JSON → pd.DataFrame

df.columns = ['url', 'date', 'title', 'paragraphs', 'figures', 'keywords','gpt_keywords', 'n_comments']

df → (pipeline) preprocessing, create features from raw data → df

df.columns['topic', 'hours', 'weekend', 'title_length', 'article_length', 'images', 'keywords', 'gpt_keywords', 'n_comments']

I chose **99th quantile** of most common words

Topics I created manually from url, and **split sport** into categories

From Date → **hours**_of_the_day, is_it_**weekend**

number of words in title, number of paragraphs in article, number of **images**

df → pipeline: one hot encoding, MultiLabelBinarizer -> df

['topic_crna-kronika', 'topic_gospodarstvo', 'topic_kolumne',
'topic_kultura', 'topic_okolje', 'topic_slovenija',
'topic_sport-atletika', 'topic_sport-citat-za-prebrat',
'topic_sport-dokovic-petic-izbran-za-najboljsega-v-evropi-doncic-na-39-mestu',
'topic_sport-formula-1',
...
'teroristični napad', 'Wimbledon', 'javno mnenje', 'preobrat',
'resolucija', 'dolgotrajna oskrba', 'title_length', 'article_length',
'images', 'n_comments'],
OneHotEncoder for all important words

X.shape = (20981, 1328)

Standardization after splitting the data into training and testing set

10 fold cross validation

Lasso (L1)

Mean Absolute Error – makes sense, by how many comments I made mistake

31.93327 is the **MAE** for **Lasso**
41.04641 is the **MAE** for **Baseline** (average)

32.98787 is the MAE for Lasso in 0th iteration
31.03187 is the MAE for Lasso in 1th iteration
30.38899 is the MAE for Lasso in 2th iteration
33.16667 is the MAE for Lasso in 3th iteration
30.97525 is the MAE for Lasso in 4th iteration
32.52818 is the MAE for Lasso in 5th iteration
31.40074 is the MAE for Lasso in 6th iteration
29.98737 is the MAE for Lasso in 7th iteration
32.18501 is the MAE for Lasso in 8th iteration
31.06465 is the MAE for Lasso in 9th iteration

With CV average score is **31.572**, which is **expected** score on new data

r2 score = 0.263

numerical features → article lentgh, number of images, title length
categorical features → hour of the day, topic, is it weekend
and categorical with words → this important words occurs in this article for every words (around 500)

1328 all together features so Lasso is necessary, and it gives positive coefs for

Positive

'Gazi': 69.06125013304744,
'Finale': 66.640987834159,
'ekipa': 53.060499174765944,
'Domen Prevc': 50.37585049172347,
'statistika': 49.162243752223404,
'oblačila': 44.732745590142116,
'model': 39.59484836654575,
'razprava': 36.912897092258746,
'hours_10': 30.687799425486297,
'tekmovalci': 29.03482047801394,
'helikopter': 28.85541684317066,
'Primož Roglič': 27.685331917063394,
'vlagatelji': 27.577599547720745,
'Irak': 25.29684727285926,
'Nato': 24.823089815732565,
'topic_sport-motosporti': 24.63243703127892,
'bolezen': 24.4854403023745,
'stroški': 23.48063575535399,

Negative

'izvoz': -21.56158988601853,
'topic_sport-atletika': -13.219868112339071,
'Las Vegas': -12.98883970827762,
'imenovanje': -7.2806735880728075,
'hours_23': -7.199392760690368,
'zmagovalci': -6.237326756275094,
'Toronto': -6.0334110072313765,
'podaja': -5.792667611690297,
'voda': -4.706384120945132,
'Tanja Fajon': -4.503339553997457,
'knjige': -4.411328299311297,

Neutral

'trgovci': -0.0,
'topic_sport-odbojka': -0.0,
'taktika': -0.0,
'stoletje': 0.0,

Score on new data is slightly worse but overall similar