


# LTM: Language Time Machine

Group 11
Felix Holz
David Wildauer
Leopold Magurano




github.com/NeXTormer/LTM-LanguageTimeMachine

# Motivation

- Language evolves over time.
- **RQ:** *Is it possible for a machine learning model to pick up on small changes in the english language to predict the publish year of text snippet?*
- Goals

  **→ predict the time period (year) in which a text snippet was published and retrieve books from a similar time period**

# The Dataset

## 01 Source

Project Gutenberg

## 02 Content

320 years of eBooks
1700-2023

## 03 Features

Title, Year, Content

## 04 Size

13.000 Unique Books

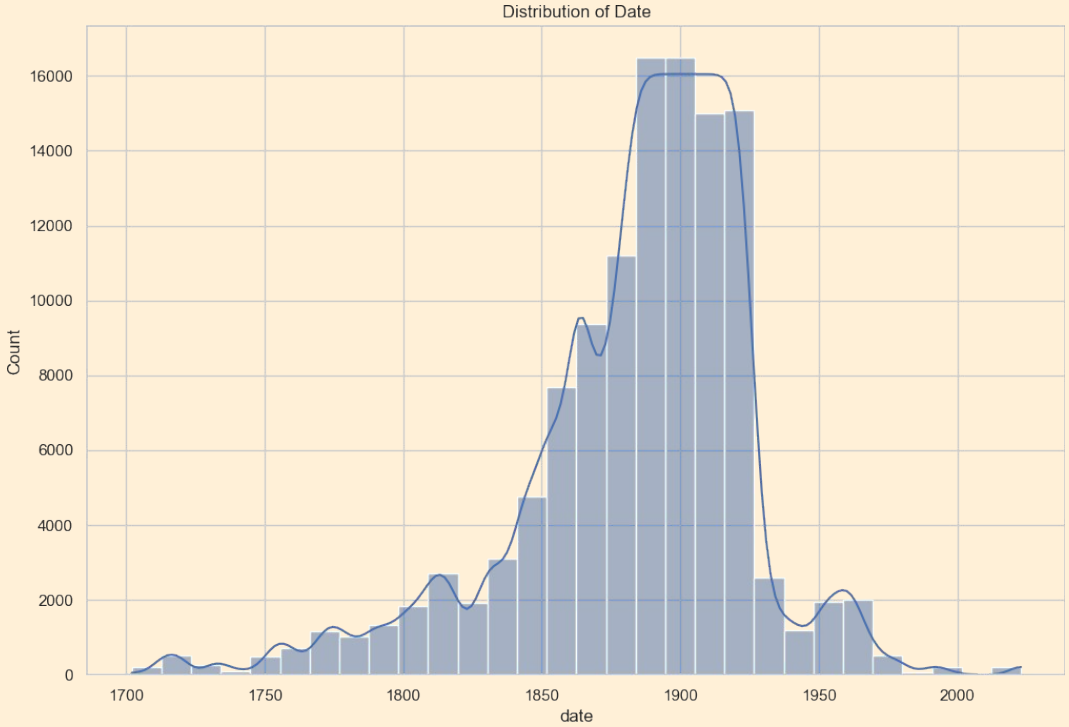
## 05 Preprocessing

94 GB of raw data

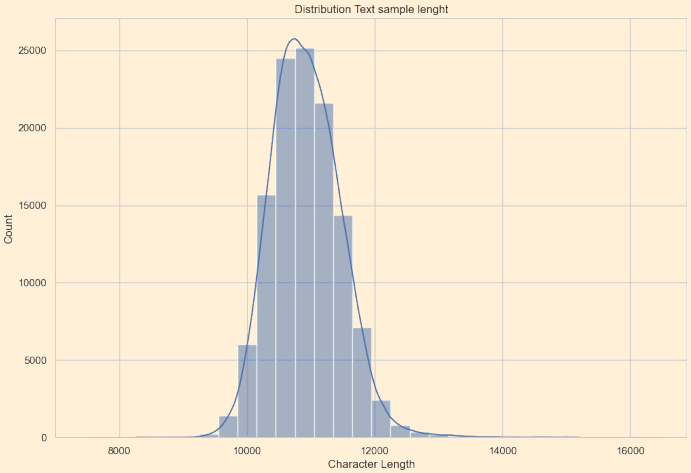
## 06 Access

www.gutenberg.org/

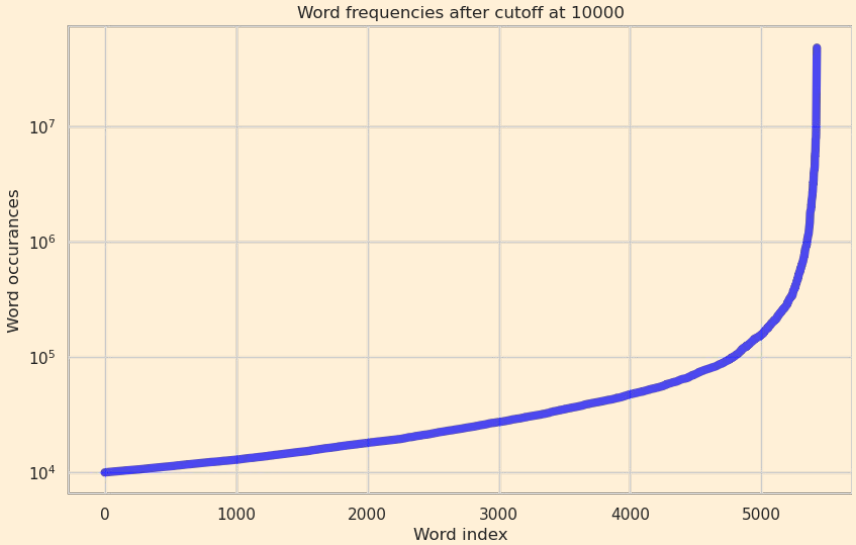
# The Dataset



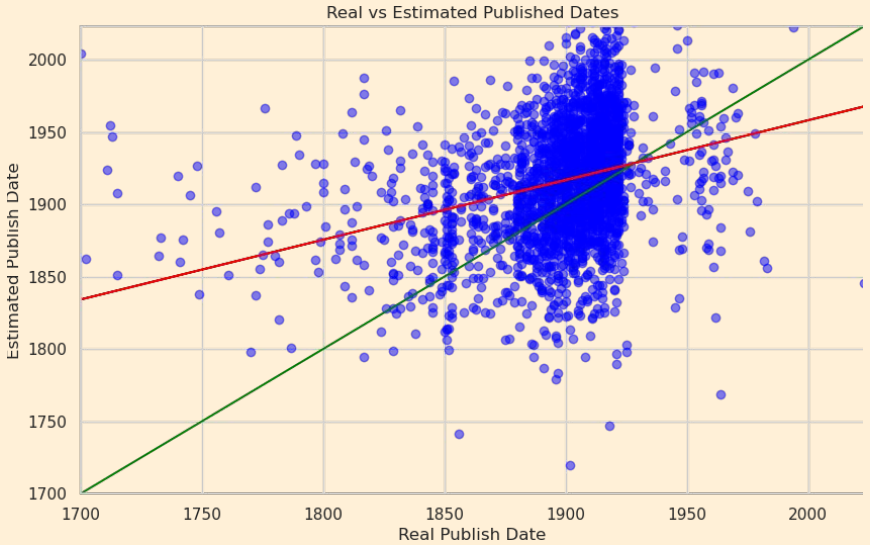

120.000 Text samples with average length of 11k characters



# Word Frequency Analysis




- Only a few words that occur often
- Words that occur under 10000 times in corpus left out

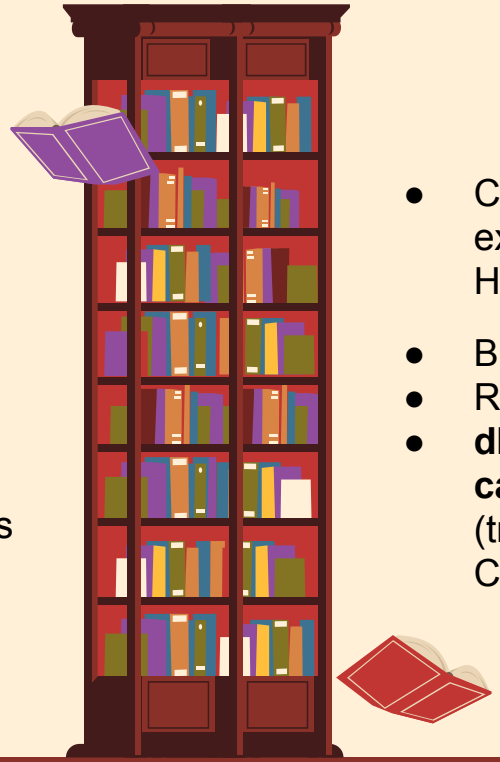



- MAE: 17.38
- Standard Deviation: 43.26
- Correct predictions within 10 years: 17.21 %

# Methods

## Doc 2 Vec

- Doc2Vec an extension of Word2Vec
- unique vector for each document (year)
- Training optimizes vector representations to predict documents in a continuous vector space

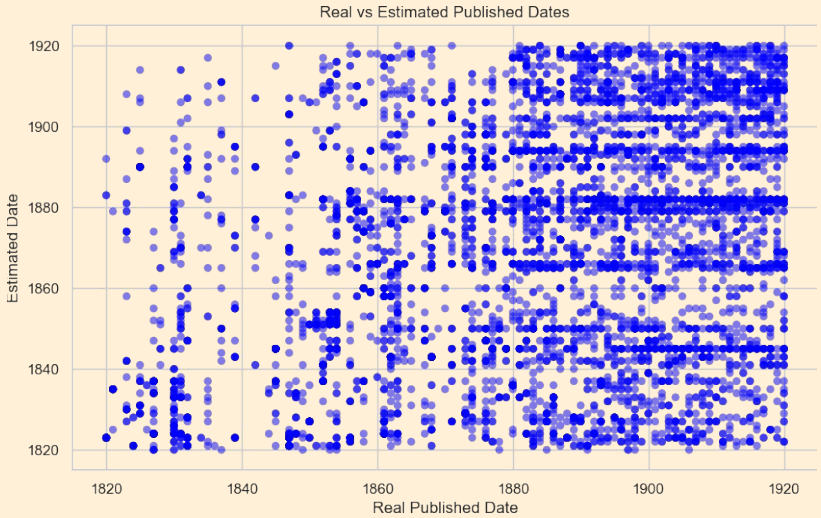


## BERT

- Comparison of standard BERT and extended models from the HuggingFace community
- BERT (bert-base-uncased)
- RoBERTa (roberta-base)
- **dbmdz/bert-base-historic-english-cased → the best one for our task** (trained with the British Library Corpus)

# Classification Results

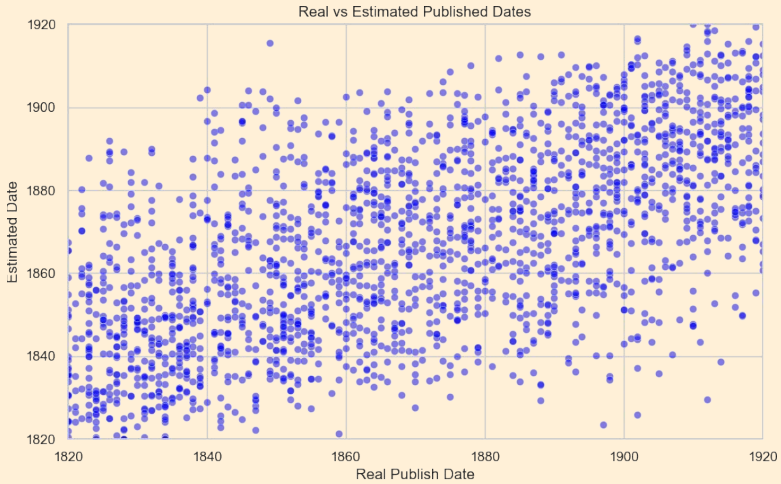
## Doc 2 Vec




- MAE: 27.49
- Standard Deviation: 29.47
- Correct predictions within 10 years: 28.2 %

## BERT

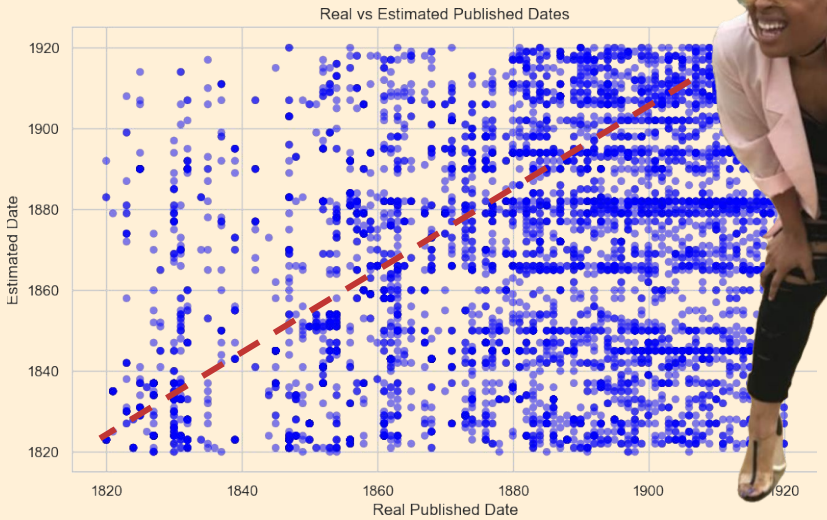
dbmdz/bert-base-historic-english-cased




- MAE: 18.85
- Standard Deviation: 23.10
- Correct predictions within 10 years: 33.9 %

# Classification Results

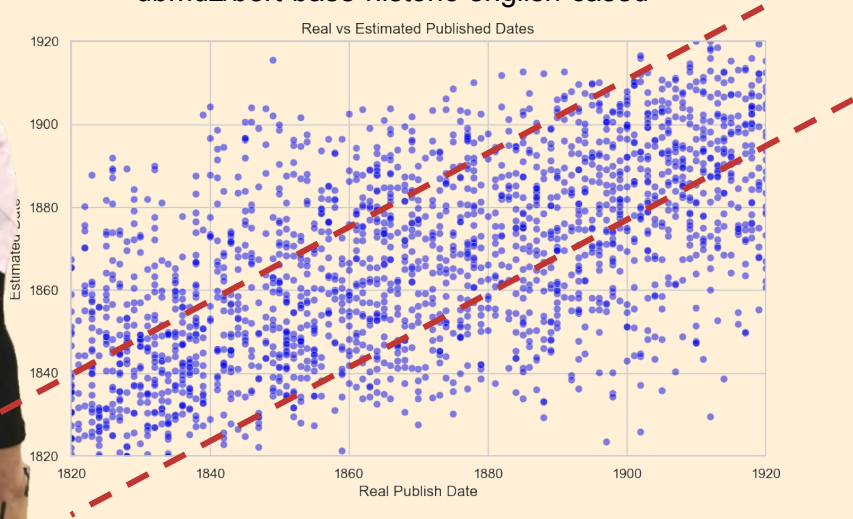
## Doc 2 Vec




- MAE: 27.49
- Standard Deviation: 29.47
- Correct predictions within 10 years: 28.2 %

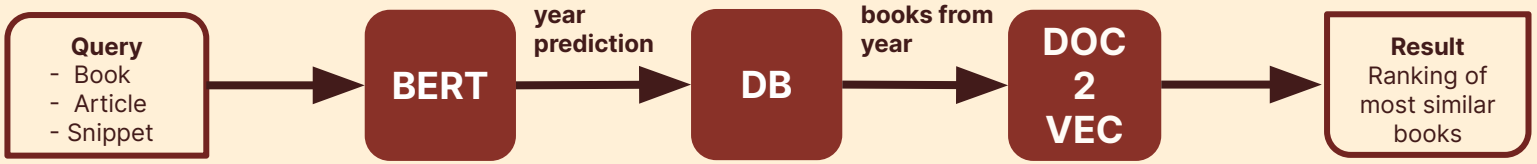
## BERT

dbmdz/bert-base-historic-english-cased




- MAE: 18.85
- Standard Deviation: 23.10
- Correct predictions within 10 years: 33.9 %

# IR Pipeline




"known to some of you i dare say as the throstle or mavis he gives the thrushwhich somehow doesnt go ..."

Similarity: 0.531 | Title: The Pearl of Orr's Island: A Story of the Coast of Maine | Author: Harriet Beecher Stowe | Publish Date: 1862

Similarity: 0.524 | Title: From the Easy Chair, Volume 1 | Author: George William Curtis | Publish Date: 1862 S62

...

(2/10)

# Conclusion (incl. limitations/biases)

- Output is only as good as the input (i.e. the dataset) → noise like not reliable publish dates can be a huge problem
- Books in train dataset must be evenly spread amongst topics and years
- Classification heavily influenced by size of dataset → training BERT model is very time consuming
- Doc2Vec not good for date prediction → focus too strongly on topical similarity → better used for re-ranking