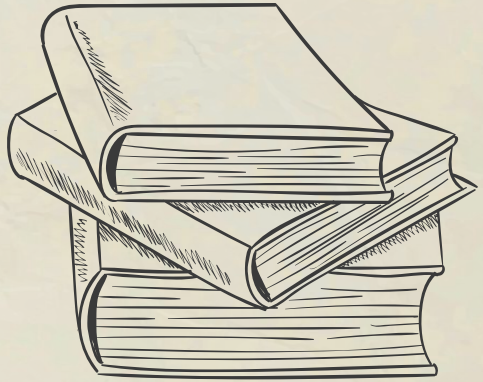
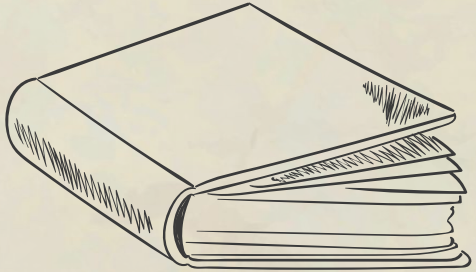
English Text Era Identification

CSCI 3832

Eddie Kiernan, Zach Conroy,
James Gashi, Turner Land





Problem Statement

- Language is dynamic, evolving over time with shifts in vocabulary, syntax, and style
  - Hwæt. We Gar-Dena in geardagum, þeodcyninga, þrym gefrunon, hu ða æþelingas ellen fremedon.
- Changes influenced by cultural, social, and historical factors, making the study of linguistic evolution an intriguing area for research.
- Understanding when a piece of text was written can provide valuable insights into literary trends, historical contexts, and cultural analysis.
  - Dating language change
- We aim to develop a model capable of estimating the date range of an English text based solely on its linguistic features.
- By analyzing stylistic and linguistic patterns like:
  - Vocabulary trends
  - Morphosyntax
  - Phrase frequencies
- Model will learn to associate textual characteristics with specific historical periods, aiming to classify a given text/excerpt with a date range.

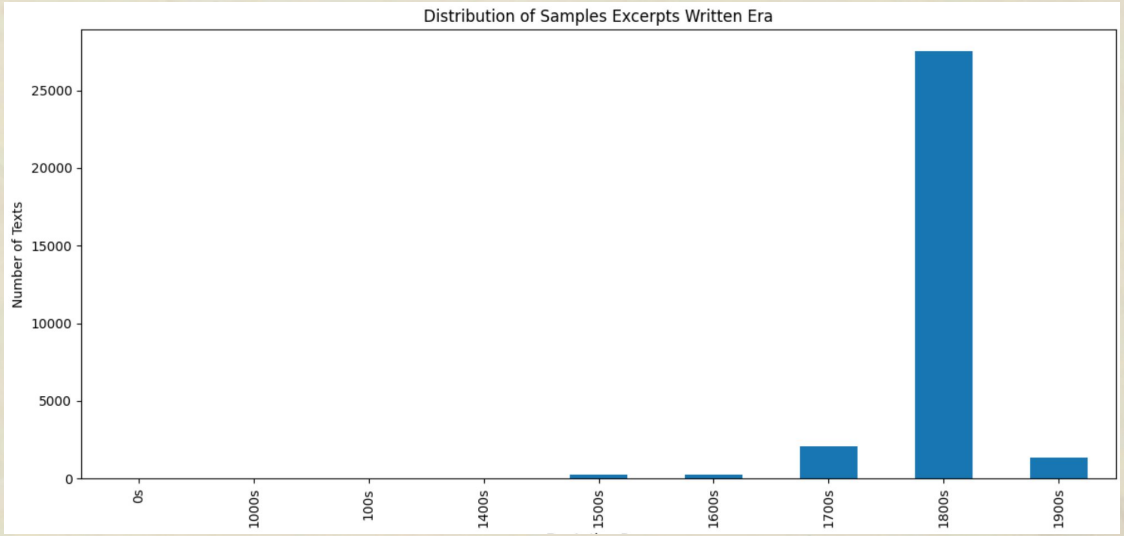
Data Collection

- Data sourced from the free ebook source Project Gutenberg.
- Retrieved all available texts: 75,627
- Metadata in .csv tied to .txt file with unique ‘Text#’
- Mainly interested in “Language”, “Authors”, and “Type”

Text#	Type	Issued	Title	Language	Authors	Subjects	LoCC	Bookshelves
1	Text	1971-12-01	The Declaration of Independence of the United ...	en	Jefferson, Thomas, 1743-1826	United States -- History -- Revolution, 1775-1...	E201; JK	Politics; American Revolutionary War; United S...
2	Text	1972-12-01	The United States Bill of Rights\r\nThe Ten Or...	en	United States	Civil rights -- United States -- Sources; Unit...	JK; KF	Politics; American Revolutionary War; United S...
3	Text	1973-11-01	John F. Kennedy's Inaugural Address	en	Kennedy, John F. (John Fitzgerald), 1917-1963	United States -- Foreign relations -- 1961-196...	E838	Browsing: History - American; Browsing: Politics
4	Text	1973-11-01	Lincoln's Gettysburg Address\r\nGiven November...	en	Lincoln, Abraham, 1809-1865	Consecration of cemeteries -- Pennsylvania -- ...	E456	US Civil War; Browsing: History - American; Br...
5	Text	1975-12-01	The United States Constitution	en	United States	United States -- Politics and government -- 17...	JK; KF	United States; Politics; American Revolutionar...

Data Issues

- No official written date provided
  - No historical English corpus that reliably tags all books with original publication date
- Limited old text (non-uniform distribution)
- OCR issues for certain texts
- Is text translated or not
- Copyright issues for modern text

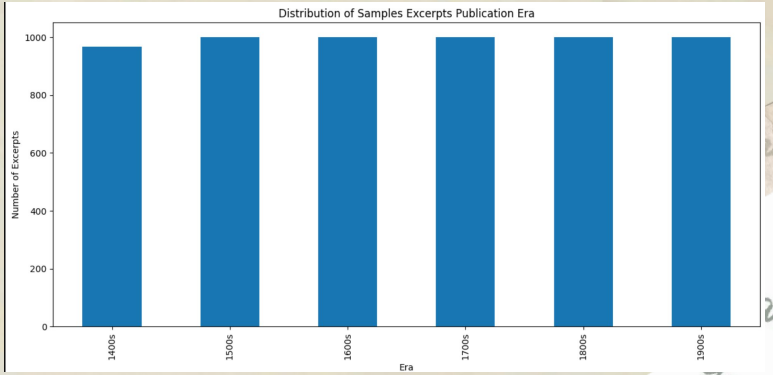


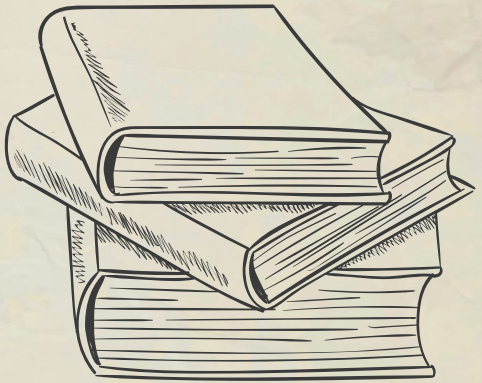
Data Cleaning

Solutions:

- Base written date on author birth/death/average
- Instead of training on full texts, training on excerpts
  - Random sampling old texts for uniform distribution
  - Avoids problems with intro's prefaces from modern day
  - 400-600 chars long
- Remove text with apparent OCR issues if too many unrecognizable chars
- Filter out any text that has multiple authors, translator, original language not in English
- Includes texts ranging from 1400s to 1900s
- 6,000 entries

	text	text_number	label
0	still. She'd got into a groove; he'd have to f...	1472	1800s
1	thinkest thou? Speak. _Filippo._ Holy Father! ...	21628	1700s
2	will burn, and I shall have a giddiness every ...	4928	1700s
3	contain'd the famous Baths of _Anastasia_, whi...	53083	1400s
4	by two sundry priests; and, further, since hat...	32155	1400s





Models

Naive Bayes

- Split 60/20/20, sklearn train_test_split()
- Feature engineering
  - Bag of Words
  - Average word size
  - Average sentence length
  - POS frequencies - future
- Bigram Model
- Used nltk naive bayes model and word tokenizer
- Goal is to prove feasibility of BERT model

Naive Bayes - Bag of words approach

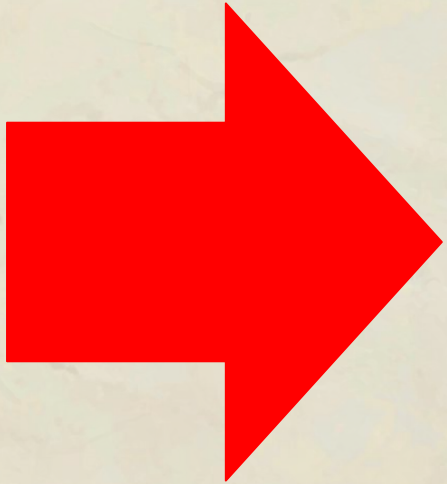
1. Find words that occur in era of training examples
2. Filter out words that appear in other eras
3. Manually refine words list
4. Use words as features and train NB model
5. Evaluate model

```
vocab_1400s = set()
words_list_1400s = []
for sample in samples_1400s:
    for word in word_tokenize(sample):
        vocab_1400s.add(word.lower())
        words_list_1400s.append(word.lower())
```

```
def contains_1400s_word_refined_list(example):
    return {word : 1 if word in example else 0 for word in refined_1400s_words}
```

Refining words lists

```
('_me._', 69),
('_ogy._', 66),
('suche', 62),
('||', 59),
('thay', 57),
('dyd', 47),...
```



```
"suche", "thay",
"wyll", "dyd",
"mynde", "woulde",
"thyng", "lyfe",
"theyr",
"thynke","muche",
"tyme","apon","nowe"
```

Filter out weird tokens, proper nouns, etc.

Initial performance

- Baseline of performance is 1/6 = **.167**, constant model
- Validation accuracy using only bag of words: **.431**
- Validation acc using only sentence length: **.216**
- Validation acc using only average word length as a feature : **.1472**
- Validation acc combining BOW, sentence length, word length : **.419**, worse than only bag of words

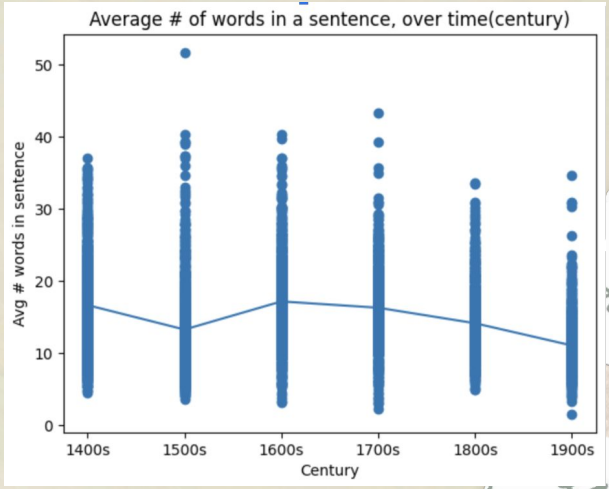
Naive Bayes - Bigram Approach

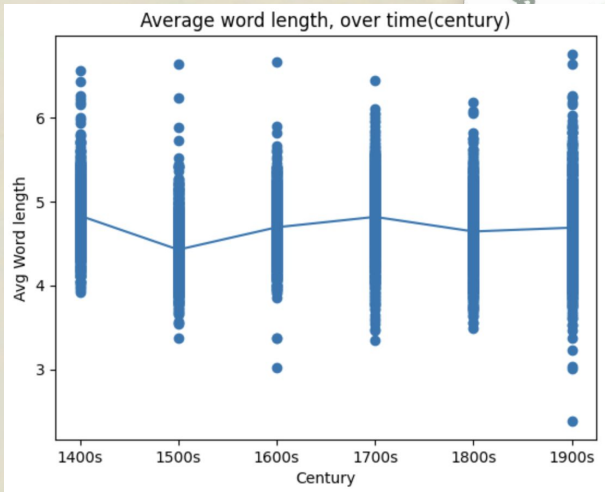
- Basing off of initial success of BOW approach
- Boosted performance to .7 validation accuracy
- Every version of the model did a great job classifying 1400s and 1500s texts
- Struggled to differentiate between 1600s, 1700s, and 1800s texts

```
def generate_bigrams(example):
    tokens = word_tokenize(example.lower())
    bigrams = []
    for i in range(len(tokens) - 1):
        bigrams.append((tokens[i], tokens[i+1]))
    return bigrams
```

Observed language trends

- Transition from middle english to modern english (Modern English ~ 1500 CE)
  - Chaucer Standard (Early Modern English) appears around 1500-1600, gaining standardization with the printing press and influence from French and Latin
- Changes in orthography (dyd to did)
  - Standardization of orthography only occurring in the late 1800s
- No clear trend in sentence length or word length
- New words invented around industrial revolution, and Victorian Era has a large corpus
- Could be due to our text bias, and lack of texts dating to earlier periods





BERT

- Google's BERT model was the harbinger of modern AI and NLP research
  - Self supervised
  - Encoder only transformer architecture
- Good at classification
- Fine tuned, 6,000 examples
  - 80% training, 10% validation, 10% test

Epoch	Training Loss	Validation Loss	Accuracy
1	0.958700	0.900812	0.645740
2	0.871200	0.861156	0.650224
3	0.830700	0.847114	0.639013
4	0.812900	0.832812	0.643498

Issues With BERT

- Long training times
- Inflexible
- Issues with our BERT implementation
    - Too large of a learning rate
    - Too big of a data scope for our purposes

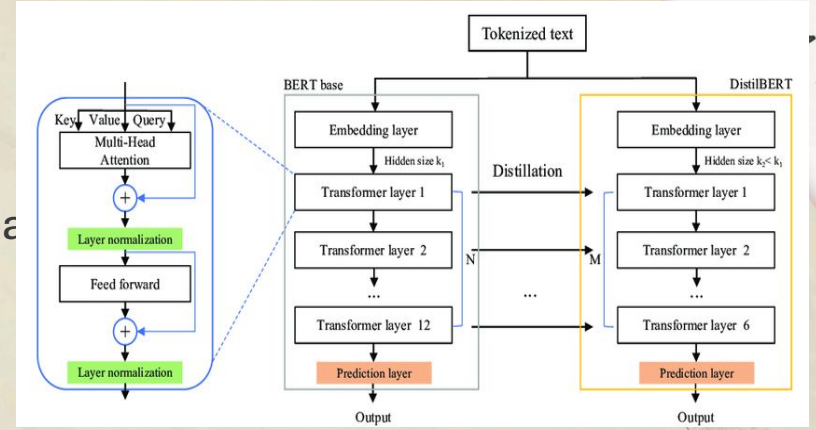
DistilBERT (uncased)

What is DistilBert?

- Transformers model
- Smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERT ba model as a teacher.
- Uncased - not case sensitive

Why DistilBert?

- DistilBERT is 60% faster than BERT.
- DistilBERT retains 97% of BERT performance
- Opted for speed during training, while trying to maintain most of performance
  - One training iteration using CURC job took ~1 hour



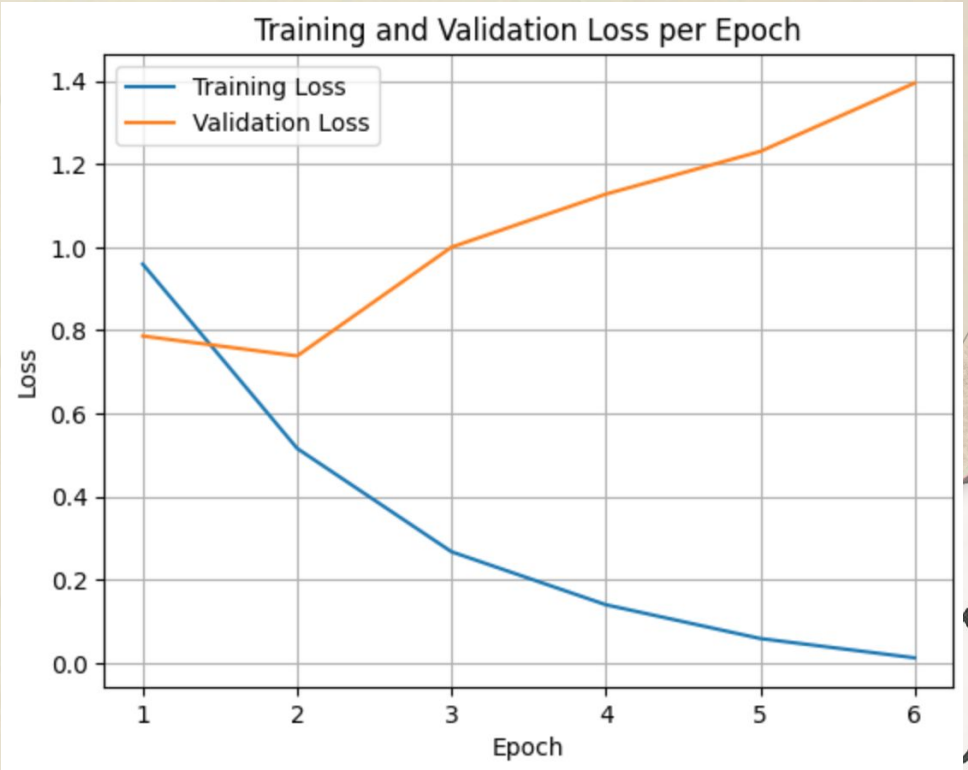
Tokenization/Training

Tokenization
- DistilBertTokenizerFast
- Max length padding of 256

Training
- 80/10/10 data split
- Learning rate=1e-4
- Weight decay=0.05
- Epochs=6

Loss
- Cross Entropy Loss

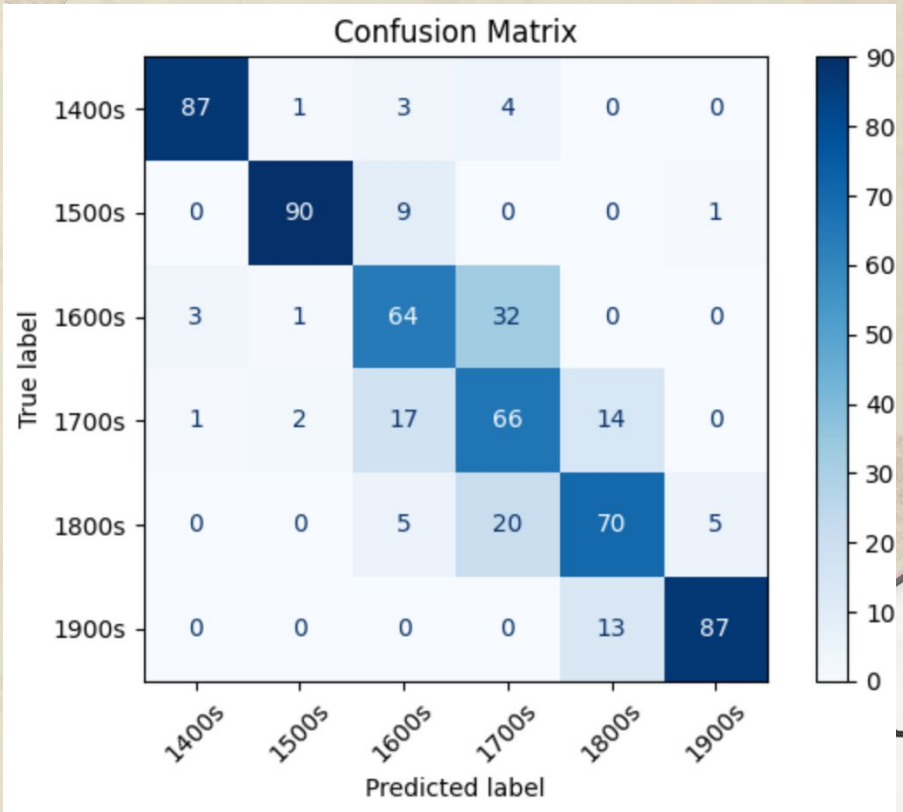
Issue
- Overfitting
  - Might look to decrease learning rate and increase weight decay
- Still long training time



Test Metrics

- Accuracy: 0.79
- Same issues as seen in NB model
  - Weakest performance 1600s-1800s
- Very strong performance 1400s-1500s and 1900s

	precision	recall	f1-score
1400s	0.90	0.87	0.89
1500s	0.87	0.94	0.90
1600s	0.64	0.67	0.65
1700s	0.60	0.59	0.59
1800s	0.68	0.67	0.68
1900s	0.92	0.86	0.89



Future Plans

- Finished BERT evaluation
  - Reduce learning rate
- Part of speech tagging for NB
- General grammar features (morphology)
- Fine tune DistilBERT to fix overfitting