analyse

June 27, 2024

1 Question

One Day during lunch my commiilitons found out that both recognized the change in languagelevel in the news during the EM. The Theory behind it was that news are now made in a way that drunk (stupid) “footballpeople” can also understand them.

Will the wording of the Tagesschau news bulletins broadcast during the European Championships be formulated in simpler language?

2 Method

To answer the question, YouTube videos from the last few months are first scraped and then automatically labeled using readability measurements ála Flesch.

3 Analyse

[ ]:
```
#pip install pandas matplotlib textstat scipy seaborn
```

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import textstat
from scipy import stats
import seaborn as sns
import re
```

[ ]:
```
df = pd.read_csv('data/data.csv')

if 'ID' not in df.columns:
    df['ID'] = range(1, len(df) + 1)

df.head()
```

[ ]:

	ID	url	path	\
0	1	https://www.youtube.com/watch?v=pWuYkzRypf8	audio/1_tagesschau.mp3	
1	2	https://www.youtube.com/watch?v=MAmyhxJoZFM	audio/2_tagesschau.mp3	
2	3	https://www.youtube.com/watch?v=AVqitc2vJ0c	audio/3_tagesschau.mp3	
3	4	https://www.youtube.com/watch?v=Nmwe8gXfu1c	audio/4_tagesschau.mp3	

```
4   5  https://www.youtube.com/watch?v=-okNLhbCcHc  audio/5_tagesschau.mp3

                                                 text  \
0   Hier ist das erste deutsche Fernsehen mit der…
1   Hier ist das erste deutsche Fernsehen mit der…
2   Hier ist das erste deutsche Fernsehen mit der…
3   Hier ist das erste deutsche Fernsehen mit der…
4   Hier ist das erste deutsche Fernsehen mit der…

                              title        date
0  tagesschau 20:00 Uhr, 25.06.2024  2024-06-25
1  tagesschau 20:00 Uhr, 24.06.2024  2024-06-24
2  tagesschau 20:00 Uhr, 23.06.2024  2024-06-23
3  tagesschau 20:00 Uhr, 22.06.2024  2024-06-22
4  tagesschau 20:00 Uhr, 20.06.2024  2024-06-20
```

[ ]:
```python
def extract_date(title):
    date_pattern = re.compile(r'\b\d{1,2}\.\d{1,2}\.\d{2,4}\b')
    match = date_pattern.search(title)
    return match.group(0) if match else None

def convert_date(date_str):
    for fmt in ('%d.%m.%Y', '%d.%m.%y'):
        try:
            return pd.to_datetime(date_str, format=fmt)
        except ValueError:
            continue
    return None

df['date'] = df['title'].apply(extract_date)
df['date'] = df['date'].apply(convert_date)

df.head()
```

[ ]:
```
   ID                                          url                    path  \
0   1  https://www.youtube.com/watch?v=pWuYkzRypf8  audio/1_tagesschau.mp3
1   2  https://www.youtube.com/watch?v=MAmyhxJoZFM  audio/2_tagesschau.mp3
2   3  https://www.youtube.com/watch?v=AVqitc2vJ0c  audio/3_tagesschau.mp3
3   4  https://www.youtube.com/watch?v=Nmwe8gXfu1c  audio/4_tagesschau.mp3
4   5  https://www.youtube.com/watch?v=-okNLhbCcHc  audio/5_tagesschau.mp3

                                                 text  \
0   Hier ist das erste deutsche Fernsehen mit der…
1   Hier ist das erste deutsche Fernsehen mit der…
2   Hier ist das erste deutsche Fernsehen mit der…
3   Hier ist das erste deutsche Fernsehen mit der…
4   Hier ist das erste deutsche Fernsehen mit der…
```

```
                                title       date
0  tagesschau 20:00 Uhr, 25.06.2024 2024-06-25
1  tagesschau 20:00 Uhr, 24.06.2024 2024-06-24
2  tagesschau 20:00 Uhr, 23.06.2024 2024-06-23
3  tagesschau 20:00 Uhr, 22.06.2024 2024-06-22
4  tagesschau 20:00 Uhr, 20.06.2024 2024-06-20
```

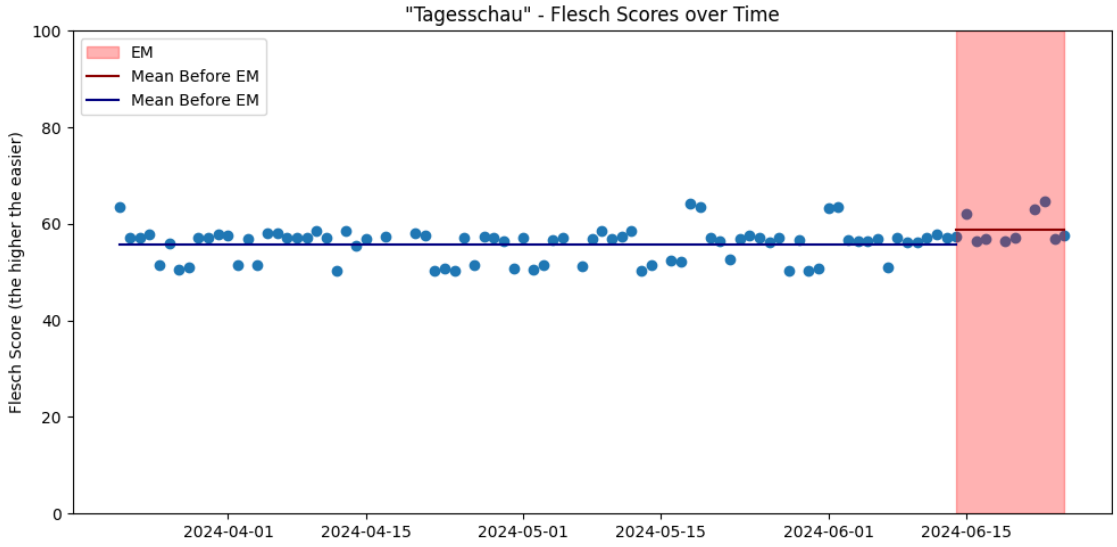
[ ]:
```
textstat.set_lang("de")
df['flesch'] = df.apply(lambda row: textstat.flesch_reading_ease(row['text']),␣
 ↪axis=1)
df.head()
```

[ ]:
```
   ID                                          url                    path  \
0   1  https://www.youtube.com/watch?v=pWuYkzRypf8  audio/1_tagesschau.mp3
1   2  https://www.youtube.com/watch?v=MAmyhxJoZFM  audio/2_tagesschau.mp3
2   3  https://www.youtube.com/watch?v=AVqitc2vJ0c  audio/3_tagesschau.mp3
3   4  https://www.youtube.com/watch?v=Nmwe8gXfu1c  audio/4_tagesschau.mp3
4   5  https://www.youtube.com/watch?v=-okNLhbCcHc  audio/5_tagesschau.mp3

                                                text  \
0   Hier ist das erste deutsche Fernsehen mit der…
1   Hier ist das erste deutsche Fernsehen mit der…
2   Hier ist das erste deutsche Fernsehen mit der…
3   Hier ist das erste deutsche Fernsehen mit der…
4   Hier ist das erste deutsche Fernsehen mit der…

                                title       date  flesch
0  tagesschau 20:00 Uhr, 25.06.2024 2024-06-25   57.65
1  tagesschau 20:00 Uhr, 24.06.2024 2024-06-24   56.85
2  tagesschau 20:00 Uhr, 23.06.2024 2024-06-23   64.60
3  tagesschau 20:00 Uhr, 22.06.2024 2024-06-22   63.00
4  tagesschau 20:00 Uhr, 20.06.2024 2024-06-20   57.25
```

Die EM startete am 14. Juni.

Score	Difficulty
90-100	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
0-29	Very Confusing

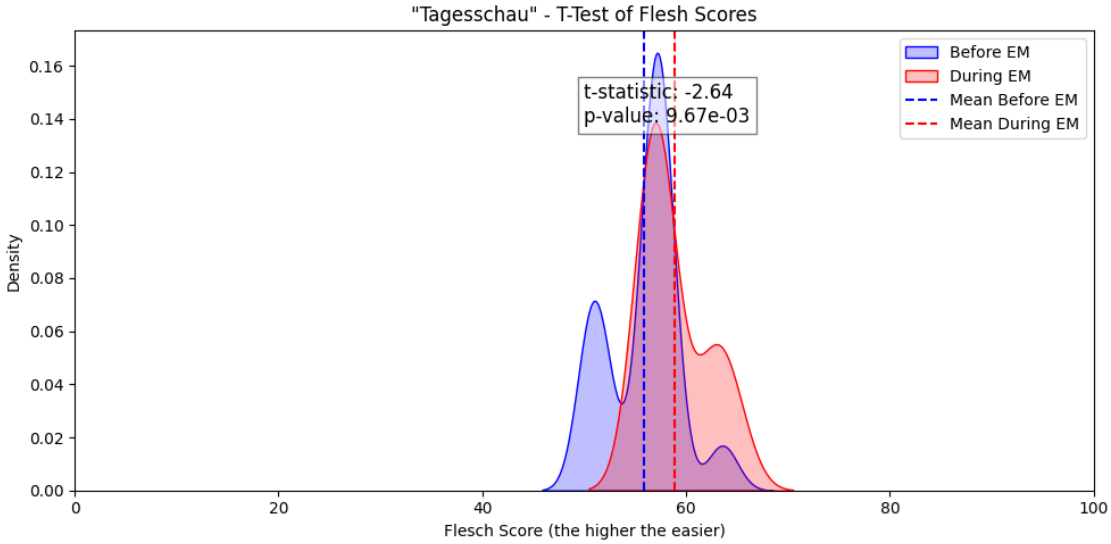
```
[ ]: df['date'] = pd.to_datetime(df['date'], errors='coerce')
     df = df.dropna(subset=['date'])

     start_date = pd.to_datetime("2024-06-14")
     end_date = pd.to_datetime(max(df['date']))


     mean_before_em = df[df['date'] < start_date]['flesch'].mean()
     mean_during_em = df[(df['date'] >= start_date) & (df['date'] <=␣
      ↪end_date)]['flesch'].mean()


     # Plotting
     plt.figure(figsize=(10, 5))
     plt.scatter(df['date'], df['flesch'], marker='o')
     plt.axvspan(start_date, end_date, color='red', alpha=0.3, label='EM')

     plt.plot([start_date, end_date], [mean_during_em, mean_during_em],␣
      ↪color='darkred', linestyle='-', label='Mean Before EM')
     plt.plot([pd.to_datetime(min(df['date'])), start_date], [mean_before_em,␣
      ↪mean_before_em], color='navy', linestyle='-', label='Mean Before EM')


     plt.ylabel('Flesch Score (the higher the easier)')
     plt.title('"Tagesschau" - Flesch Scores over Time')
     plt.grid(False)
     plt.ylim(0, 100)
     plt.tight_layout()
     plt.legend()

     plt.show()
```



3.1 T-Test

[ ]:
```
before_em = df[df['date'] < pd.to_datetime("2024-06-14")]['flesch']
during_em = df[(df['date'] >= pd.to_datetime("2024-06-14")) & (df['date'] <= pd.
 ↪to_datetime(max(df['date'])))]['flesch']

t_stat, p_value = stats.ttest_ind(before_em, during_em)
```

[ ]:
```
plt.figure(figsize=(10, 5))
sns.kdeplot(before_em, label='Before EM', color='blue', fill=True)
sns.kdeplot(during_em, label='During EM', color='red', fill=True)
plt.axvline(before_em.mean(), color='blue', linestyle='--', label='Mean Before␣
 ↪EM')
plt.axvline(during_em.mean(), color='red', linestyle='--', label='Mean During␣
 ↪EM')

plt.annotate(f't-statistic: {t_stat:.2f}\np-value: {p_value:.2e}',
             xy=(0.5, 0.8), xycoords='axes fraction', fontsize=12,
             bbox=dict(facecolor='white', alpha=0.5))
plt.xlim(0, 100)

plt.xlabel('Flesch Score (the higher the easier)')
plt.ylabel('Density')
plt.title('"Tagesschau" - T-Test of Flesh Scores')
plt.grid(False)

plt.legend()
plt.tight_layout()

plt.show()
```



The densityplot shows that the difference in means probably occurs because of the outliers, wich are also seen in the first plot.

To answer the question definitely we have to wait until a few weeks after the EM and see if the languagelevel goes *back to normal lows*.

Until then we can conclude the general impression that languagelevel changed is only partly legitimized.