

sentiment analysis of polish texts

Redko Anna

Dataset



EOSC CO-CREATION
secretariat.eu

A dataset of media releases (Twitter, News and Comments, Youtube, Facebook) from Poland related to COVID-19 for open research

15/01/2020-31/07/2020



```
"", "text", "time"  
'1", "Tez uwazam ze falszywa pandemia #koronawirus jest jak stan wojenny. Ja juz mam nerwice, czuje stres jak widze zamaskowanych. https://t.co/TtlrJ  
'2", "Podawajcie dalej! #koronawirus #KoronawirusWPolsce #zajob #wybory2020 #DUDA2020 #Trzaskowski2020 # https://t.co/9PDDfYrpFp", 2020-07-09 19:48:39  
'3", "Precz z pandemia!!! #koronawirus #koronawiruswpolsce #koronawiruspolska #zajob https://t.co/N8bTjei2zt", 2020-07-10 17:58:43  
'4", "Ludzie zra sie na ulicach przez te kanalie #szumowski #koronawirus #koronawiruswpolsce dlatego nie ide na wybory https://t.co/NRfe9MQ0xH", 2020-  
'5", "Lekarze to najwiekszy zawyd tej pseudoepidemii #koronawirus https://t.co/32BkrWrHTa", 2020-07-13 20:08:11  
'6", "Rekordowa frekwencja? Przeciez mamy WIELKA EPIDEMIE! Ludzie otworzcie oczy! Bo sami sobie przeczycie! #koronawirus #koronawiruswpolsce #zajob #  
'7", "Polacy przeciwko #zajob #falszywapandemia #koronawirus #koronawiruspolska #COVID19 #KoronawirusWPolsce #wybory2020 #Trza #Trzaskowski2020 #DUDA  
'8", "Niewykluczone, ze jednak sprzybuje pokonac wstret i oddam glos - oczywiscie na #PAD . Przekonal mnie apel @GadowskiWitold . Mam nadzieje, ze jes  
'9", "65 uzytych respiratoryw w calym kraju! Polsko otworz oczy! #koronawirus https://t.co/luXYCuZ5am", 2020-07-12 19:55:42  
'10", "Polacy przeciwko #zajob #falszywapandemia #koronawirus #koronawiruspolska #COVID19 #KoronawirusWPolsce #wybory2020 #Trza #Trzaskowski2020 #DUD
```



Dataset

coronavirus | 2020 Poland coronavirus data | Dataset library

by dtandev • Python • Version: Current • License: MIT

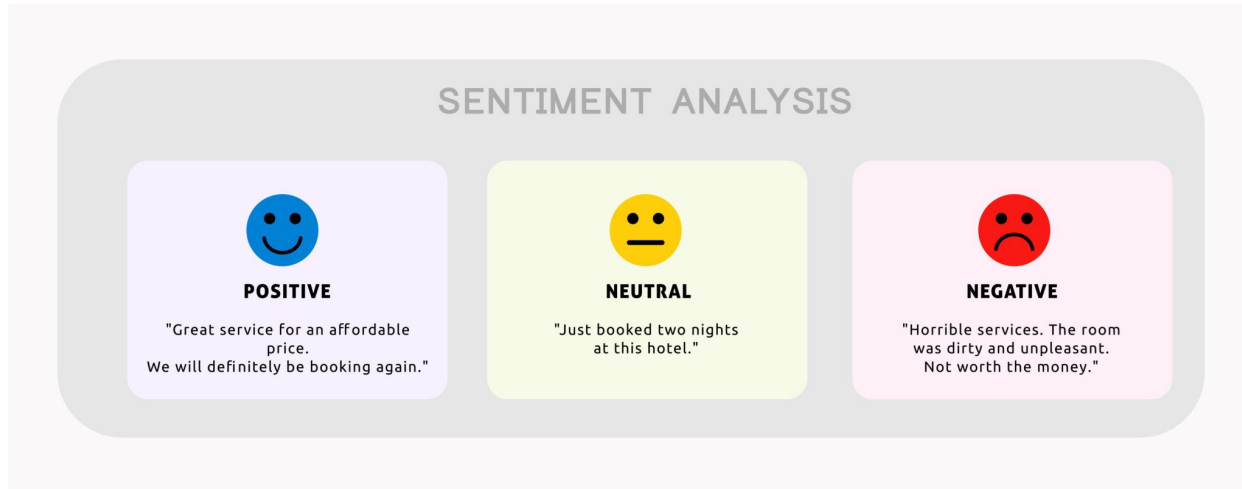
2020 Poland coronavirus data (COVID-19 / 2019-nCoV)

Timestamp	Confirmed	Deaths	Recovered	In_the_hospital	In_quarantine	Under_medical_supervision	Number_of_tests_carried_out
03-03-2020	0	0	0	68	316	4459	559
04-03-2020	1	0	0	65	349	4540	584
05-03-2020	1	0	0	92	490	5647	676
06-03-2020	5	0	0	128	1299	6184	855
07-03-2020	6	0	0	146	1548	6409	856
08-03-2020	12	0	0	168	932	7122	1154
09-03-2020	17	0	0	467	1014	7110	1384
10-03-2020	22	0	0	220	1055	9366	1630
11-03-2020	31	0	0	317	1193	11524	2024



Sentiment analysis

Sentiment analysis is the automated process of tagging data according to their sentiment, such as positive, negative and neutral. Sentiment analysis allows companies to analyze data at scale, detect insights and automate processes.



Idea

- to analyze data from social media(e.g. Twitter) in Polish that would be related to the coronavirus
- Compare results with statistics data for the:
Deaths,Recovered,In_the_hospital,In_quarantine,Under_medical_supervision

Model

twitter_sentiment_pl_base

Model is based on HerBERT (Polish version of BERT) since it receives state-of-the-art results in the area of text classification.

It was trained on the translated version of TweetEval by Barbieri et al., 2020 for 10 epochs on single RTX3090 gpu

The model will give you a three labels: positive, negative and neutral.

Model

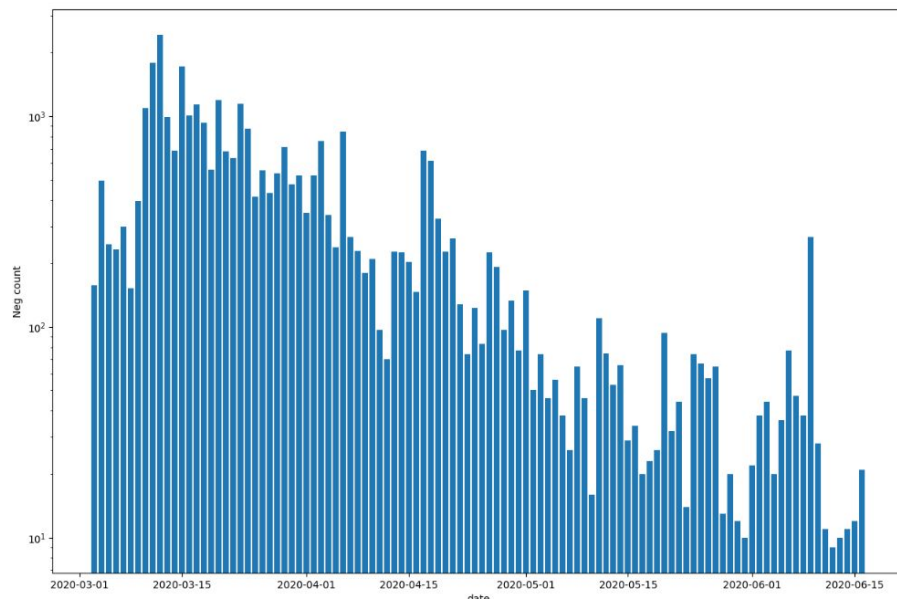
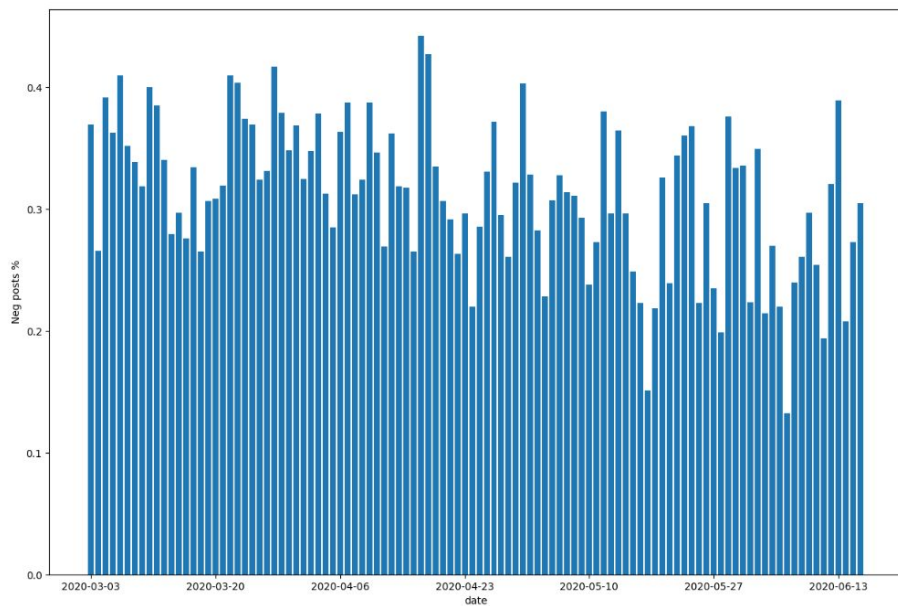
twitter-xlm-roberta-base-sentiment-finetuned

This is multilingual XLM-Roberta model sequence classifier fine tuned and based on Cardiff NLP Group sentiment classification model.

supports Polish

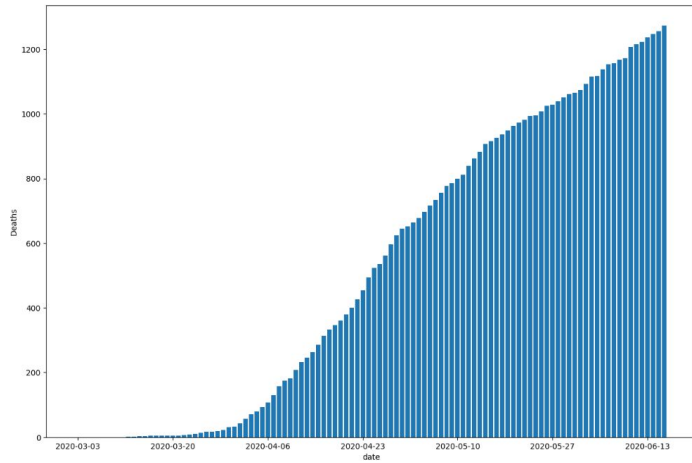
Results

Neg posts percentage / count

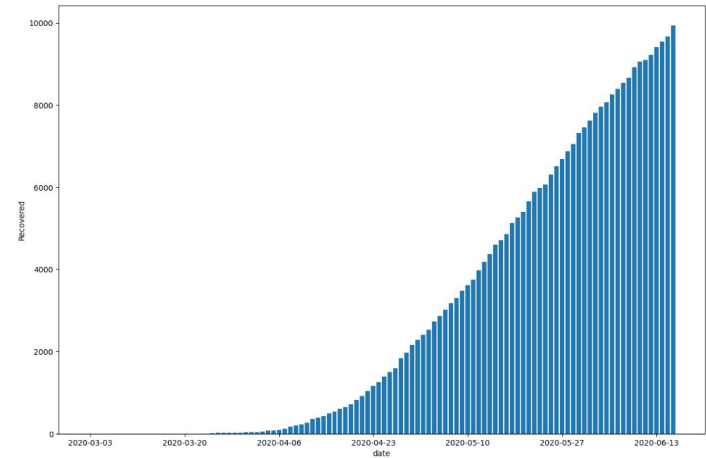


Results

Death curve

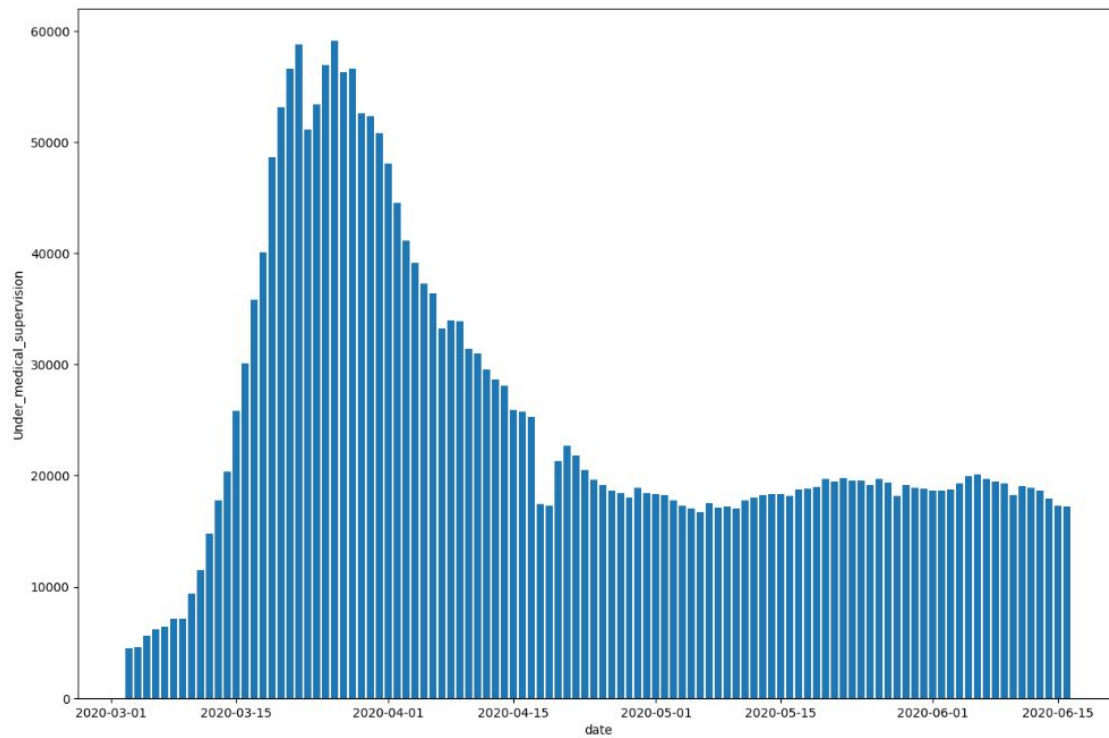


Recovered



Results

Under medical supervision



Results

Let's have a look at Pearson correlations of negative posts and dataset variable

	Recover ed	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.48	-0.50	0.25	-0.49	0.002	-0.21
neg count	-0.62	-0.74	0.46	-0.73	-0.36	-0.58

twitter_sentiment_
pl_base

	Recover ed	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.169	-0.191	0.071	-0.193	-0.117	-0.12
neg count	-0.45	-0.549	0.221	-0.55	-0.426	-0.525

twitter-
xlm-roberta-
base-sentiment-
finetuned

Let's have a look at Pearson correlations of negative posts and dataset variables

Results

sliding window smooth

	Recover ed	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.948	-0.966	0.276	-0.961	-0.37	-0.641
neg count	-0.76	-0.85	0.46	-0.85	-0.45	-0.67

twitter_sentiment_
pl_base

	Recover ed	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.527	-0.652	0.432	-0.640	-0.380	-0.574
neg count	-0.672	-0.779	0.349	-0.783	-0.538	-0.696

twitter-xlm-roberta-
base-sentiment-fin
etuned

Results

Let's have a look at Pearson correlations of negative posts and dataset variables

*twitter_sentiment_
pl_base*

sliding window
smooth

	Recover ed	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.948	-0.966	0.276	-0.961	-0.37	-0.641
neg count	-0.76	-0.85	0.46	-0.85	-0.45	-0.67

	Recover ed	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.48	-0.50	0.25	-0.49	0.002	-0.21
neg count	-0.62	-0.74	0.46	-0.73	-0.36	-0.58

Results

Let's have a look at Pearson correlations of negative posts and dataset variables

twitter-xlm-roberta-base-sentiment-finetuned

sliding window
smooth

	Recovered	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.527	-0.652	0.432	-0.640	-0.380	-0.574
neg count	-0.672	-0.779	0.349	-0.783	-0.538	-0.696

	Recovered	Deaths	Under medical supervision	Confirmed	In quarantine	In the hospital
neg %	-0.169	-0.191	0.071	-0.193	-0.117	-0.12
neg count	-0.45	-0.549	0.221	-0.55	-0.426	-0.525

Conclusion

From the conclusions, it is strange that negative posts have no correlation with the number of deaths. this is most likely due to population fatigue over time.

Thank you for your attention! Python notebook and a github blog are also provided