Sentiment Analysis on Mastodon Posts

Predicting Election Outcomes with Elixir?

Sebastian Heiden

Harz University of Applied Sciences,
Student: M.Sc. Data Science, 3rd Semester

February 8, 2024

About Me

Working Life

topics:

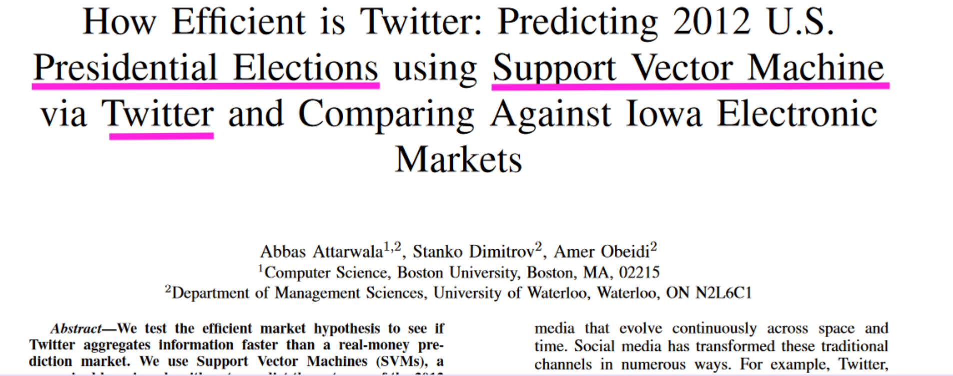
- heat demand and PV cadastres

using:

- geodata (Raster, Vector)
- Python: GeoPandas, Numpy
- land usage; coverage, property register

Privat Life

Election Monitoring with on Twitter[1]



How Efficient is Twitter: Predicting 2012 U.S. Presidential Elections using Support Vector Machine via Twitter and Comparing Against Iowa Electronic Markets

Abbas Attarwala[1,2], Stanko Dimitrov[2], Amer Obeidi[2]
[1]Computer Science, Boston University, Boston, MA, 02215
[2]Department of Management Sciences, University of Waterloo, Waterloo, ON N2L6C1

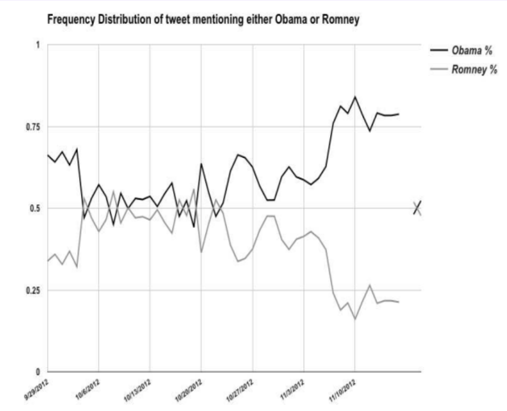
***Abstract*—We test the efficient market hypothesis to see if Twitter aggregates information faster than a real-money prediction market. We use Support Vector Machines (SVMs), a**

media that evolve continuously across space and time. Social media has transformed these traditional channels in numerous ways. For example, Twitter,

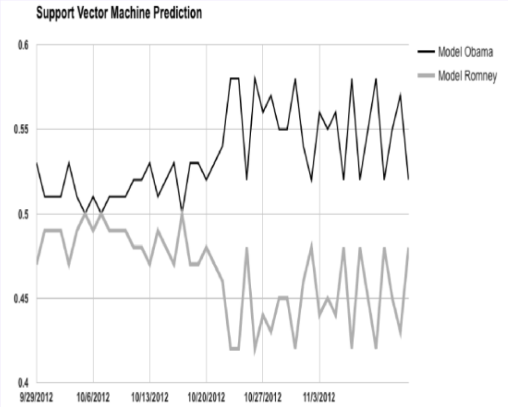
[1] Attarwala, A. *et al.* *How efficient is Twitter: Predicting 2012 U.S. presidential elections using Support Vector Machine via Twitter and comparing against Iowa Electronic Markets.* in *2017 Intelligent Systems Conference (IntelliSys)* 2017 Intelligent Systems Conference (IntelliSys) (Sept. 2017), 646–652. https://ieeexplore.ieee.org/document/8324363 (2024).

Prediction

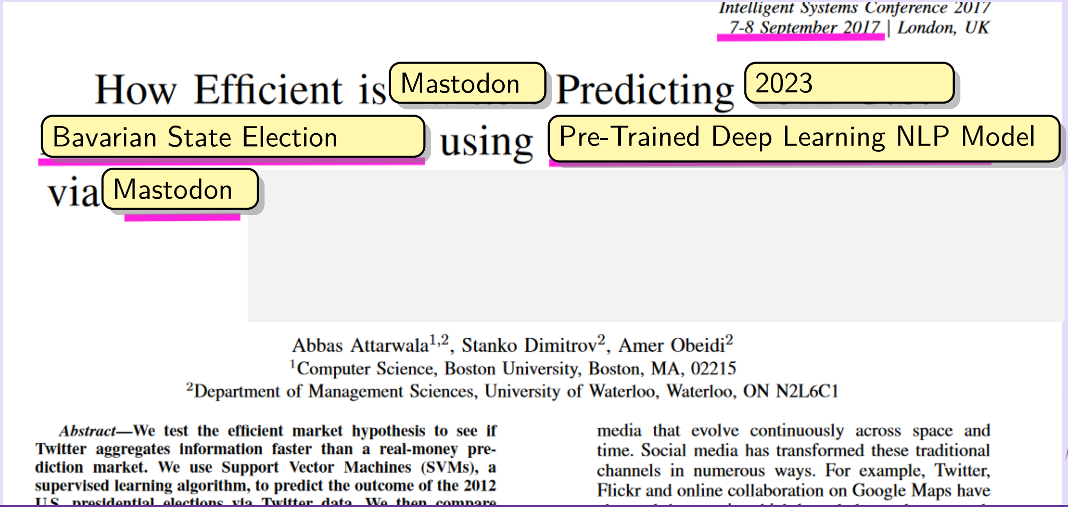
Frequency



Prediction



Election on Mastodon



Intelligent Systems Conference 2017
7-8 September 2017 | London, UK

How Efficient is Mastodon Predicting 2023 Bavarian State Election using Pre-Trained Deep Learning NLP Model via Mastodon

Abbas Attarwala[1,2], Stanko Dimitrov[2], Amer Obeidi[2]
[1]Computer Science, Boston University, Boston, MA, 02215
[2]Department of Management Sciences, University of Waterloo, Waterloo, ON N2L6C1

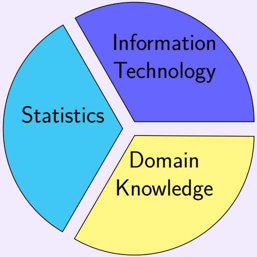
***Abstract*—We test the efficient market hypothesis to see if Twitter aggregates information faster than a real-money prediction market. We use Support Vector Machines (SVMs), a supervised learning algorithm, to predict the outcome of the 2012 U.S. presidential elections via Twitter data. We then compare**

media that evolve continuously across space and time. Social media has transformed these traditional channels in numerous ways. For example, Twitter, Flickr and online collaboration on Google Maps have

Data Science & Big Data [2]

DS Perspective

Statistics
- Modelling
- Model evaluation
- Causality
vs. Correlation



Information Technology
- Data preparation
- Data processing
- Implementation
- Algorithms

Domain Knowledge
- Business Practice
- Economic Value
- Communication
- Practical Implementation

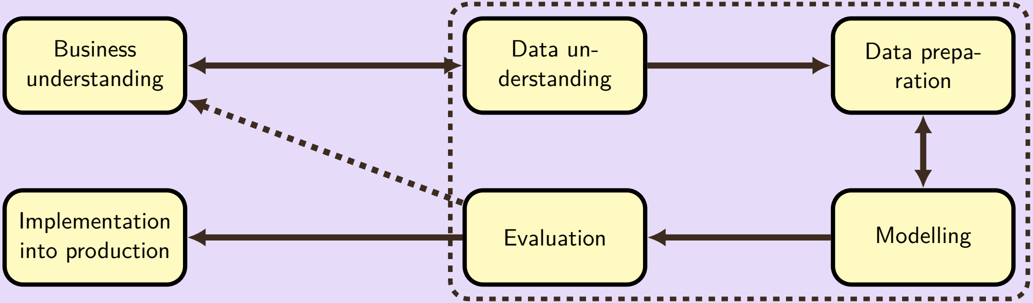
Figure: High Level

Big Data

- ▶ Volume
- ▶ Velocity
- ▶ Variety
- ▶ Veracity
- ▶ Value
- ▶ Validity

[2] Courtesy Prof. F. Transchel, Harz University of Applied Sciences.

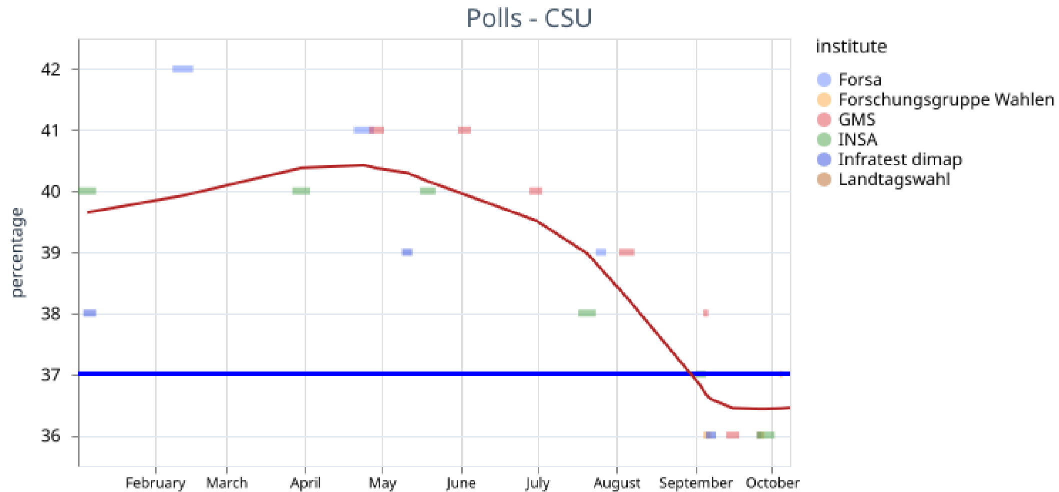
CRISP-DM



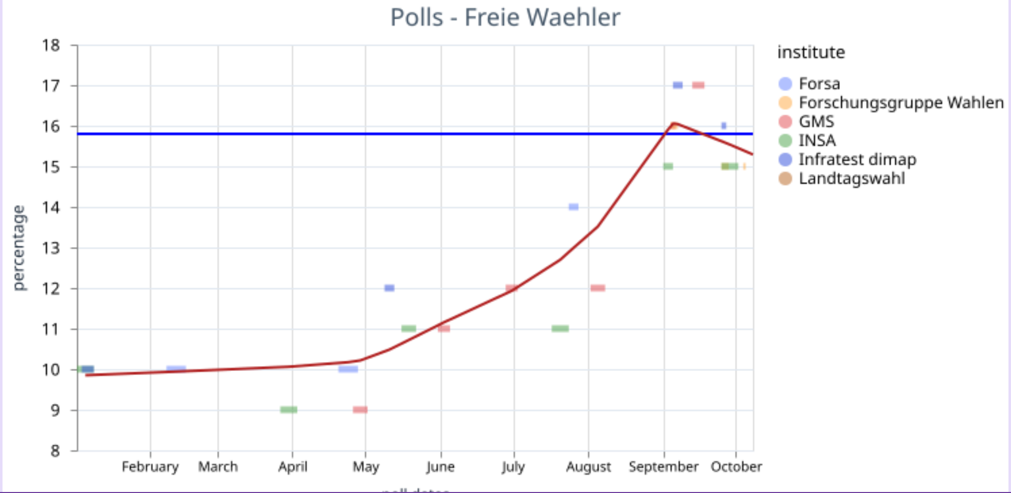
Source: P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth (2000); CRISP-DM 1.0 Step-by-step data mining guides

Figure: CRISP-DM: Cross Industry Standard Process for Data Mining

Polls CSU


Polls - CSU
percentage
42
41
40
39
38
37
36
February
March
April
May
June
July
August
September
October
institute
● Forsa
● Forschungsgruppe Wahlen
● GMS
● INSA
● Infratest dimap
● Landtagswahl

Polls Freie Waehler



Research Question

Predict the voting result of the 2023 Bavarian State election with Mastodon.

- time period: six weeks before, 4 weeks after election
- differentiate Bavarian from other users
- sample size
- selection bias: socio-economics, gender, age
- what about X|Twitter?

Data Collection

Endpoints

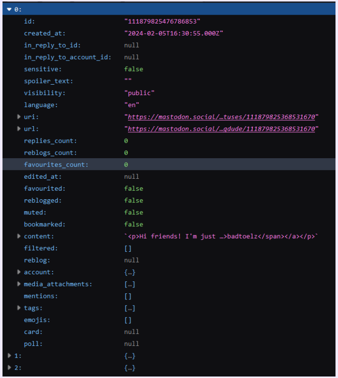
Tags: `{{instance_url}}/api/v1/timelines/tag/{{tag_name}}`

- used public timeline

Search: `{{instance_url}}\/api\/v2\/search?q={{search_word}}`

- opt-in
- log-in
- finished role out 2 days before election

Tags



```
▼ 0:
    id:                        "111879825476786853"
    created_at:                "2024-02-05T16:30:55.000Z"
    in_reply_to_id:            null
    in_reply_to_account_id:    null
    sensitive:                 false
    spoiler_text:              ""
    visibility:                "public"
    language:                  "en"
  ▶ uri:                       "https://mastodon.social/…tuses/111879825368531670"
  ▶ url:                       "https://mastodon.social/…gdude/111879825368531670"
    replies_count:             0
    reblogs_count:             0
    favourites_count:          0
    edited_at:                 null
    favourited:                false
    reblogged:                 false
    muted:                     false
    bookmarked:                false
  ▶ content:                   `<p>Hi friends! I'm just …>badtoelz</span></a></p>`
    filtered:                  []
    reblog:                    null
  ▶ account:                   {…}
  ▶ media_attachments:         […]
    mentions:                  []
  ▶ tags:                      […]
    emojis:                    []
    card:                      null
    poll:                      null
▶ 1:                           {…}
▶ 2:                           {…}
```

Data Understanding

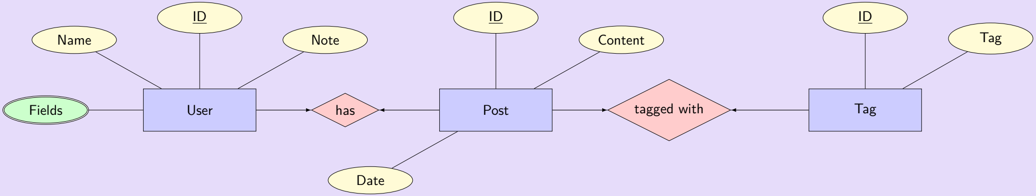
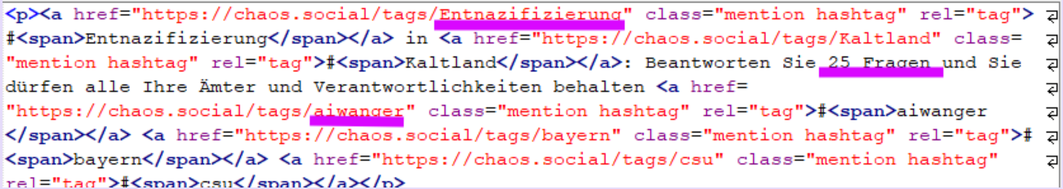


Figure: ER Diagram

Data Understanding 2



```
<p><a href="https://chaos.social/tags/Entnazifizierung" class="mention hashtag" rel="tag">
#<span>Entnazifizierung</span></a> in <a href="https://chaos.social/tags/Kaltland" class=
"mention hashtag" rel="tag">#<span>Kaltland</span></a>: Beantworten Sie 25 Fragen und Sie
dürfen alle Ihre Ämter und Verantwortlichkeiten behalten <a href=
"https://chaos.social/tags/aiwanger" class="mention hashtag" rel="tag">#<span>aiwanger
</span></a> <a href="https://chaos.social/tags/bayern" class="mention hashtag" rel="tag">#
<span>bayern</span></a> <a href="https://chaos.social/tags/csu" class="mention hashtag"
rel="tag">#<span>csu</span></a></p>
```

Figure: Example Post

Data Cleaning

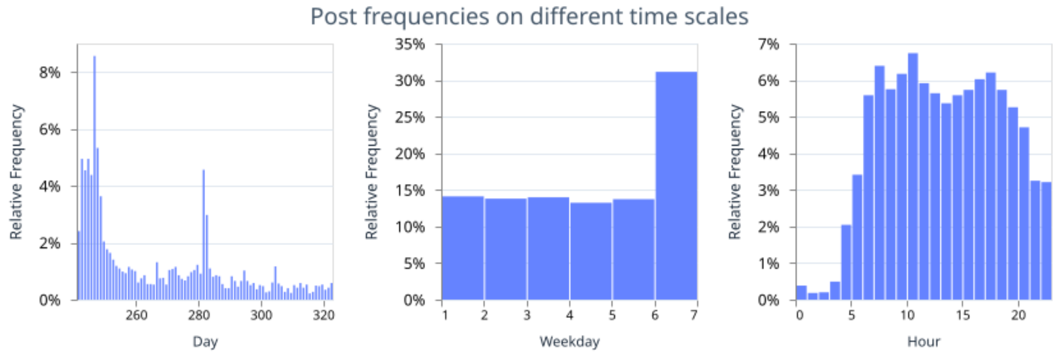
Text Cleaning

- html tags
- links
- special characters
- double spaces

Post Selection

- regional filter
  - name local entity
  - name any candidate
- party attribution filter
  - single party in post
  - party highest frequency in post
- text length

Post Frequencies



Region Classification

Text Cleaning

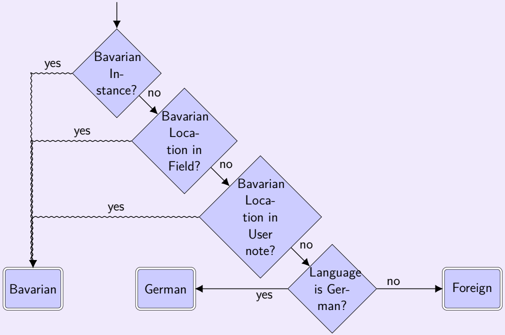


Figure: Classification: Bavarian

Post Selection

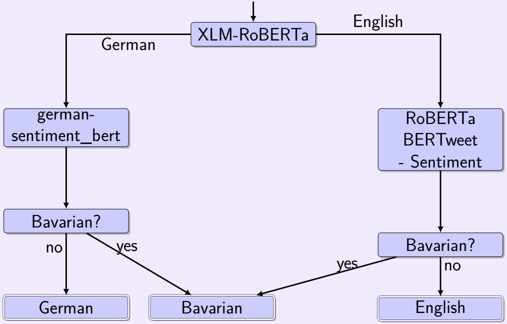
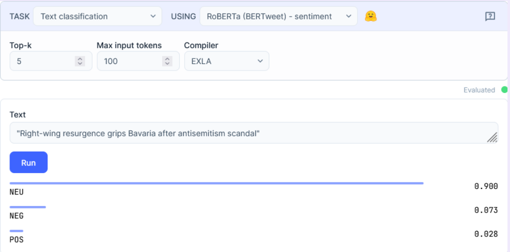


Figure: Language Classification

Smart Cells

Smart Cell



TASK Text classification USING RoBERTa (BERTweet) - sentiment

Top-k: 5

Max input tokens: 100

Compiler: EXLA

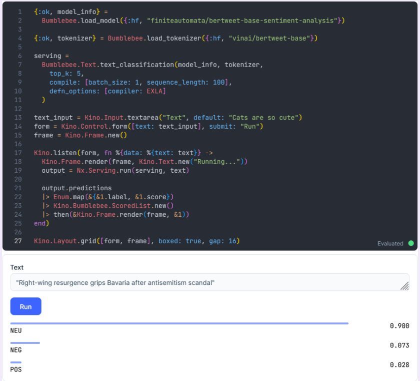
Evaluated

Text

"Right-wing resurgence grips Bavaria after antisemitism scandal"

```
NEU    0.900
NEG    0.073
POS    0.028
```

Code



```elixir
{:ok, model_info} =
  Bumblebee.load_model({:hf, "finiteautomata/bertweet-base-sentiment-analysis"})

{:ok, tokenizer} = Bumblebee.load_tokenizer({:hf, "vinai/bertweet-base"})

serving =
  Bumblebee.Text.text_classification(model_info, tokenizer,
    top_k: 5,
    compile: [batch_size: 1, sequence_length: 100],
    defn_options: [compiler: EXLA]
  )

text_input = Kino.Input.textarea("Text", default: "Cats are so cute")
form = Kino.Control.form([text: text_input], submit: "Run")
frame = Kino.Frame.new()

Kino.listen(form, fn %{data: %{text: text}} ->
  Kino.Frame.render(frame, Kino.Text.new("Running..."))
  output = Nx.Serving.run(serving, text)

  output.predictions
  |> Enum.map(&{&1.label, &1.score})
  |> Kino.Bumblebee.ScoredList.new()
  |> then(&Kino.Frame.render(frame, &1))
end)

Kino.Layout.grid([form, frame], boxed: true, gap: 16)
```

Evaluated

Text

"Right-wing resurgence grips Bavaria after antisemitism scandal"

```
NEU    0.900
NEG    0.073
POS    0.028
```

Sentiment Analysis English[3]

```
{:ok, model_info} = Bumblebee.load_model({:hf,
↪  "finiteautomata/bertweet-base-sentiment-analysis"})
{:ok, tokenizer} = Bumblebee.load_tokenizer({:hf, "vinai/bertweet-base"})

english_sentiment_serving = Bumblebee.Text.text_classification(model_info, tokenizer,
↪  compile: [batch_size: 128, sequence_length: 130], defn_options: [compiler: EXLA])

Kino.start_child({Nx.Serving, serving: english_sentiment_serving, name: EngSentimentServer})

english_toots_df = DF.filter(single_party_toots_df, detected_languages == "en")
english_toots = S.to_list(english_toots_df["cleared_content"])
eng_predictions = Nx.Serving.batched_run(EngSentimentServer, english_toots)
```

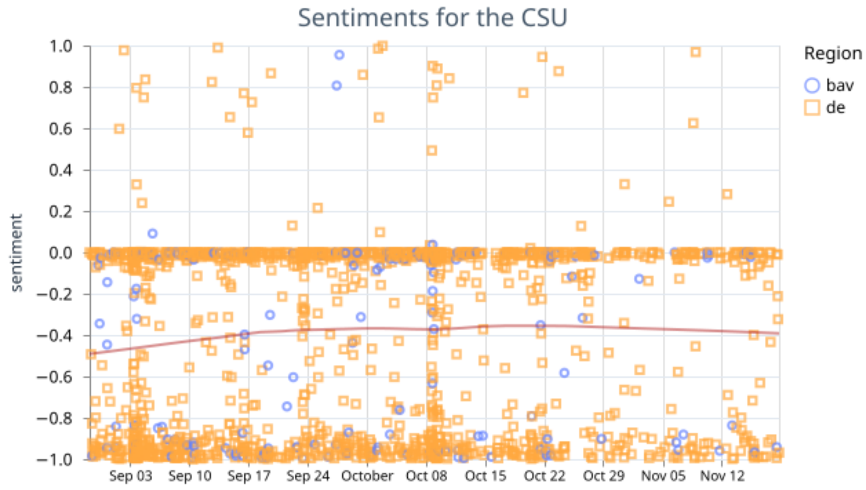
[3] Pérez, J. M. *et al.* *pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks.* Oct. 25, 2023. arXiv: 2106.09462[cs]. http://arxiv.org/abs/2106.09462 (2024).

Sentiment Analysis German[4]

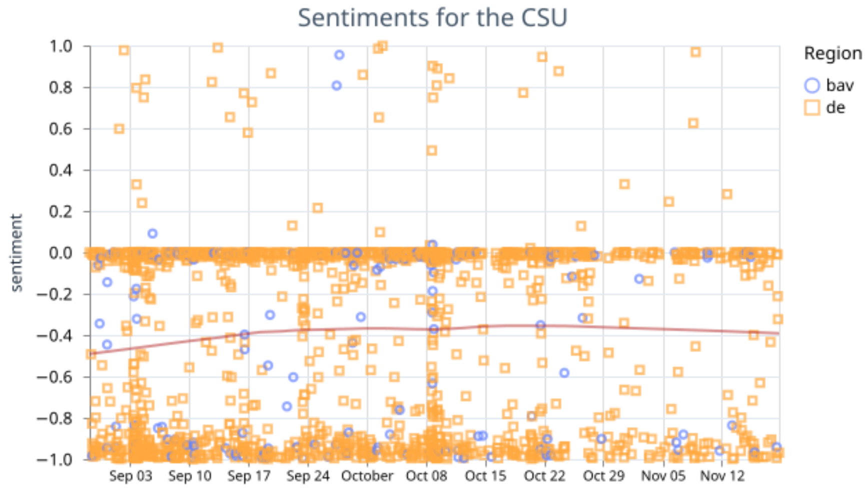
```
{:ok, ger_sent_model_info} = Bumblebee.load_model({:hf, "oliverguhr/german-sentiment-bert"})
{:ok, ger_sent_tokenizer} = Bumblebee.load_tokenizer({:hf, "bert-base-german-cased"})

ger_sent_serving = Bumblebee.Text.text_classification(ger_sent_model_info,
↪  ger_sent_tokenizer, compile: [batch_size: 128, sequence_length: 512], defn_options:
↪  [compiler: EXLA])

Kino.start_child({Nx.Serving, serving: ger_sent_serving, name: GerSentimentServer})

german_toots_df = DF.filter(single_party_toots_df, detected_languages == "de")
german_toots = S.to_list(german_toots_df["cleared_content"])
ger_predictions = Nx.Serving.batched_run(GerSentimentServer, german_toots)
```

[4] Guhr, O. *et al.* *Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems.* in *Proceedings of the Twelfth Language Resources and Evaluation Conference* LREC 2020 (eds Calzolari, N. *et al.*) (European Language Resources Association, Marseille, France, May 2020), 1627–1632. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.202 (2024).

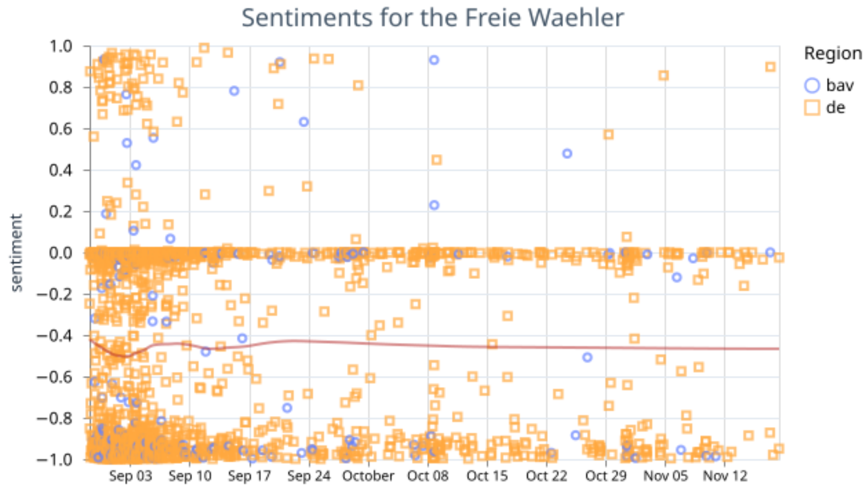
Sentiment CSU



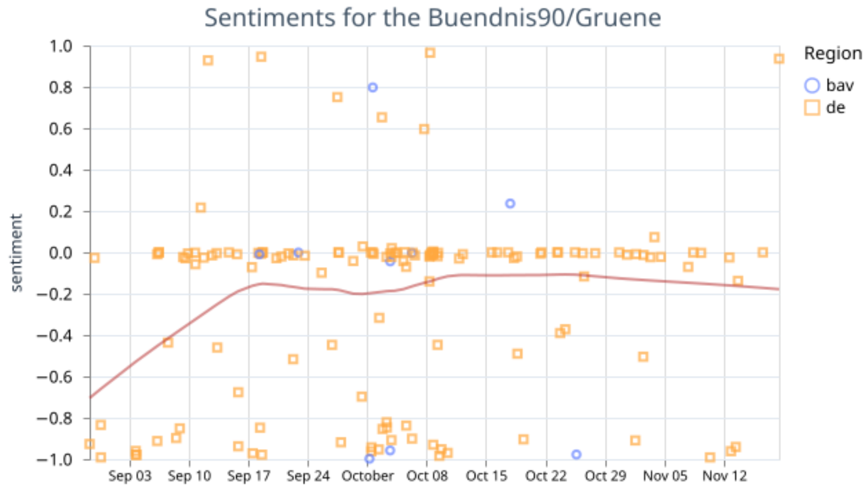
Sentiment CSU



Sentiment Frei Waehler



Sentiment Buendnis 90/Gruene



Frequency

Table: Frequencies how often the parties are mentioned.

Party	Mentioned	Mentioned Bavaria	Election
AFD	11.7 %	11.0 %	14.6 %
CSU	30.7 %	32.6 %	37.0 %
FDP	1.9 %	1.3 %	3.0 %
FW	47.9 %	49.1 %	15.8 %
Gruene	3.0 %	1.8 %	14.4 %
Linke	1.3 %	1.8 %	1.5 %
SPD	3.7 %	2.1 %	8.4 %

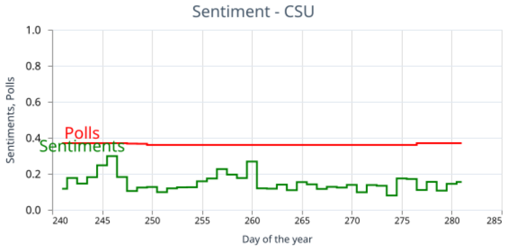
Frequency Enhanced

Table: Frequencies of mentions, after Sept. 17th and most positive post per author.

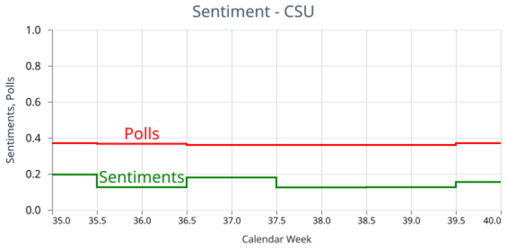
Party	Mentioned	Mentioned Bavaria	Election
AFD	18.7 %	16.1 %	14.6 %
CSU	45.9 %	38.1 %	37.0 %
FDP	1.0 %	n/a	3.0 %
FW	22.0 %	14.1 %	15.8 %
Gruene	6.4 %	8.1 %	14.4 %
Linke	1.4 %	1.6 %	1.5 %
SPD	4.6 %	4.8 %	8.4 %

Timeline CSU

Daily

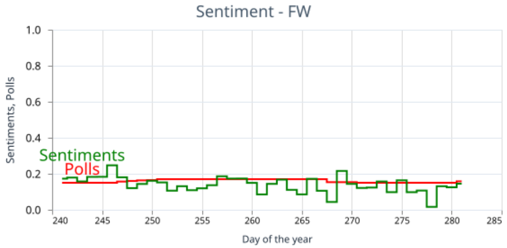


Weekly

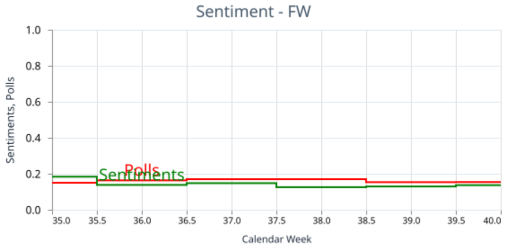


Timeline Freie Waehler

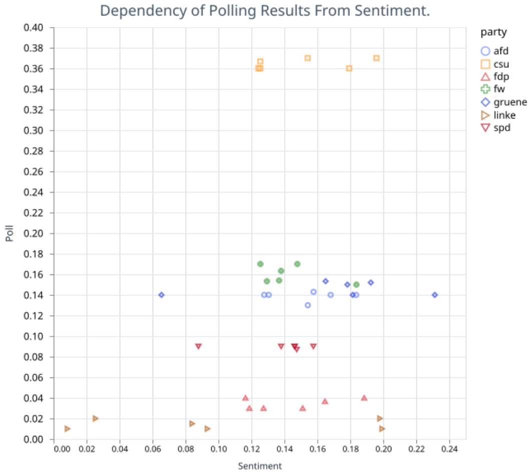
Daily



Weekly



Sentiment vs Polls



Ease of Use

The Good

- It's Elixir
- BumbleBee (Hugging Face)
- Livebook > Jupyter Notebook
- It's Possible

The Enhancing

- Help from Forum is Great:
  - Released it last week.
  - It's on Github, not in hexdocs, yet.
- Graphics: Tucan vs. Vega Lite.
- Scholar (ML) not as complete .
- A lot of Progress in most libraries.