representation. The first follow-on experiment restricted the top-n feature selection to only a language specific corpus. For instance, when training an English classifier, only the English corpus was used to select the top-n most prevalent unigrams and bigrams. The features were still represented as counts in the vector form of each sample. The same experiment was then run using binary indicators for feature presence rather than counts evaluate which representation performed better.

A fourth experiment was run using intra-document frequency as the feature representation value, thereby incorporating the length of the sample into the metric. Finally, a fifth experiment occurred to evaluate the relevance of word order by *requiring* high frequency word bigrams only. No unigrams permitted.

3.4 Lexically Similar Languages

Throughout the previous experiments a pattern continued to emerge where the Romance languages Spanish and French required measurably more features to successfully classify as compared to English and German. A second series of experiments was executed using a corpus assembled from Spanish, French, Italian and Portuguese paragraphs using a subset of the tests described previously, drawing features from a corpus using all 4 languages, then language specific corpora with binary feature representation only.

A follow on experiment was executed using a chi-square test for feature significance between the language being classified and the other languages in the corpus. This was done to provide perspective on what optimal classification performance might look like if feature extraction took class labels into account. Finally, an evaluation was done using *character* n-grams of 2-5 characters in length, but as the results were considerably worse than word n-grams these were not published for brevity.

4. Background and Related Work

The original inspiration for this study was drawn from the challenges Chee (2011) faced in distinguishing English and non-English messages from one another when using Yahoo's online health forums for drug safety signal detection. Chee's study required distilling a forum corpus to English only messages, as the vocabulary for describing adverse events in drug safety - Medra - is in English. A simple vocabulary based technique using English words drawn from OpenOffice.org is used to count instances of English and non-English words in each message, then scored English/non-English based on their prevalence. However, Chee does allude to common words and n-gram based techniques as possible alternatives.

Dunning (1994) discusses common words, supporting this technique when enough text is present that *closed class* words have an opportunity to appear. These words provide structure to text, as they serve to join other less common words and phrases. Examples of closed class words in English would be *and, or, this, that* and we do see various closed class words appear as features selected in this study. Dunning highlight's the limitations of this approach by constructing short text fragments that are clearly English (example: "man immunodeficiency") but lack common words. Dunning's approach treats language as a Markov Decision Process, where the probability of a word or phrase in a specific language is dependent on those preceding it. This study remains focused on the admittedly simpler task of n-gram discovery based on frequency in the corpora.

Pang and Lee (2008) deal with the topic of textual feature representation in their study on opinion mining and sentiment analysis, citing preference to binary representations of word features as preferable to term-frequency, inverse document frequency (TF-IDF). These statements motivated contrasting binary and term-frequency representations in this study. Forman (2003) evaluates multiple feature selection metrics used in text classification for identifying features that differentiate between 2 or more classes: Chi-Squared, Information Gain, Bi-Normal Separation (BNS), etc. BNS yielded the best performance. Unfortunately no BNS implementation was available in the sci-kit learn packages. Chi-Squared was opted for comparison as an alternative to BSN and the mutual information approaches seen in other studies (Sahami et al., 1998).

4.1 Naïve Bayes Classifiers

Naïve Bayes classifiers see considerable use in text classification problems. Sahami et al. (1998) evaluate Naïve Bayes classifiers for SPAM recognition via message word features, features for specific phrases, and non-word features derived from email headers, approaching near 100% precision. Sahami uses the mutual information scores of word features to select those words most relevant to the SPAM/non-SPAM class labels.

The assumption in Naïve Bayes classification is that all features (word n-grams) are independently distributed within the classes of interest. The classification technique evaluates the *conditional probability* $p(C|D)$ - the probability of the class C given the document D. This study trains a single language classifier, giving us two classes: $C_{language}$ and C_{other}. The probability of the document given the classification is represented as $p(D|C) = \prod_i p(w_i|C)$, where w_i represents the ith word in a given document. We construct a classifier by selecting the class having the highest probability:

$$\hat{y} = argmax_{k \in (language, other)} p(C_k) \prod_{i=1}^{n} p(w_i|C_k)$$

4.2 Support Vector Machines

Support Vector machines work by discovering the maximum margin hyperplane(s) separating the feature space of 2 or more classes. They were originally formulated by Vapnik and discussed at length in Vapnik and Cortes (1995). The term *support vector* describes any training sample found that functions as a boundary of the hyperplane(s), to which test samples are compared when generating a prediction. Joachims (1998) explains why SVMs are expected to work well in text classification, highlighting their ability to scale well given large, sparse feature spaces derived from text. Joachims compares SVM text classifiers using RBF and polynomial kernels to Naïve Bayes, KNN, and C4.5, showing consistently higher precision/recall break even points when used with two popular test corpora.

The SVM applied in this study uses the Radial Basis Function for a kernel. The classifier can be formulated as:

$$y_i(w \cdot \Phi x_i + b) \geq 1 - \xi_i$$

where y_i is a class label $y \in \{-1, 1\}$ for sample x_i. ξ_i is a misclassification error for sample x_i, and w, b are the weights and bias to be learned by the SVM. The kernel function Φ projects x_i into a new dimensional space according to the RBF definition $\exp(-\gamma||x_i - x_j||^2),\ \gamma > 0$. γ may be specified when training the SVM. Not shown in this formulation is the regularization parameter C which is adjusted to regulate accumulated classification error $C \sum_{i=1}^{n} \xi_i$.

5. Results

Experimental results are published in the subsequent subsections with brief summaries. Further discussion and analysis of the results is reserved for section 6.

5.1 Features drawn from combined corpora

The first experiment using features selected from the union of the French, English, German and Spanish corpora showed considerably different improvement rates in accuracy and

AUC for Germanic languages (German and English) as opposed to Romance (Spanish and French). Figure 2 shows the rate at which the Accuracy and AUC improved English and French as the number of top-n features extracted from the combined corpora. The German and Spanish plots (omitted for brevity) were comparable to English and French respectively. Table 3 shows the number of features necessary to cross a threshold of .95 for accuracy and AUC on the lower bound of the 95% confidence interval for these metrics. The top 25 word features by frequency in the corpora are presented in Table 4.

Figure 2: Unigram and Bigram word Features from all Corpora

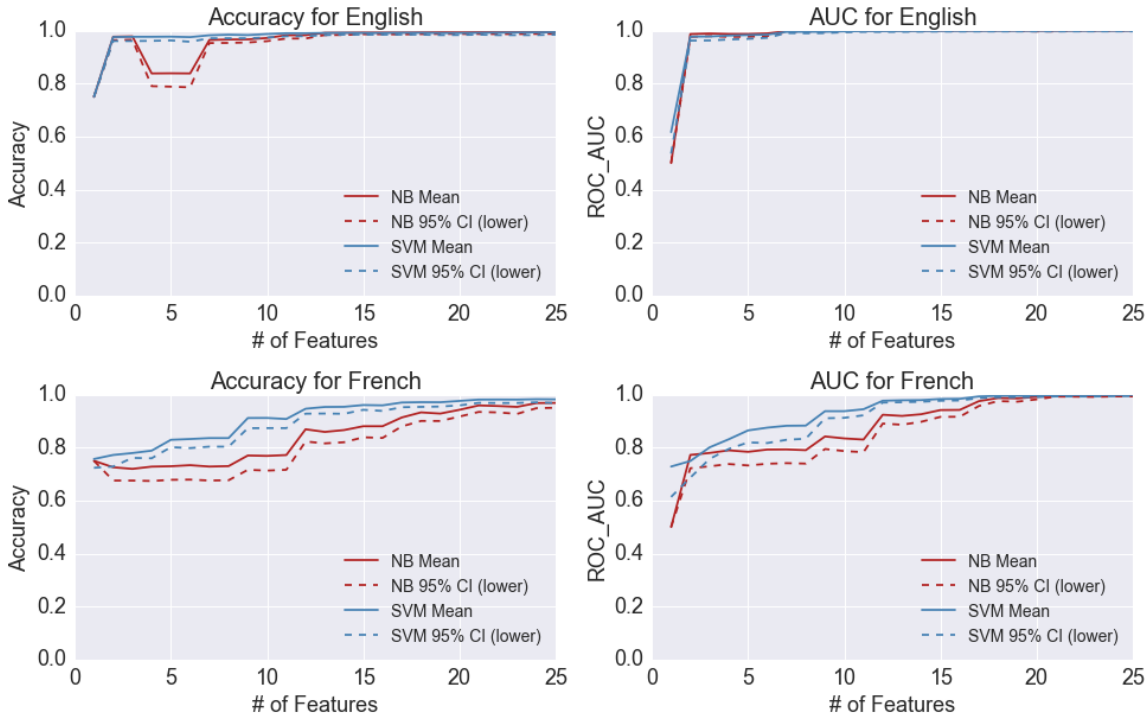


Table 3: # of Features when drawn from all corpora

Language	Accuracy > .95 (Lower Bound)		AUC > .95 (Lower Bound)	
	Naïve Bayes	SVM	Naïve Bayes	SVM
English	2	2	2	2
French	25	17	17	12
German	8	7	7	5
Spanish	21	15	15	12

5.2 Language Specific Features, Feature Representations

Selecting word features specific to the language being classified shows a marked change in the number of features required to classify French and Spanish, with a slight reduction for German and increase for English. Table 5 tabulates the number of top-n most frequent

Table 4: Baseline: Top 25 word features (all copora)

Rank	Words				
1-5	de	the	la	in	en
6-10	a	der	of	y	and
11-15	die	el	und	un	des
16-20	que	le	l	to	á
20-25	et	is	es	les	se

unigram and bigram word features required for crossing the lower bound threshold for each language, representation (language$_{representation}$) and classifier combination. The top 10 terms for each language are presented in Table 6.

Table 5: Top-N Features by Feature Representation

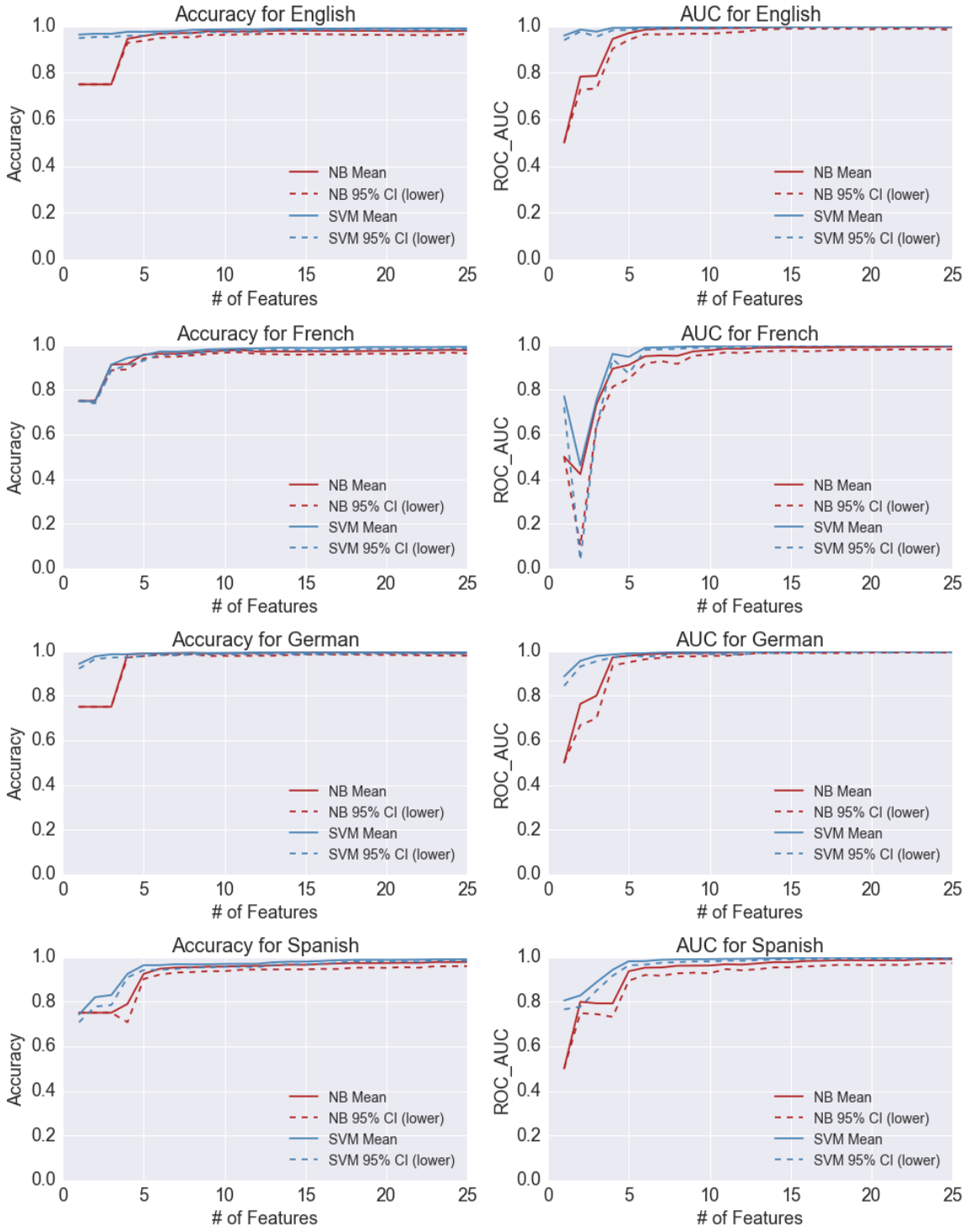
$Language_{representation}$	Accuracy > .95 (Lower Bound)		AUC > .95 (Lower Bound)	
	Naïve Bayes	SVM	Naïve Bayes	SVM
English$_{counts}$	7	3	6	3
French$_{counts}$	7	7	8	7
German$_{counts}$	4	2	5	5
Spanish$_{counts}$	15	10	13	5
English$_{binary}$	7	2	6	2
French$_{binary}$	8	6	9	6
German$_{binary}$	4	3	5	3
Spanish$_{binary}$	18	8	14	5
English$_{tf}$	10	4	6	3
French$_{tf}$	17	8	13	5
German$_{tf}$	4	2	5	3
Spanish$_{tf}$	25	16	14	5
All (Bigrams only)$_{counts}$	>25	>25	>25	>25

Table 6: Top 10 word features by language

Language	Word Rank									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
English	the	in	of	and	a	to	is	was	as	for
French	de	la	le	et	l	á	les	en	du	est
German	der	und	die	in	das	den	von	im	ist	mit
Spanish	de	la	en	y	el	que	a	del	las	una

Figure 3 shows the classifier (SVM) and feature representation (binary) combination achieving the .95 mark for each performance metric in the lowest number of word features. Naïve Bayes is plotted for comparison purposes.

Figure 3: Unigram & Bigram Word Features - Binary Representation



5.3 Lexically Similar Languages

Using the results with the English, French, German and Spanish corpora, experiments with the lexically similar languages were limited to only the SVM classifier with a binary feature representation. Table 7 shows the results. The "(2-char min)" descriptor denotes tests where the minimum length of the words in unigrams and bigrams was at least 2 characters, to evaluate if removing single letter words improved the feature space. Figure 4 contains the plots for this particular case, as this configuration showed the lowest number of word features required to cross the .95 threshold. Requiring bigrams and trigrams was tested as well. A final experiment using the top-n features found via a chi-squared ($\tilde{\chi^2}$) separation evaluation is reported at the bottom of Table 7.

Table 7: Top-N Features for Lexically Similar Romance Languages

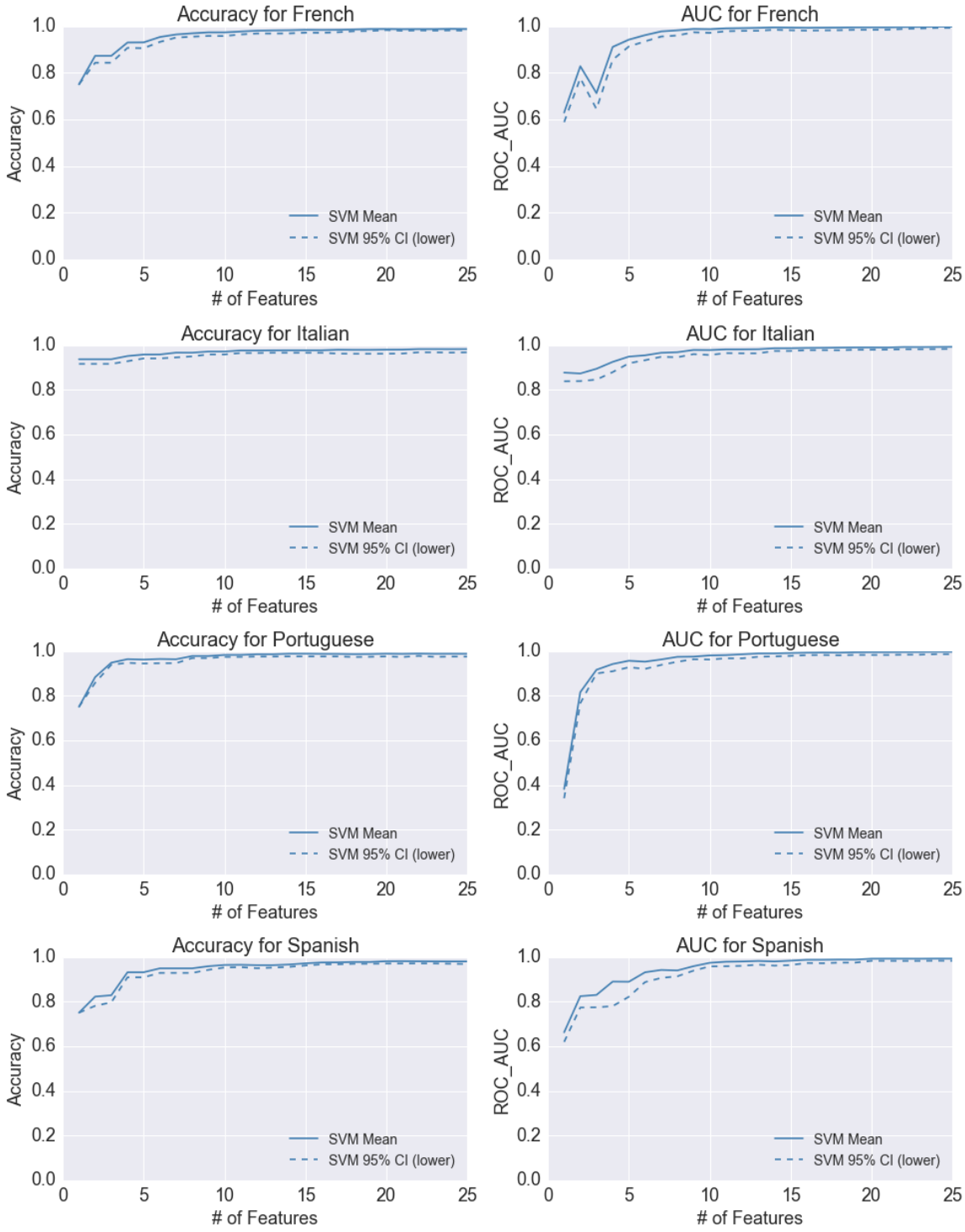
Language$_{representation}$	Accuracy > .95 (Lower Bound)	AUC > .95 (Lower Bound)
	SVM	SVM
French$_{binary}$	8	8
Italian$_{binary}$	7	10
Portuguese$_{binary}$	7	7
Spanish$_{binary}$	17	11
French (2-char min)$_{binary}$	7	7
Italian (2-char min)$_{binary}$	8	9
Portuguese (2-char min)$_{binary}$	8	8
Spanish (2-char min)$_{binary}$	10	10
All(Bigrams)$_{binary}$	> 25	> 25
All(Trigrams)$_{binary}$	> 25	> 25
French($\tilde{\chi^2}$)$_{binary}$	3	6
Italian($\tilde{\chi^2}$)$_{binary}$	4	5
Portuguese($\tilde{\chi^2}$)$_{binary}$	3	5
Spanish($\tilde{\chi^2}$)$_{binary}$	5	4

6. Discussion and Analysis of Results

6.1 Analysis of Features drawn from all Corpora

The first experiment shows the challenge inherent in how to use a corpus appropriately to identify textual features for classification. We can see that reaching the accuracy and SVM cutoff of .95 is achieved very quickly per Figure 2 with English, within 2 words: *de* and *the* per the word ranks in Table 4. *The* is only present in the English vocabulary and we know the absence of *de* would rule out Spanish and French. German requires slightly more words per Table 3. Spanish and French require considerably more, lending some support to the challenges expected in differentiating languages considered lexically similar per Table 1. We can also immediately observe the benefits of the SVM classifier over Naïve Bayes, reducing the number of features by a measurable margin for French and Spanish. This could be attributed to the RBF kernel in the SVM's ability to model relationships between multiple

Figure 4: Top-N 2+char word n-grams, Lexically Similar Romance Languages



terms (features) when projecting them into a higher dimension feature space, whereas the Naïve Bayes classifier assumes independence between all features.

We have to consider the plausibility of identifying word features in this fashion. Clearly having counter example languages in the corpus lends informative word features, but we also risk assembling a feature space that favors words with high relative popularity in their language. There is also the practical aspect of training a language classifier. Should such a classifier really be trained to incorporate word features pulled from counter examples? Expecting to have access to samples for every non-English language when training an English classifier seems an inconvenient requirement.

6.2 Analysis of Language Specific Features & Representation

We can immediately see the more intuitive word feature selections for each language per table 6, favoring the most common words of each language. Despite feature selection permitting both unigram and bigram word features, no bigram features appeared in the top-25 for these experiments. The forced use of bigrams present in the bottom row of table 6 shows a dramatic increase in the number of features required, as presumably not all bigrams were present in all paragraphs for the language being modeled. This is unfortunate as it was hypothesized that bigrams would be stronger features for differentiating languages, but the relatively short length of the paragraphs in the corpora may limit the opportunity for even the most common bigrams to show up as features in a training or test sample.

We observe a reasonable reduction in the number of terms required for classifying French, German and Spanish when using the *count* and *binary* feature representations per Table 5, thanks to the focus on word features derived from the specific language and elimination of more frequent terms from other languages. Recapping, the *count* feature representation is a literal count of occurrences in the sample being classified, whereas *binary* was just an indicator value (0,1) that the word feature was present. There appears to be a marginal benefit to using the binary representation with the SVM, though admittedly it is only a reduction of 1-2 features. The use of the *term frequency (tf)* representation was disappointing, increasing the number of features required for Spanish, French and English for both classifier types. As term frequencies would be in the range 0.0-1.0, they may conflict with the assumption the Naïve Bayes classifier makes regarding the use of integers only.

The best performing combination of SVM and binary feature representation was noted from these experiments and used for the follow-on study focused on lexically similar romance languages.

6.3 Separating Lexically Similar Languages

First, per Table 7 we can observe the new training and testing corpora of French, Italian, Portuguese and Spanish does has a negative effect on the ability to separate French and Spanish, increasing them 6 to 8 and 8 to 17 terms when using the binary representation in the SVM classifier. This is attributed to the higher similarity between Spanish, Portuguese and Italian as seen in Tables 1 and 2. Interestingly, a small change requiring that word features be at least 2 characters in length (*2-char min*) significantly reduces the number of features required for Spanish from 17 to 10, while having a negligible impact on French and Italian and Portuguese. Upon looking at the individual results for these experiments

more closely it was determined that while the choice of 95% as a target accuracy is useful as a cut-off for discussing a minimum number of features, fluctuation in the mean and 95% CI values of the performance metrics during cross-validation tests prevented crossing the .95 threshold (values of .940 to .946 were prevalent) despite selecting the identical set of 10 word features in both experiments: *con, de, de la, del, el, la, los, que, se*. The experiment requiring 2-char minimum lengths was re-run again for Spanish and the number of features required to cross the .95 threshold was reduced to 10, confirming this suspicion.

Experiments requiring bigrams and trigrams were run for comparative purposes, demonstrating once again the challenges in finding high frequency multi-word phrases in light of the smaller length text samples. The final experiment in Table 7 using the Chi-Squared was conducted to understand the ideal situation where the feature selection and training corpus for a language has counter examples in it for other languages and we opt to use a strong feature selection technique for class differentiation. Recall that Chi-Squared feature selection works by quantifying the probability that a particular feature is positively correlated with a with a particular class label (the language) or is otherwise uniformly distributed across all class labels (e.g. no correlation to the language). This approach would ensure the top-n features selected were those most closely correlated to the language being classified. As anticipated, there is a marked reduction in the number of features required for each language when meeting the accuracy and AUC thresholds. The most important terms chosen via this technique are in Table 8.

Table 8: $\tilde{\chi^2}$ Selected Features

Language	Word Rank					
	1st	2nd	3rd	4th	5th	6th
French	dans	des	du	est	et	une
Italian	che	di	in	per	è	-
Portuguese	com	do	em	um	uma	-
Spanish	el	es	las	los	y	-

7. Future Work

Extension points for this study are numerous. Character based n-grams for features were experimented with only briefly and tangentially in this study, and yielded unsatisfactory results, frequently requiring more than 25 features to cross the target performance threshold. However, little attention was given to how these n-grams were selected via tokenization of the corpus, as many contained punctuation or whitespace from word boundaries. A more in-depth focus to choose features based on common character patterns within the words of a language would be interesting.

Feature selection in conjunction with elimination is another readily apparent extension point. An ablation study to remove features after reaching a target performance metric could be done to better isolate the minimum feature set. Features selected in this study were done in a strictly additive sense - with each experimental run adding the next feature based on its frequency in the feature sampling corpus subsequent any tokenization restrictions. A

high frequency term may be shared across languages (ex: *de* in French and Spanish) and thus add little value in differentiating them.

The most challenging aspect to language classification would seem the assembly of a good feature selection corpus in light of the context in which any trained classifier might be applied. The emphasis in this study was placed on attempting to identify and usefully represent features derived only from a language specific corpus, with features pulled from a corpus including counter examples offered only to provide contrast and highlight the difficulty. Expanding the feature space beyond vocabulary to include other structural elements of the language, such as the presence/absence of word gender or certain punctuation might be useful. Finally, all text in this study was stored in UTF-8 encoding, making a unicode code point readily available for lookup to understand if the character represented is in Latin script (most languages in this study), Cyrillic, Hebrew or other languages that are clearly dissimilar.

References

B. Chee. *Exploring Machine Learning Techniques Using Patient Interactions in Online Health Forums to Classify Drug Safety.* Phd dissertation, University of Illinois at Urbana-Champaign, 2011.

T. Dunning. Statistical identification of language. Technical report, Computing Research Laboratory - New Mexico State University, March 1994.

Ethnologue. Lexical similarity. `https://en.wikipedia.org/wiki/Lexical_similarity`. Online, accessed 13-April-2018.

G. Forman. An extensive empirical study of feature selection metrics for text. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: EMCL-98*, pages 137 – 142, Berlin, Heidelberg, 1998. Spring Berlin Heidelberg. ISBN 978-3-540-69781-7.

Linguatools. Wikipedia comparable corpora. :http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/. Online, accessed 13-April-2018.

Medra. Medical dictionary for regulatory activities. `https://www.meddra.org/`. Online, accessed 13-April-2018.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. 1998.

V. Vapnik and C. Cortes. Support-vector networks. *Machine Learning*, 20:273 – 297, 1995.

J. M. Williams. *Origins of the English Language: A Social and Linguistic History.* The Free Press, New York, New York, 1975.