ASSIGNMENT 1: [COMP550]

JOSEPH D. VIVIANO

1. Ambiguity

1.1. **Phonology.** https://books.google.ca/books?id=eJUvDGXRW7AC&pg=PA92#v=onepage&q&f=false

"Mr Buyer, I understand that you do not have the money to write me a check for the retainer tonight. When would it be all **right** for you to **write** me a check for the retainer?"

Right and *write* sound identical in English, but have very different meanings. Indeed, *right* after the word "all" means something potentially different than *right* on it's own. Here, *right* means "correct", and adjective, whereas *write* mean "to use a hand-held utensil for the purposes of placing words on a medium such as paper", a verb. For a reader to disambiguate these words, one could look to the word before right, i.e., "all right", to see that right in this case means correct, where as one could look at the words directly after "write", i.e., "write me a check", to see that this word refers to the act of visual letter production.

1.2. **Morphology.** https://twitter.com/JVM/status/1020084244490670080

"We are **live** in New York City's Brooklyn Bridge Park at the high-profile photography exhibit The Fence 2018, an exciting photo series with an animal rights message starring rescued chickens."

Live by itself 'happening right now' or 'a place that I call home', and are written the same way although they sound differently when spoken. The reader must use the words immediately before it, "we are" to determine that this word "live" refers to "happening right now".

1.3. **Syntax.** http://journals.sagepub.com/doi/abs/10.1177/1744987109358836?journalCode=jrnb

"The aims of the present study were both to explore how the oldest of **old** men and women with estimated high resilience talk about experiences of becoming and being old, and to discuss the analysis of their narratives in terms of the foundational concepts of the Resilience Scale (RS)."

Date: Sept 2018.

Old in this case implies that the men are old, but it isn't clear whether this sentence applies only to old women, or all women with an estimated high resiliance. A non-ambiguous version of this sentence would be "oldest of old men and old women". In this case, the reader would have to look further to "talk about experiences of becoming and being old" to infer that all of the women being discussed here are old.

1.4. **Semantics.** `https://en.wikipedia.org/wiki/Domestication_of_the_Syrian_hamster#Capture_of_live_hamsters`

"The domestication of **the Syrian hamster** began in the late 1700s when naturalists cataloged the Syrian hamster, also known as Mesocricetus auratus or the golden hamster."

The in this case could refer to either a single hamster of the Syrian species, or the entire species of Syrian hamsters. A non-ambiguous version of the sentence would explicity use the word 'species' i.e., "The domestication of the Syrian hamster species began in the late 1700s". To disambiguate these meanings, the reader would have to have knowledge of the most probable use of "the Syrian hamster" in English, or knowledge from the sentences before (e.g., a particular hamster was introduced).

1.5. **Pragmatics.** `https://grassrootsmotorsports.com/forum/off-topic-discussion/i-hate-squirrels/33189/page2/`

"**I chased** a squirrel around the yard. Up and down the yard, back and forth, back and forth and the squirrel ran up a tree. So now the car's totaled. – Emo Phillips (a joke)."

I chased without context suggests a person persuing a squirrel around the yard on foot, because that is most likely, but in fact any method of pursuit could have been employed because no particular method was defined. In this case, the reader must continue to the punchline, which makes the meaning of the first sentence concrete, i.e., the person was in an automobile. This is a cause for levity in some.

2. FST for Spanish Verbal Conjugation

Infinitive	**1st sg**	**2nd sg**	**3rd sg**	**1st pl**	**2nd pl**	**3rd pl**
andar	and a:o r:ϵ	anda r:s	anda r:ϵ	anda r:m ϵ:o ϵ:s	and a:á r:i ϵ:s	anda r:n
contestar	contest a:o r:ϵ	contesta r:s	contesta r:ϵ	contesta r:m ϵ:o ϵ:s	contest a:á r:i ϵ:s	contesta r:n
beber	beb e:o r:ϵ	bebe r:s	bebe r:ϵ	bebe r:m ϵ:o ϵ:s	beb e:é r:i ϵ:s	bebe r:n
correr	corr e:o r:ϵ	corre r:s	corre r:ϵ	corre r:m ϵ:o ϵ:s	corr e:é r:i ϵ:s	corre r:n
vivir	viv i:o r:ϵ	viv i:e r:s	viv i:e r:ϵ	vivi r:m ϵ:o ϵ:s	viv i:í r:s	viv i:e r:n
recibir	recib i:o r:ϵ	recib i:e r:s	recib i:e r:ϵ	recibi r:m ϵ:o ϵ:s	recib i:í r:s	recib i:e r:n
infinitive	1-sg-irreg	2-sg-irreg	3-sg-irreg	1-pl-irreg	2-pl-irreg	3-pl-irreg
ser	s e:o r:y	s:e e:r r:e ϵ:s	s:e e:s r:ϵ	s e:o r:m ϵ:o ϵ:s	s e:o r:i ϵ:s	s e:o r:n
haber	h a:e b:ϵ e:ϵ r:ϵ	ha b:s e:ϵ r:ϵ	ha b:ϵ e:ϵ r:ϵ	h a:e b:m e:o r:s	hab e:é r:i ϵ:s	ha b:n e:ϵ r:ϵ

Please find the FST diagrams attached to the end of this document.

3. Sentiment Analysis

3.1. **Experiment Description.** To train our sentiment analysis classifier, I built a pipeline using Scikit Learn to compare various preprocessing strategies and the Naive Bayes (NB), Support Vector Machine with a Linear kernel (SVC), and Logistic Regression Classifiers (LR).

To preprocess the corpus, ngram counts were collected for each sentence and encoded in a one-hot manner. The data were then split into training (90%) and test (10%) sets using 10-fold cross validation. All results presented are the mean $\pm$ standard deviation across all 10 folds. This procedure was performed independently for the NB, SVC, and LR classifiers.

To select the best performing hyperparameters, we used Randomized Cross Validation to search over the possible settings. 100 randomized settings were tried per outer fold. To evaluate these 100 randomized settings, 3 inner-fold cross validation was performed on the training set only. The best set of the 100 best settings were then applied to the test set, ensuring no leakage between the training and test set. The pipeline was encouraged to maximize the F1 score (macro) during hyperparameter tuning.

Preprocessing settings explored included removing infrequently occurring words (less than 1-10 times in the corpus), computing unigram, unigram+bigram, and bigram only models, and whether it is best to remove stopwords or not.

For the models themselves, I sampled alpha values uniformly between 0.5 and 2 for the NB classifier, and I sampled C values between 1×10^{-5} and 100 uniformly for the SVM and LR models.

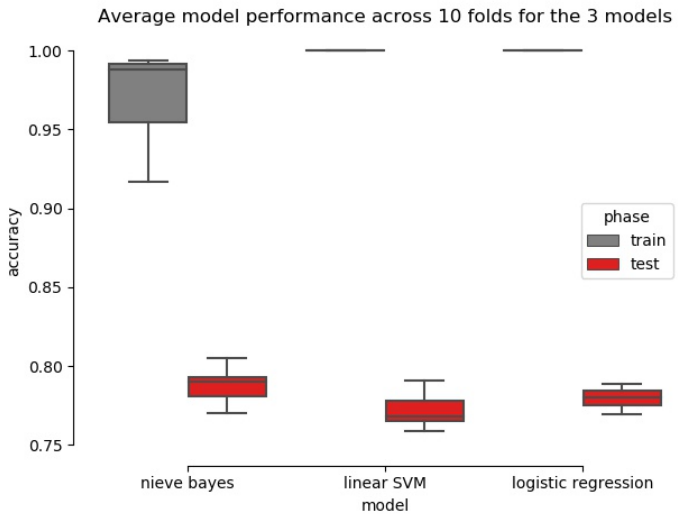
Accuracy scores on the test sets, averages across all 10 folds, for all experiments are shown below. All models performed well above chance (50% as there were an equal number of positive and negative examples in the dataset), and the NB model performed best. All models show some evidence of overfitting (nearly perfect performance on the training set with a large generalization gap) and would benefit from stronger regularization.

Inner-fold cross-validation chose the following settings most often for the NB classifier: alpha=1.18, do not remove stop words, use a unigram+bigram model, and to not remove infrequently occurring words. These latter three preprocessing hyperparameters were consistent across the three classification models used, even though these optimizations were done completely independently.

3.2. **Error Analysis.** The confusion matrix at the end of this section details the performance of the best performing model (the Naive Bayes classifier):

Note that the algorithm appears to be making a similar number of errors on positive and negative examples, so luckily our algorithm isn't too biased towards a particular answer. To get a better sense of the kinds of errors the algorithm made, we stored a small number of misclassified sentences across the 10 folds. What follows

FIGURE 1. All classification results.



are some examples:

- director hoffman , his writer and kline's agent should serve detention
- allen se atreve a atacar , a atacarse y nos ofrece gags que van de la sonrisa a la risa de larga duracian
- i have two words to say about reign of fire . great dragons !
- the cast is uniformly excellent . . . but the film itself is merely mildly charming .
- mildly entertaining .
- thekids will probably stay amused at the kaleidoscope of big , colorful characters . mom and dad can catch some quality naptime along the way .

We make a few observations. First, the model isn't picking up on what I'll call 'veiled criticisms'. For example, saying that the filmmakers should *serve detention* or that the film is *mildly entertaining* is an indirect criticism in the first case, and actually a positively-valanced sentiment in the second case. A better algorithm would need a fuller understanding of language beyond bigrams to understand that 'mildly entertaining' might actually not be an endorsement in the context of a film review. Second, at least one review in the corpus is not in English. Any other foreign language reviews we would expect to be predicted at chance level (50%). Third, some reviews use a lot of positive words in isolation as a form of either irony or politeness (e.g., *mom and dad [can nap]*, *great dragons*, *mildly charming*). These sentiments imply a negative review to the reader ('are the dragons the only good part of this film?'), but a bigram model might read 'great' or 'great dragons' and reasonably score this as a positive review. Fourth, it appears that some words were not separated properly during preprocessing (i.e., *thekids* and *naptime*). These would be treated as new words in the corpus, which isn't appropriate, as these errors aren't likely to occur often.

	Positive	**Negative**
Positive	427.0±10.35	106.1±10.28
Negative	122.5±8.88	410.6±8.85

I suggest: better data cleaning to remove foreign language reviews, a more sophisticated preprocessing pipeline that will catch joined words and other similar errors, and a more sophisticated model that understand language in context beyond what a bigram model can encode.

McGill University
Email address: joseph@viviano.ca