Profiling Poets from Spanish Sonnets

Alysa Meng

About the presenter

Alysa Meng - CSE MS student

- **CS, Math, Spanish** as an undergraduate
- Found the **Textual Studies** program last year :D

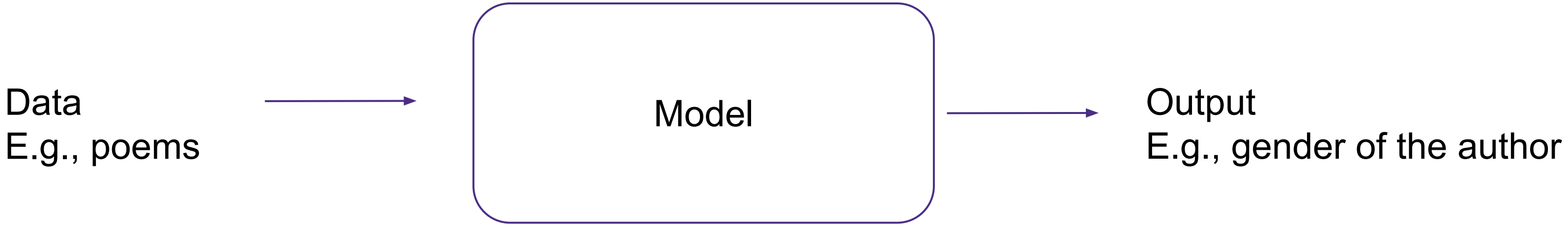
About the project

- Started in a deep learning for computer vision course (CSE599G)
  - No computer vision but…

    - **Machine Learning** (CS + Math)
    - **Sonnets** (Spanish)
    - **Applications** (Textual Studies)

Crash course on machine learning

Goal: Build a "good" model to do some task.



Data
E.g., poems

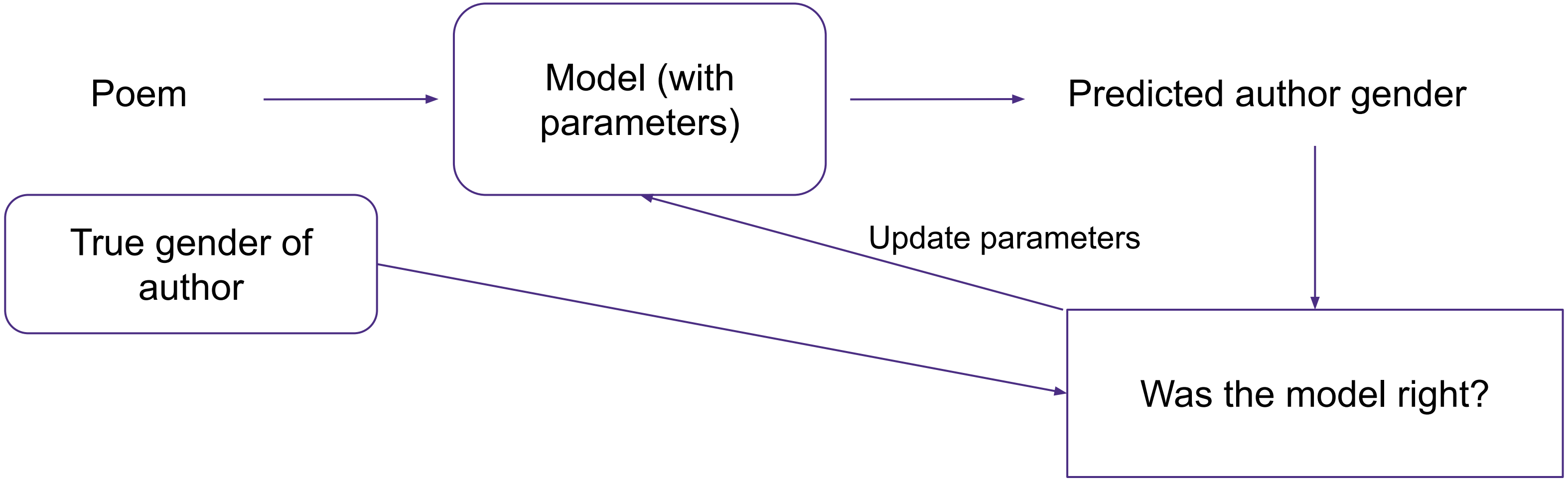
Model

Output
E.g., gender of the author

- Many choices for internal structure (architecture)
- Lots of parameters ("knobs" to turn)

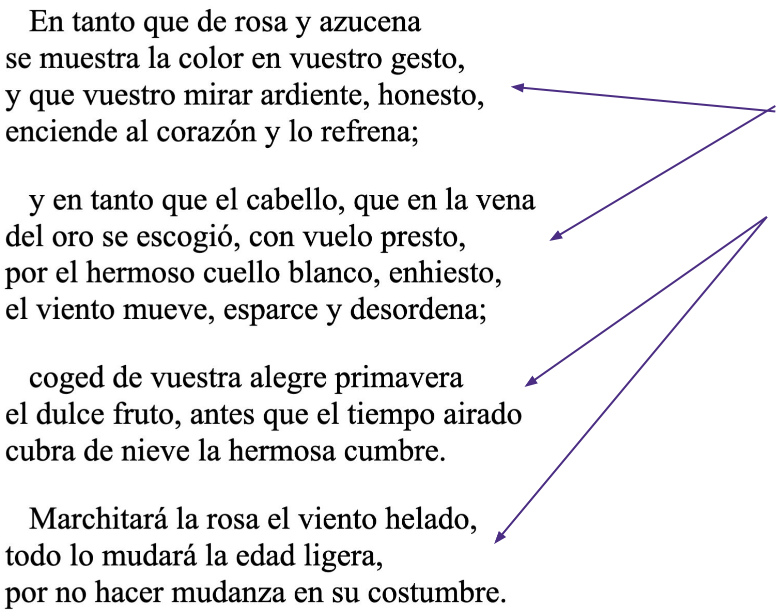
Crash course on machine learning

Supervised learning for classification
Training (trying to turn the “knobs” to make a good model)



UNIVERSITY *of* WASHINGTON

Crash course on sonnets



En tanto que de rosa y azucena
se muestra la color en vuestro gesto,
y que vuestro mirar ardiente, honesto,
enciende al corazón y lo refrena;

y en tanto que el cabello, que en la vena
del oro se escogió, con vuelo presto,
por el hermoso cuello blanco, enhiesto,
el viento mueve, esparce y desordena;

coged de vuestra alegre primavera
el dulce fruto, antes que el tiempo airado
cubra de nieve la hermosa cumbre.

Marchitará la rosa el viento helado,
todo lo mudará la edad ligera,
por no hacer mudanza en su costumbre.

14 verses broken into 2 quatrains (4 lines), 2 tercets (3 lines)

1 verse → 11 syllables

Verses rhyme, 10th syllable is stressed

Garcilaso de la Vega's Soneto XXIII

Dataset

DISCO - Diachronic Spanish Sonnet Corpus

~4k richly annotated sonnets

- Metadata about author
- Notes on poetic features
- From Biblioteca Virtual Miguel de Cervantes, Wikisource

Class imbalance

- E.g., ~90% of the dataset is male poets, ~60% from 19th century, ~60% from Spain

Task: Given a Spanish sonnet, what biographical information can we infer about its author?

3 classification tasks

Gender
Birth country
Time period (birth century)

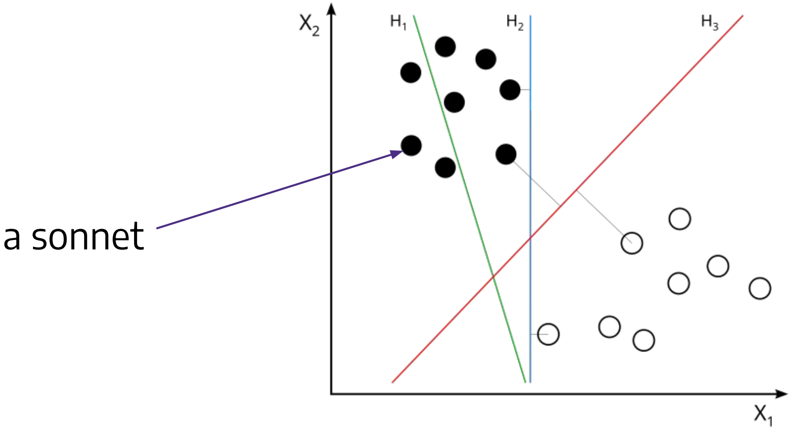
3 approaches per task

Support Vector Machine
Long Short-Term Memory
Transformer

Applications: authorship attribution, style comparisons, better indexing on similarity, metadata generation, etc
In this case: Understand how ML models represent Hispanic poetry

Architecture: Support Vector Machine

Learn linear boundaries (divide space with lines for classification).

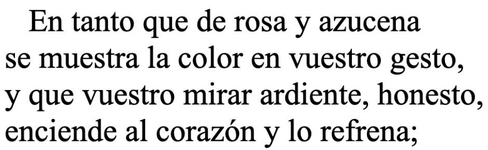


a sonnet

Turning sonnets into numbers

Idea 1: Create features

(measurable characteristics about the sonnet)



En tanto que de rosa y azucena
se muestra la color en vuestro gesto,
y que vuestro mirar ardiente, honesto,
enciende al corazón y lo refrena;

Garcilaso de la Vega's Soneto XXIII

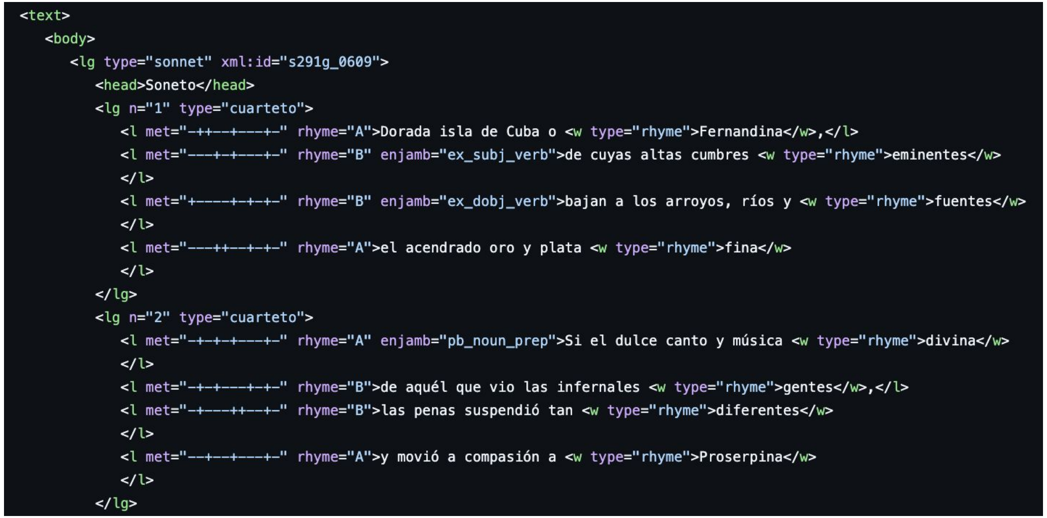
"Bag of words"

- Count frequencies of words.
  - We have 3 "y"s, 2 "en"s, 1 "rosa", and so on
- Every syllable in a verse is stressed (1) or unstressed (0)

How do we get this? It seems hard…

Turning sonnets into numbers

DISCO has TEI files!



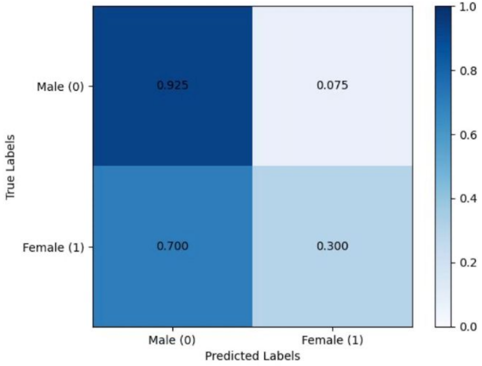
```
<text>
   <body>
      <lg type="sonnet" xml:id="s291g_0609">
         <head>Soneto</head>
         <lg n="1" type="cuarteto">
            <l met="-++--+---+-" rhyme="A">Dorada isla de Cuba o <w type="rhyme">Fernandina</w>,</l>
            <l met="---+-+---+-" rhyme="B" enjamb="ex_subj_verb">de cuyas altas cumbres <w type="rhyme">eminentes</w>
            </l>
            <l met="+----+-+-+-" rhyme="B" enjamb="ex_dobj_verb">bajan a los arroyos, ríos y <w type="rhyme">fuentes</w>
            </l>
            <l met="---++--+-+-" rhyme="A">el acendrado oro y plata <w type="rhyme">fina</w>
            </l>
         </lg>
         <lg n="2" type="cuarteto">
            <l met="-+-+-+---+-" rhyme="A" enjamb="pb_noun_prep">Si el dulce canto y música <w type="rhyme">divina</w>
            </l>
            <l met="-+-+---+-+-" rhyme="B">de aquél que vio las infernales <w type="rhyme">gentes</w>,</l>
            <l met="-+---++--+-" rhyme="B">las penas suspendió tan <w type="rhyme">diferentes</w>
            </l>
            <l met="--+--+---+-" rhyme="A">y movió a compasión a <w type="rhyme">Proserpina</w>
            </l>
         </lg>
```

UNIVERSITY *of* WASHINGTON

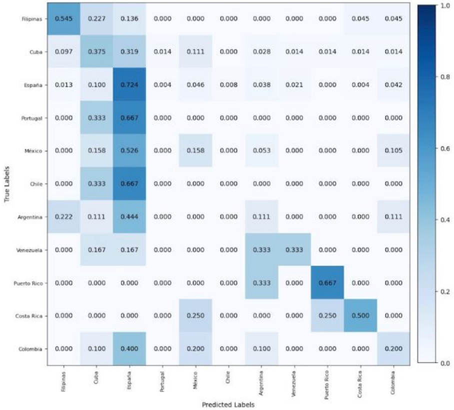
Support Vector Machine Results

Normalized Confusion Matrices

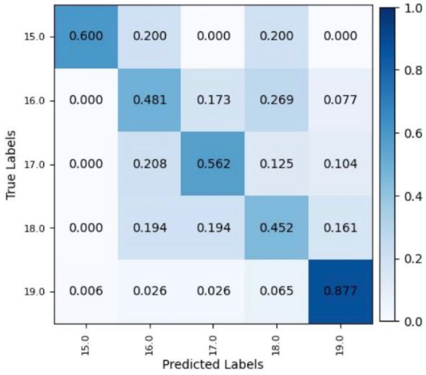
Gender



Country



Century

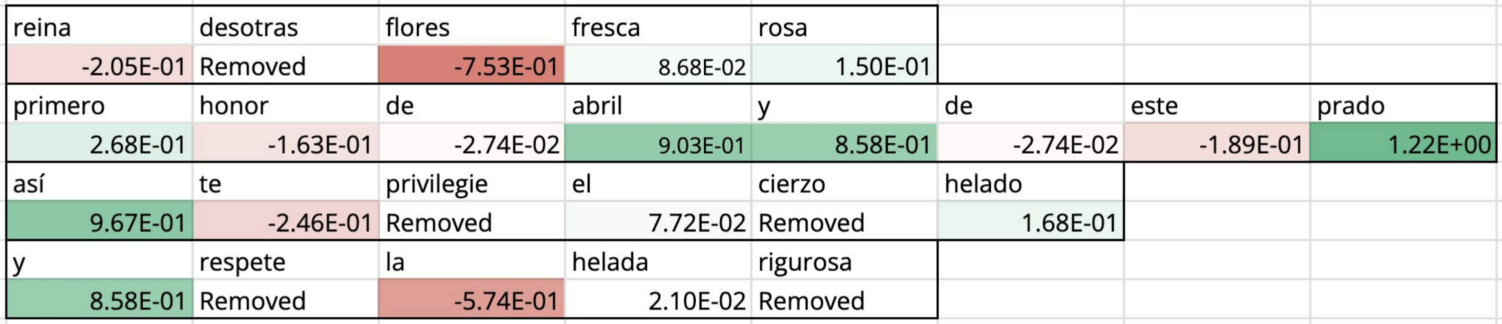


UNIVERSITY *of* WASHINGTON

Understanding Support Vector Machines

Model learns templates for each class. The template that "fits" the best will be the one it choses.

First quatrain of Spain's Luis Martín de la Plaza’s “Fresca rosa” (16th century weights)



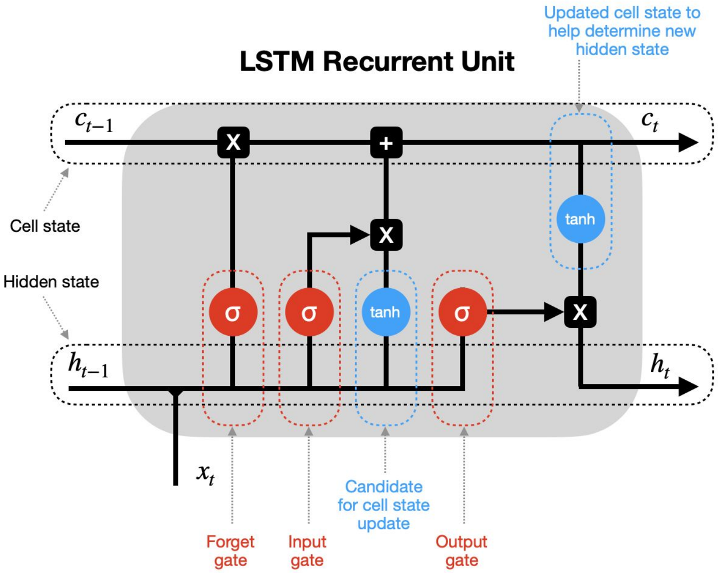
reina	desotras	flores	fresca	rosa			
-2.05E-01	Removed	-7.53E-01	8.68E-02	1.50E-01			
primero	honor	de	abril	y	de	este	prado
2.68E-01	-1.63E-01	-2.74E-02	9.03E-01	8.58E-01	-2.74E-02	-1.89E-01	1.22E+00
así	te	privilegie	el	cierzo	helado		
9.67E-01	-2.46E-01	Removed	7.72E-02	Removed	1.68E-01		
y	respete	la	helada	rigurosa			
8.58E-01	Removed	-5.74E-01	2.10E-02	Removed			

Support Vector Machine Limitations

- Limited vocabulary
- Not all things can be classified with linear boundaries*
- Features are tedious and require expert knowledge
- Hard to capture sequential information

*Technically, there are ways to fix this.

Architecture: Long Short-Term Memory

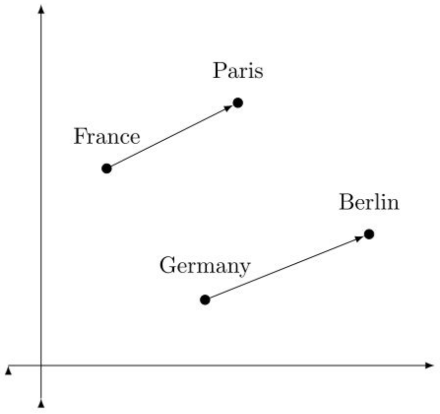


Process text one word at a time. "Remember" what was processed when you need it.

Image by Saul Dobilas

Turning sonnets into numbers

Idea 2: Word Embeddings (can we translate semantics into numbers?

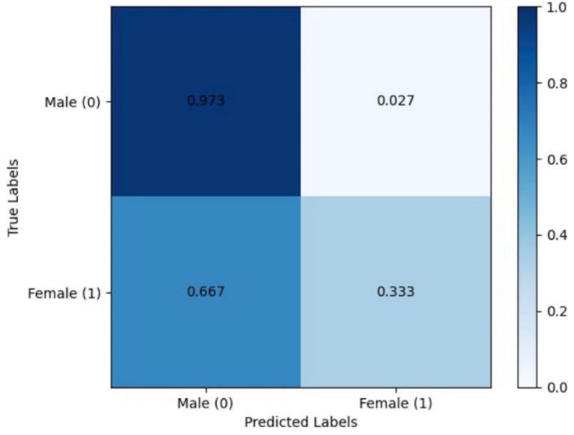


LSTM should learn these.

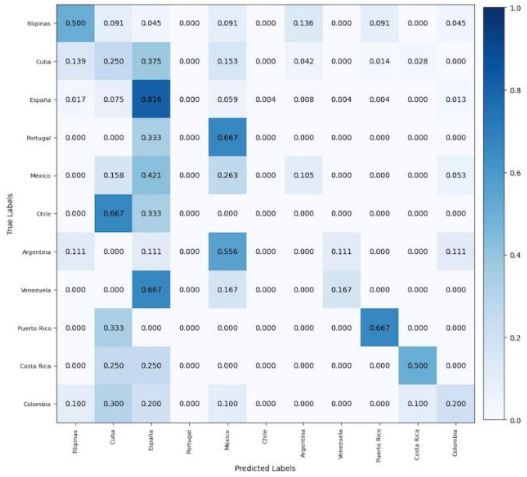
We can kickstart the process by using pre-trained embeddings (word embeddings someone else got from training a model — probably).

LSTM Results

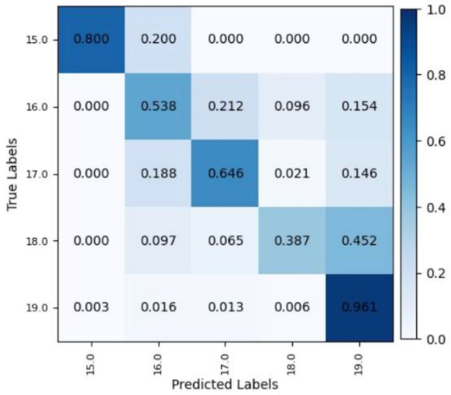
Gender



Country



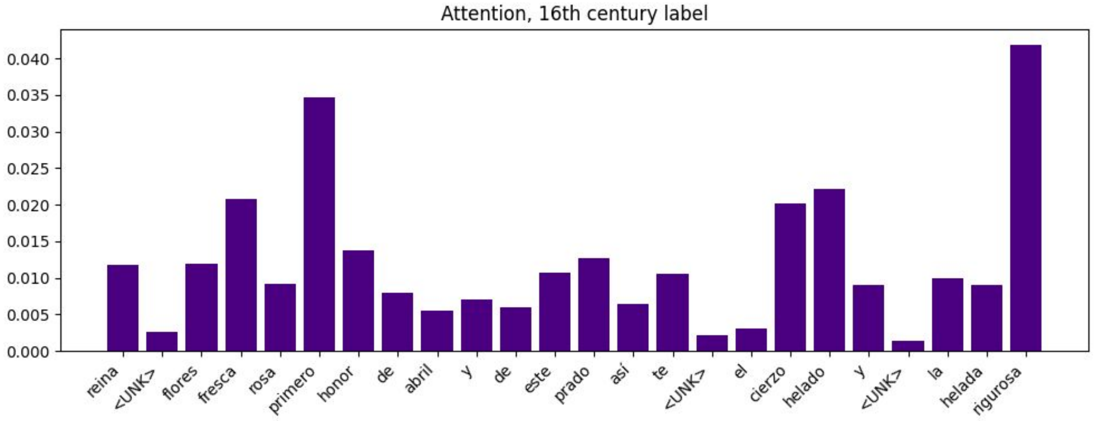
Century



UNIVERSITY *of* WASHINGTON

Understanding LSTMs

Optionally, we include an attention layer. Helpful for understanding what the model is "looking at" (maybe).



UNIVERSITY *of* WASHINGTON

Architecture: Transformer

Words have different meanings in different contexts.

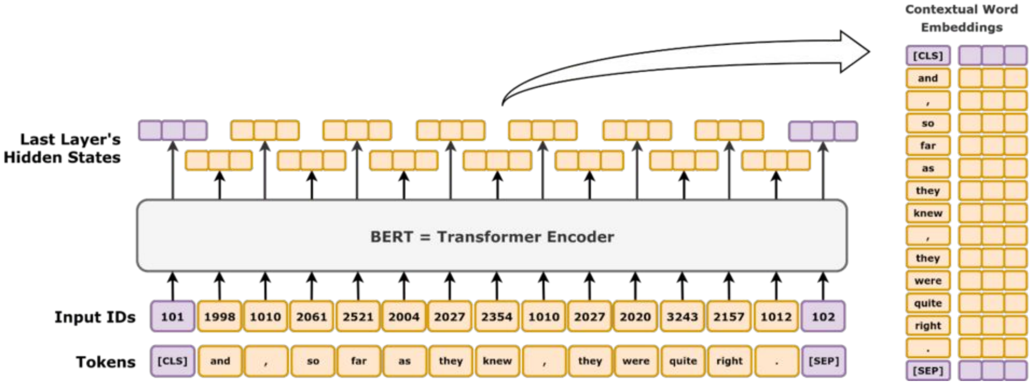
I'm a ________ at UW.

Bidirectional Encoder Representation from Transformers (BERT)

Turning sonnets into numbers

"Fine-tuning"

Idea 3: Use an encoder model that someone made that has already been trained on related data



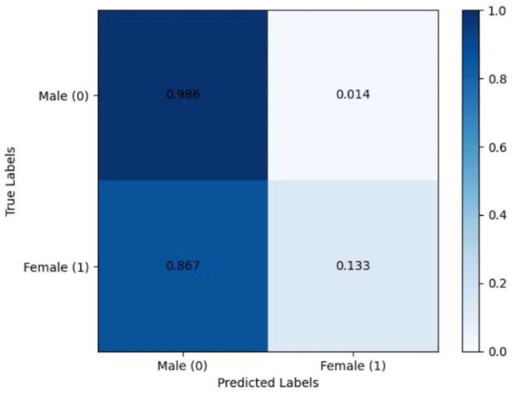
Instead of words, use tokens (chunks of words).

UNIVERSITY *of* WASHINGTON

Image by Daniel Godoy

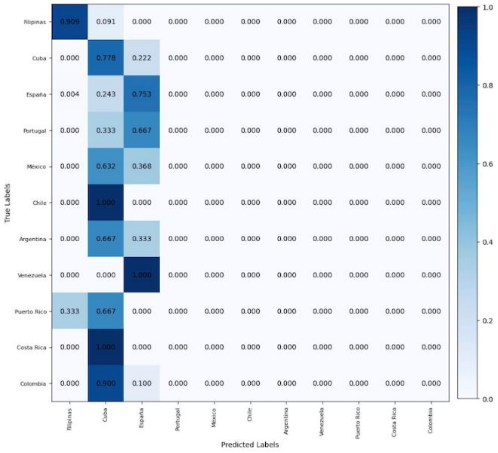
Transformer Results (Fine-tuned BETO)

Gender



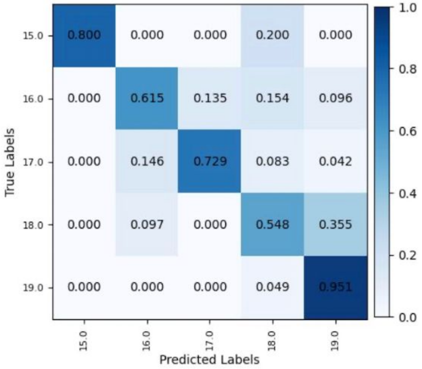
True Labels \ Predicted Labels	Male (0)	Female (1)
Male (0)	0.986	0.014
Female (1)	0.867	0.133

Country



True Labels \ Predicted Labels	Filipinas	Cuba	España	Portugal	México	Chile	Argentina	Venezuela	Puerto Rico	Costa Rica	Colombia
Filipinas	0.909	0.091	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cuba	0.000	0.778	0.222	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
España	0.004	0.243	0.753	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Portugal	0.000	0.333	0.667	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
México	0.000	0.632	0.368	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Chile	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Argentina	0.000	0.667	0.333	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Venezuela	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Puerto Rico	0.333	0.667	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Costa Rica	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Colombia	0.000	0.900	0.100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Century



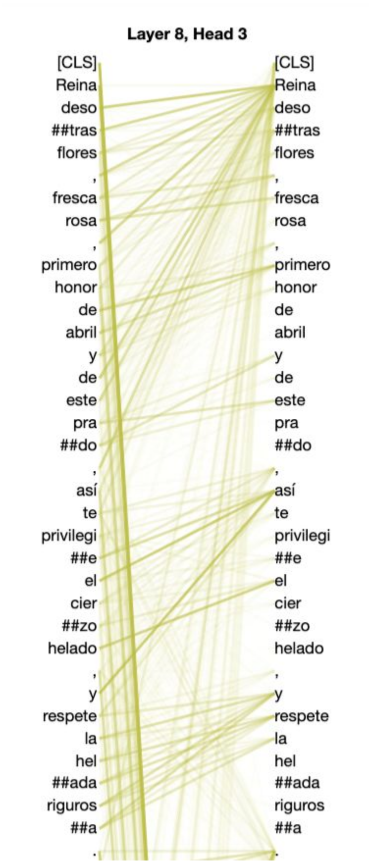
True Labels \ Predicted Labels	15.0	16.0	17.0	18.0	19.0
15.0	0.800	0.000	0.000	0.200	0.000
16.0	0.000	0.615	0.135	0.154	0.096
17.0	0.000	0.146	0.729	0.083	0.042
18.0	0.000	0.097	0.000	0.548	0.355
19.0	0.000	0.000	0.000	0.049	0.951

Understanding Transformers

Also use an attention mechanism.

Later layers = higher level features?

- Punctuation matters.
- Capturing relationships within and between verses.



How would a human perform?

I tried to classify the author birth century for **6** sonnets.

3 correct!

Why was this so hard?

And you may still be asking:

What does a CSE project have to do with **Textual Studies** ?

The data matters.

- Who is represented in this dataset?
- Where does the data and metadata come from?
- Who has access to it?

The model matters.

- What does it cost to train a model?
- How about to store it? Where to store it?
- How do AI architectures shape literary analysis?
  - Turning human text into numbers

The people matter.

- Who was the author and original audience?
- How are human decisions embedded in the models?
- Who is going to use models and what for?

Thank you!

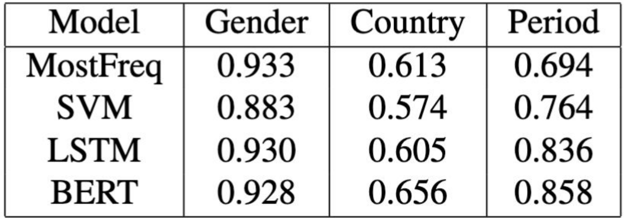
Especially to Geoffrey Turnovsky and the Textual Studies program.

Any questions?

Contact: Alysa Meng (menga@uw.edu)

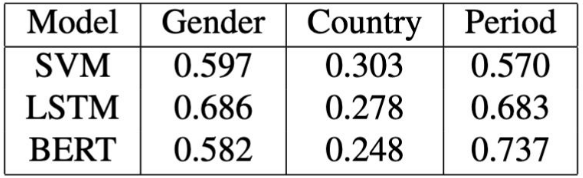
Appendix

Model Accuracy, Macro F1 Scores



Model	Gender	Country	Period
MostFreq	0.933	0.613	0.694
SVM	0.883	0.574	0.764
LSTM	0.930	0.605	0.836
BERT	0.928	0.656	0.858

Table 1. Model Accuracies (MostFreq denotes a classifier that always chooses the most frequent label)



Model	Gender	Country	Period
SVM	0.597	0.303	0.570
LSTM	0.686	0.278	0.683
BERT	0.582	0.248	0.737

Table 2. Model Macro F1 Scores