Hemingway Next Word Prediction

Using DistilBERT with *The Sun Also Rises*

The Context

Hemingway

As a famous writer, Ernest Hemingway's became famous for his modernist novels. Elements of his sparse style are ripe for Natural Language Processing.

NLP

As Natural Language Processing (NLP) has advanced since the 1970s, methods were changed from rule based to statistical approaches. The advent of large datasets allowed for larger language models.

BERT

Google's Bidirectional Encoder Representations from Transformers (BERT) is a large language model designed for masked word and next sentence prediction. It is the prediction engine for this project.

The process

Data Wrangling → Exploratory Data Analysis → Unmasking & Training

The Sun Also Rises

- 68,315 words
- Mixed case
- Full punctuation range
- Epilogue
- 19 chapters

Frequency Patterns

- Word cloud
- Nouns
- Verbs
- Adjectives
- Sentence length
- Ngrams

Large Language Models

- Tokenization
- Vectors
- spaCy
- Matched spans
- DistilBERT
- Similarity scores

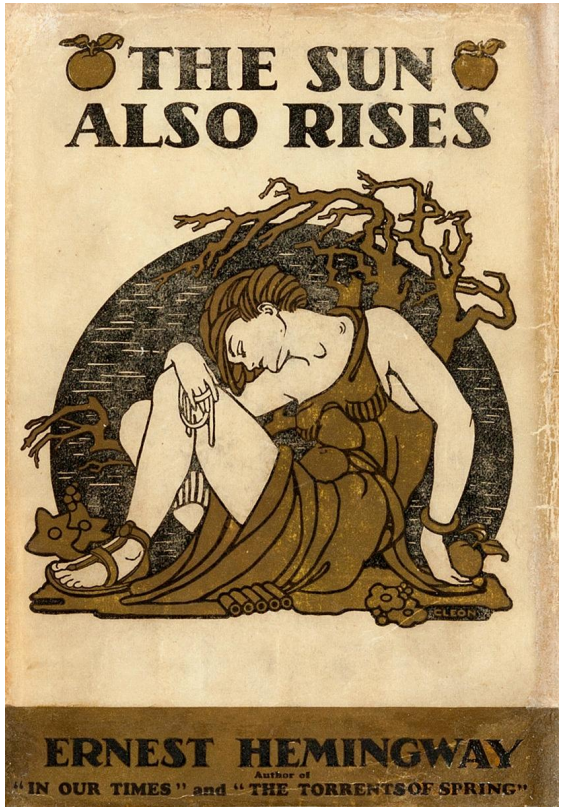
Technical Solution

spaCy & BERT

- Compare similarity scores between true and predicted words from both untrained and trained DistilBERT models.
- If the trained model's similarity score is significantly different from the untrained model's similarity score . . .
- And the similarity scores are higher for the trained model . . .
- Training DistilBERT on the novel resulted in better predictions.

Data Wrangling

Original cover & sample text from *The Sun Also Rises*



"Nobody ever lives their life all the way up except bull-fighters."

"I'm not interested in bull-fighters. That's an abnormal life. I want to go back in the country in South America. We could have a great trip."

"Did you ever think about going to British East Africa to shoot?"

"No, I wouldn't like that."

"I'd go there with you."

"No; that doesn't interest me."

"That's because you never read a book about it. Go on and read a book all full of love affairs with the beautiful shiny black princesses."

Cleaning the data

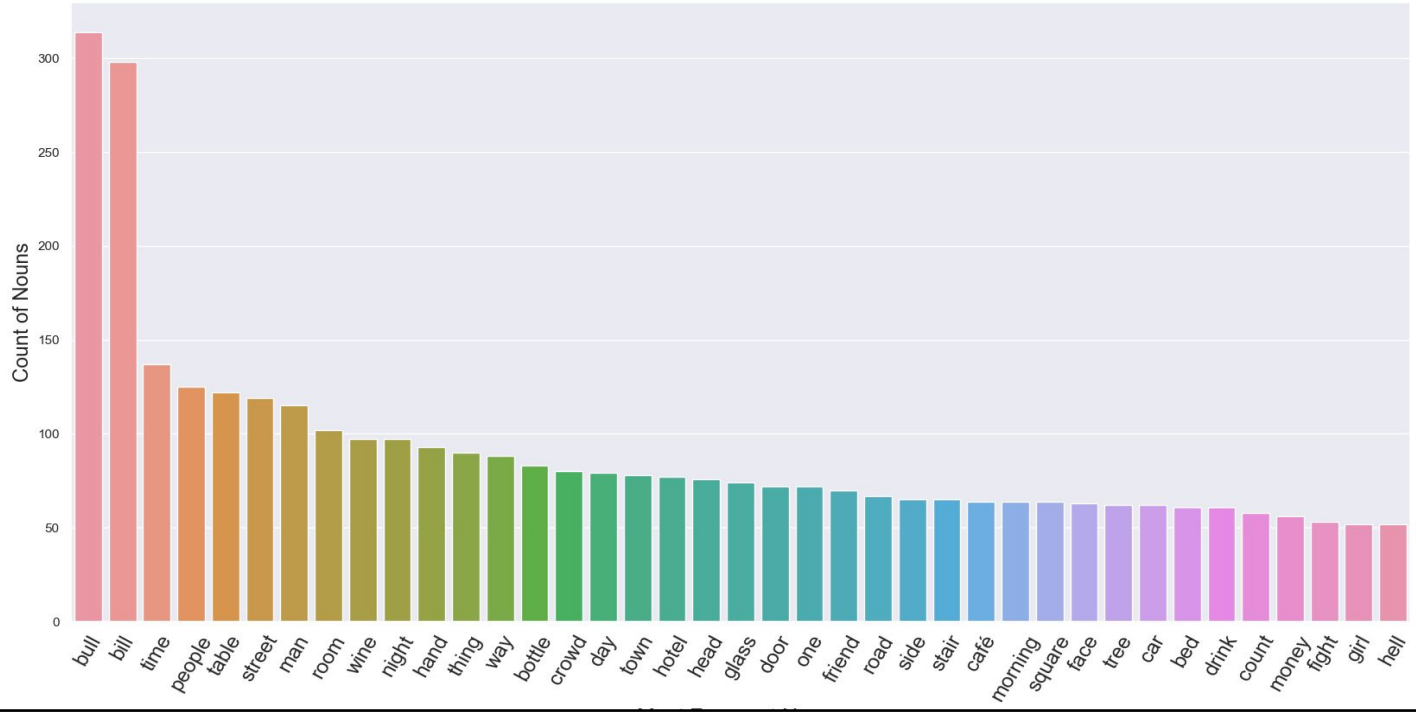
- The novel text comes from Project Gutenberg: https://www.gutenberg.org/.
- Legal front matter from Project Gutenberg was removed.
- For most NLP tasks in this project:
  - Punctuation was removed.
  - Stopwords were removed.
  - The text was lowercased.
- One side effect of removing the stop words is that the narrative flow was often destroyed.
- This project was run with the stop words in place.

Exploratory Data Analysis

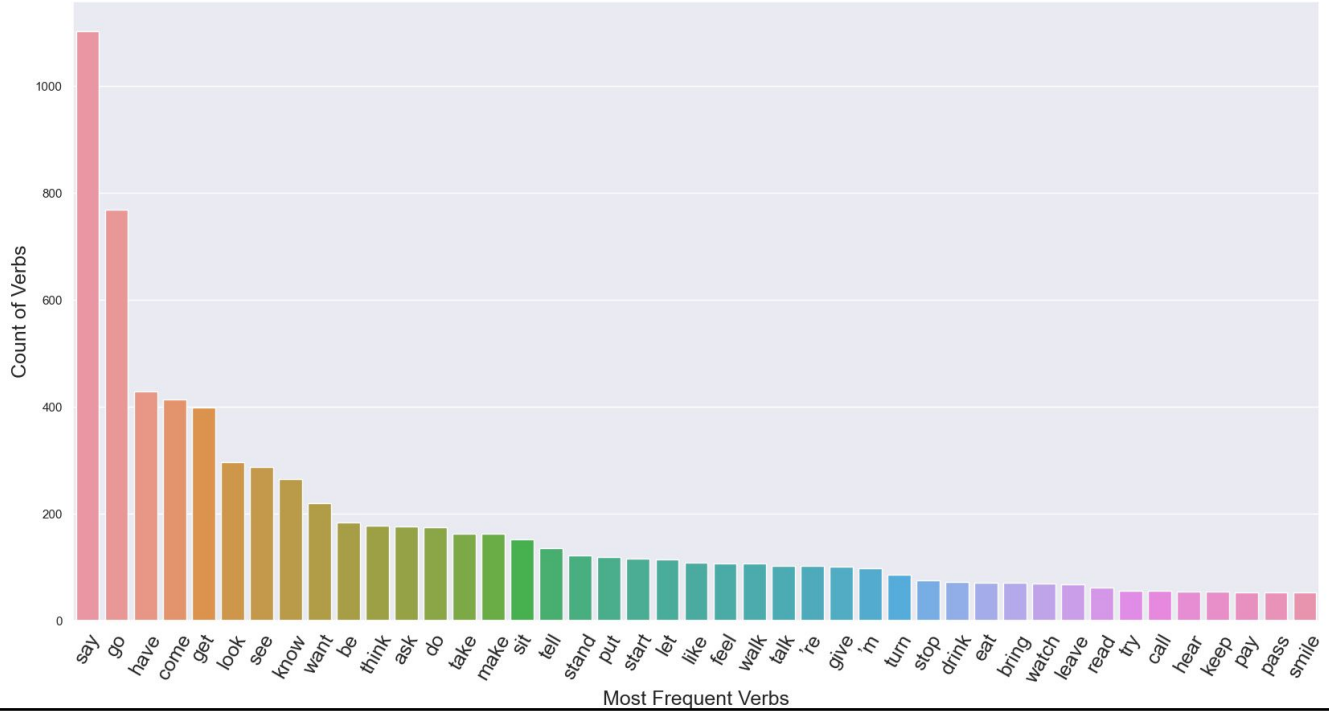
Word Cloud


said
brett
went
s
don t
came
bill
one
go
back
know
good
asked
looked
bill said
mike
want
m
look
now
three
hand
little
old
room
ll
something
started
friend
take
away
hell
hotel
tell
town
well
romero
took
wanted
didn t
stood
come
right
turned
bed
jake
saw
night
always
count
put
cohn
make
drink
square
way
see
head
made
time
let s
got
around
sat
brett said
going
table
think
ve
big
road
thing
long
many
day
door
stairs
mike said
never
oh
much
people
tree
d
say
yes
bottle
us
re
two
crowd
face
wine
street
café
nice
really
looking
another

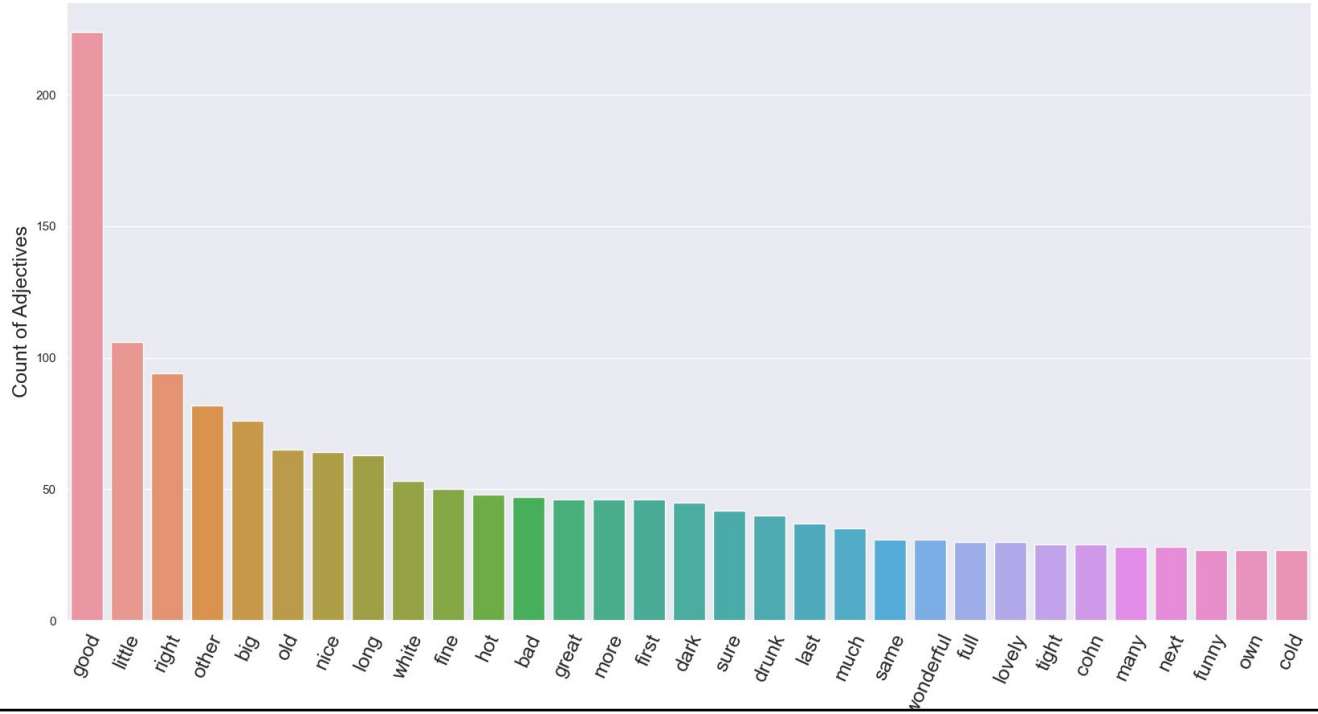
Most frequent nouns



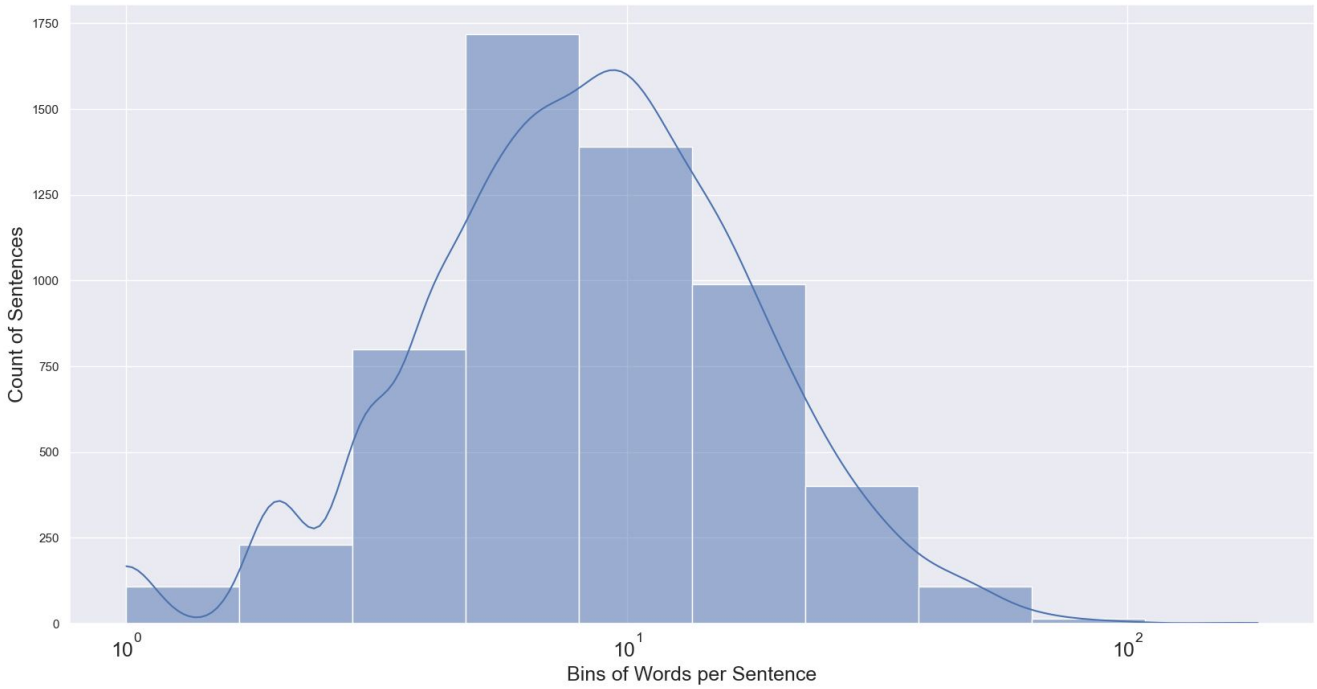
Most frequent verbs



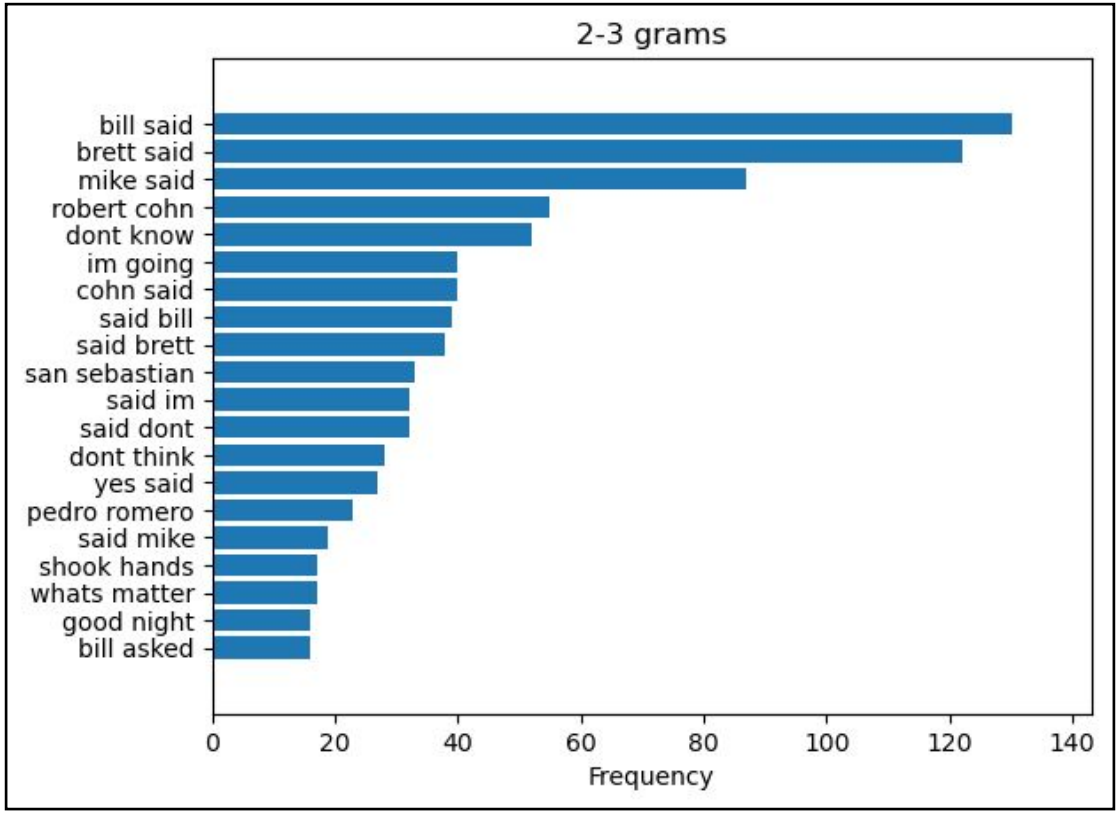
Most frequent adjectives



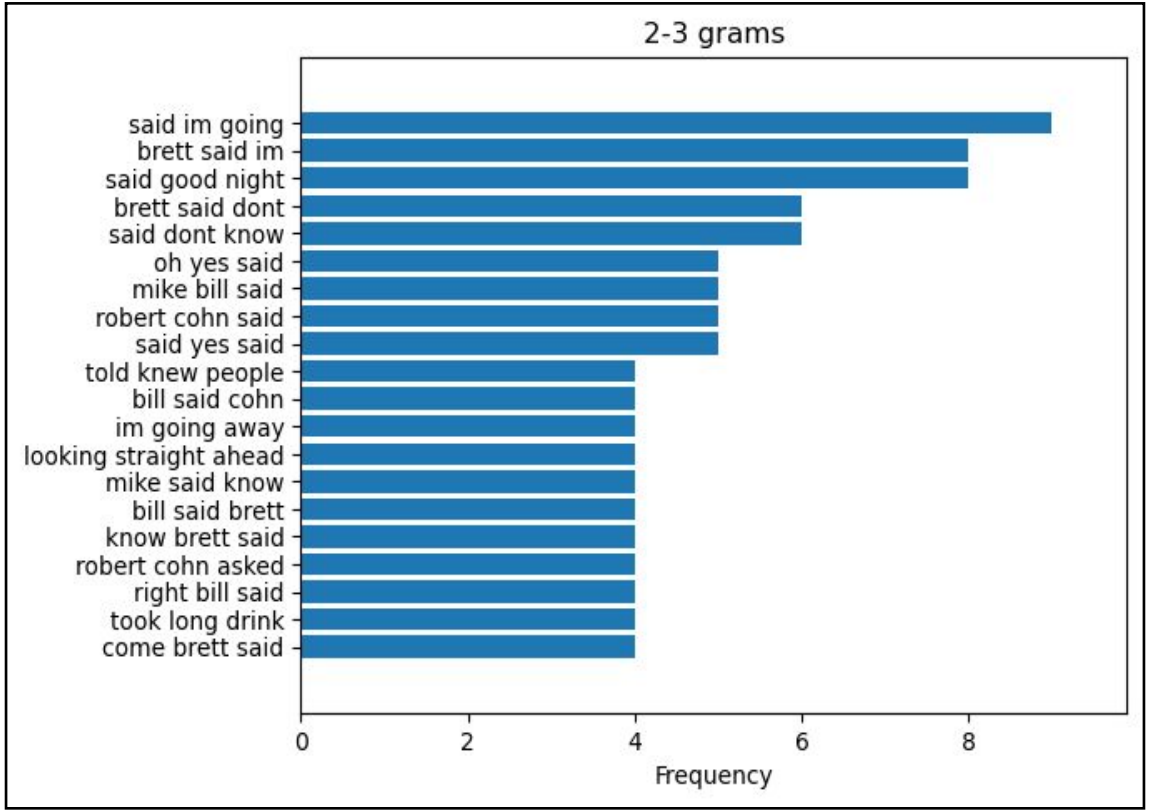
Sentence length (log transformed)



Bigrams


2-3 grams
bill said
brett said
mike said
robert cohn
dont know
im going
cohn said
said bill
said brett
san sebastian
said im
said dont
dont think
yes said
pedro romero
said mike
shook hands
whats matter
good night
bill asked
0
20
40
60
80
100
120
140
Frequency

Trigrams


2-3 grams
said im going
brett said im
said good night
brett said dont
said dont know
oh yes said
mike bill said
robert cohn said
said yes said
told knew people
bill said cohn
im going away
looking straight ahead
mike said know
bill said brett
know brett said
robert cohn asked
right bill said
took long drink
come brett said
0
2
4
6
8
Frequency

Unmasking & Training

Vectorization with the word *worry*

```
worry [ 2.9654     2.7027    -2.4714    -1.0352    -0.079387  -0.72472
  0.53368    3.8428    -1.3603     2.3727     0.22586    0.75641
 -4.1174     0.83035    0.73932   -1.948      0.27187   -3.479
 -1.3285     1.1281    -0.53185    1.2922     0.53875   -3.2607
 -2.4816    -1.059     -0.81081   -2.788     -3.3112     4.8863
  0.13508   -5.8231     0.74262   -2.3775     3.5092     1.4314
  0.092381   0.64814    5.9549     4.9481    -1.8308     0.028828
  2.5295     1.3199    -2.2144    -0.91651    2.4754    -2.894
 -2.5452     0.48501    0.091526  -3.0365     2.2395    -3.6397
 -2.7671     1.0851     0.60659    3.4097     0.0090092  1.5626
  3.1725    -0.69717   -1.6433    -1.811     -1.0557    -1.0214
  0.58334    0.46305    2.307      1.2694     1.1162     2.6585
 -2.9381    -1.3663     1.7924    -2.3868     1.9842     0.44663
  3.5436     0.32388   -0.87528   -0.96563    2.767      2.3729
  2.0346    -0.8488     0.015217  -5.0707     0.21105    2.6214
 -1.3929     2.8451    -3.3657     1.1677    -0.441     -5.6514
 -0.59095    1.7608    -0.74169   -1.6772     0.52922    1.4826
  0.21566    1.1929    -0.3963     1.2581    -0.44234    1.3542
 -0.15589    0.40392   -2.3522    -1.302     -1.555     -4.6697
 -0.23593    4.8172    -5.3577    -3.7915     2.7631     0.63667
 -1.5845    -1.6256    -5.1559    -1.2487    -3.9772     3.1195
  1.1596     5.6142    -2.2377    -0.84486   -0.24609    1.9577
 -1.1008     1.7674     2.3359    -2.0569    -4.6275    -3.7086
  5.1245    -4.2225    -0.75321    2.0614     1.9893    -0.23126
 -4.3271    -2.7021    -1.9307    -1.1503    -1.0111    -0.2155   ]
```

Masked sentences

- Actual Hemingway Sentence (after stopword removal):
  - 'bullfights good bullfighters stayed montoyas hotel'

- Masked Hemingway Sentence:
  - bullfights good bullfighters stayed montoyas [MASK]

Sample of Similarity Score Results

Base DistilBERT			Fine-tuned Model	
Predicted Word	Similarity Scores	True Word	Similarity Scores	Predicted Word
carries	0.2918593414	put	0.2918593414	carries
collects	0.1621982666	put	0.1621982666	collects
pulls	0.5215437336	put	0.2968177014	handles
handles	0.2968177014	put	0.5215437336	pulls
opens	0.253190749	put	0.2420324967	steals
prepares	0.2417159155	brought	0.2417159155	prepares
drinks	0.124599894	brought	0.124599894	drinks
orders	0.06745186402	brought	0.06745186402	orders
delivers	0.1889055538	brought	0.4605056586	brings

.1940

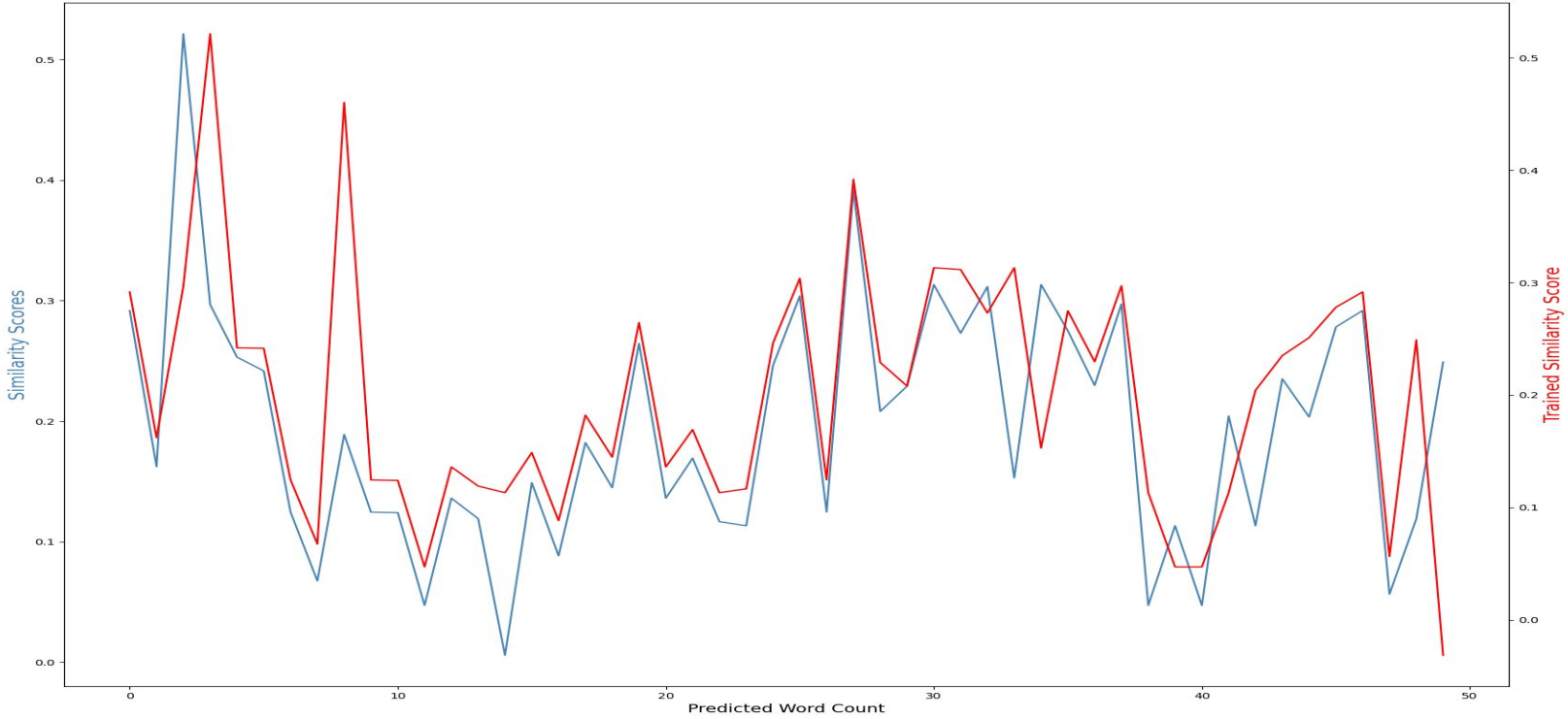
Mean Similarity Score for Untrained Model

.1993

Mean Similarity Score for Trained Model

With the Mann-Whitney test, the two distributions of similarity scores are not significantly different with a probability of $p < 0.97$. With the trained model not having higher similarity scores, we can conclude that training the model on the novel did not improve its predictive performance.

Similarity Score Comparison



Impact

Trained BERT Models

For an effectively trained model, a great deal more text would be needed to render a change in predictive power.