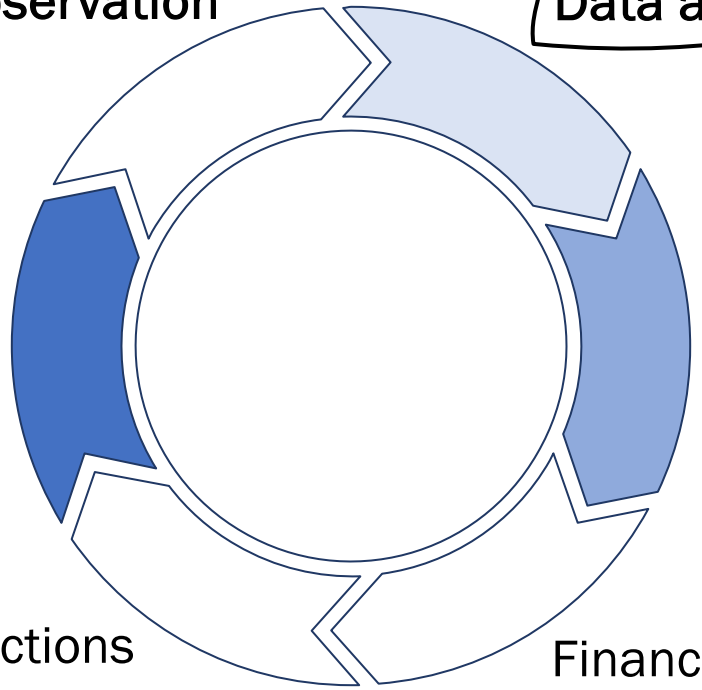
Prof. Dr. Martin Schlather
Lehrstuhl für Stochastik und ihre Anwendungen

Prof. Dr. Leif Döring
Lehrstuhl für Stochastik

UNIVERSITÄT MANNHEIM

How to predict stock market prices from daily news articles?

How to become a billionaire



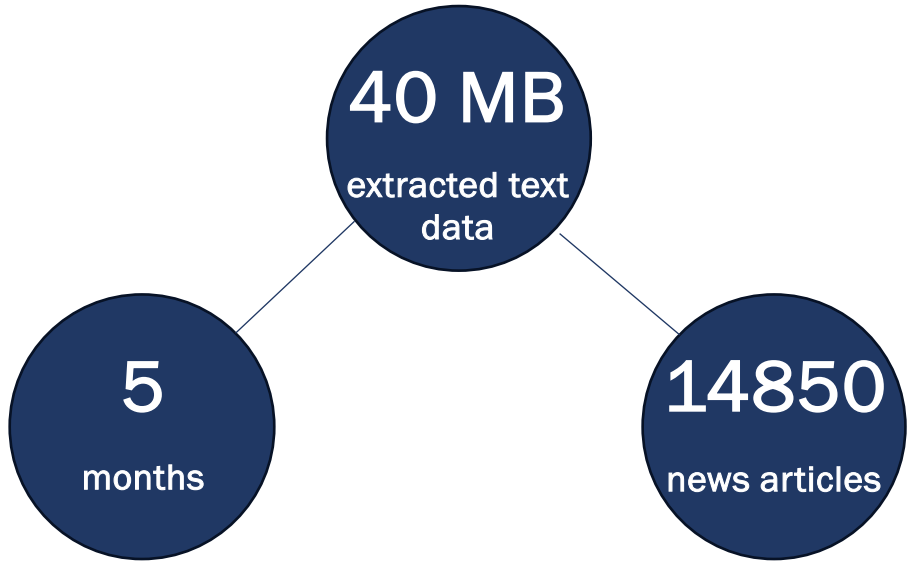
$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(predicted_i - actual_i)^2}{N}}$$

$$\vec{v} = \begin{pmatrix} \text{political instability} \\ \text{geopolitical factors} \\ \text{currency fluctuations} \\ \text{investment demand} \\ \text{supply and demand} \\ \text{industrial demand} \\ \text{natural disasters} \end{pmatrix}$$

Web Scraping like a Pro

Problem:
Getting blocked after scraping a couple of articles

Solution:
Imitate a Google web crawler bot
to avoid getting blocked

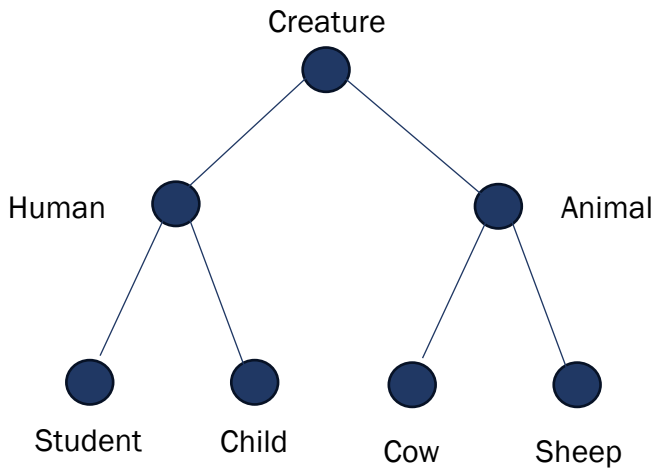



```
User-Agent:
"Mozilla/5.0 AppleWebKit/537.36
(KHTML, like Gecko; compatible;
Googlebot/2.1;
+http://www.google.com/bot.html)"
```

Deep Dive Feature Extraction

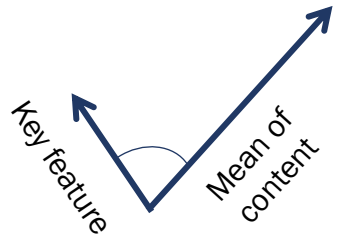
WordNet

1. Tokenize title of article
2. Calculate similarity to key features



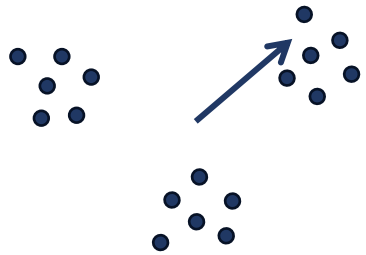
Glove/Sentence-Transformer/Doc2Vec

1. Vectorize words, sentences or documents
2. For each key feature:
   - Calculate cosine similarity of key features and vectorized entries



BERTopic

1. With BERTopic:
   1. Create document embedding
   2. Dimensionality reduction
   3. Cluster with HDBSCAN
2. Calculate similarities between embeddings of key features and document embedding



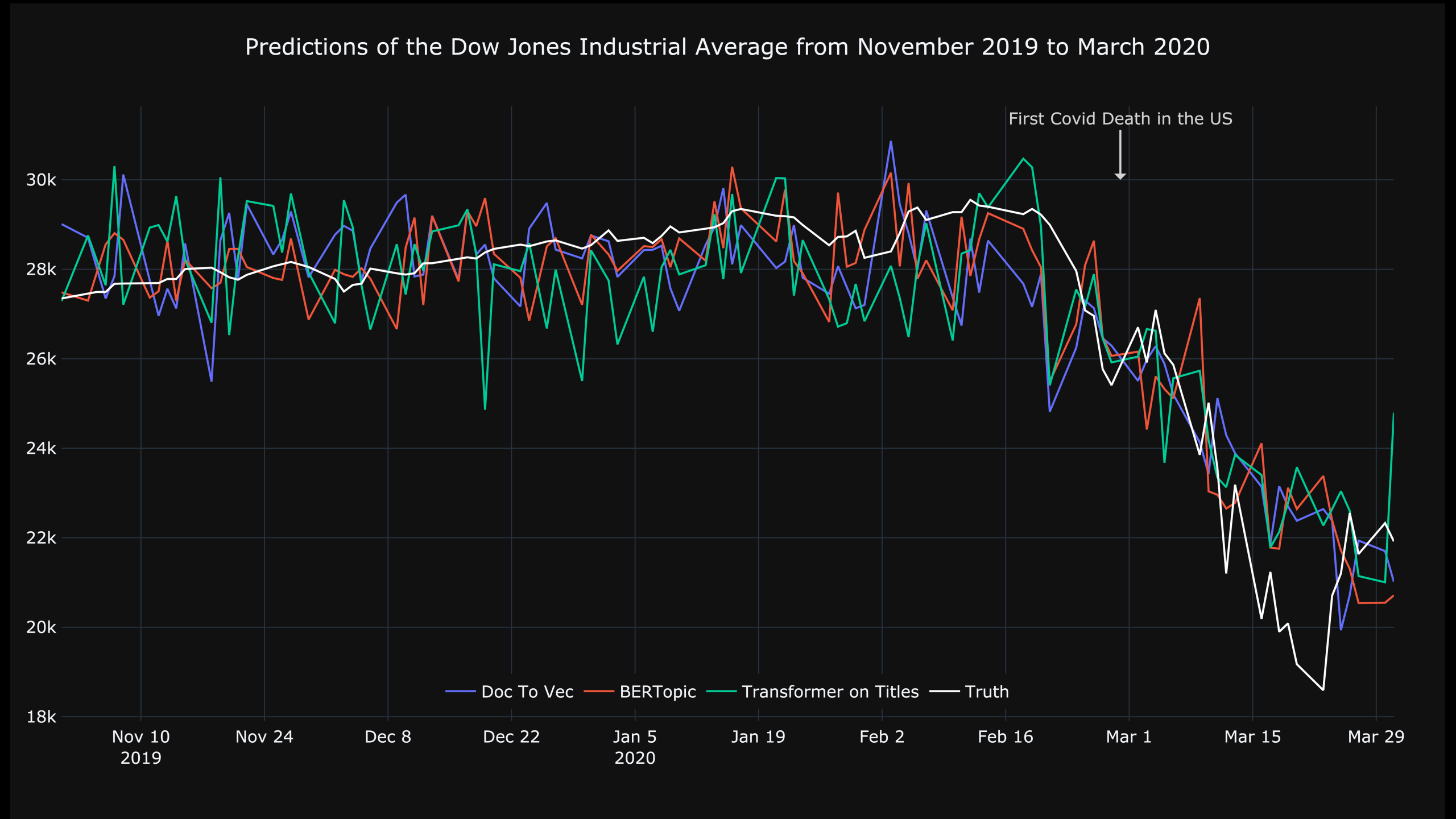
Evaluation on Linear Regression Model

Ranking	Model	Idea	RMSE
1	**BERTopic**	Unsupervised topic generation by clustering similar document and similarity calculation	1.555.377
2	**Doc2Vec** (self-trained)	Vectorize news articles / headlines / nouns and verbs in newspaper articles and calculate cosine similarity	1.910.070
3	**Sentence-BERT** (pretrained)		2.273.107
4	**Glove** (pretrained)		2.947.910
5	**WordNet**	Find shortest path between two words in the WordNet tree	3.690.413

1 Reimers and Gurevych, 2019, [1908.10084] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (arxiv.org)

2 Le and Mikolov, 2014
3 Grootendorst, 2022
4 Pennington et al., 2014

5 Pedersen, Patwardhan and Michelizzi, 2004, WordNet::Similarity - Measuring the Relatedness of Concepts


Predictions of the Dow Jones Industrial Average from November 2019 to March 2020
First Covid Death in the US
30k
28k
26k
24k
22k
20k
18k
Nov 10
2019
Nov 24
Dec 8
Dec 22
Jan 5
2020
Jan 19
Feb 2
Feb 16
Mar 1
Mar 15
Mar 29
Doc To Vec
BERTopic
Transformer on Titles
Truth

Questions? Discussion!