Laiko eilučių modelių modeliavimas panaudojant natūralios kalbos modelius

Matas Gaulia

Darbo vadovas: Linas Petkevičius
Vilniaus Universitetas
Matematikos ir Informatikos fakultetas

2023

CLIP modelis

CLIP (Contrastive Language-Image Pre-Training) modelis - "OpenAI" kompanijos sukurtas dirbtinio intelekto modelis kuris sugeba susieti vaizdus ir juos apibūdinančius tekstus.
Išleidimo data - 2021 metai
Iki tol nebuvo modelio kuris tiesiogiai susietų tekstą ir vaizdus be duomenų modifikavimo

Learning transferable visual models from natural language supervision
A Radford, JW Kim, C Hallacy… - International …, 2021 - proceedings.mlr.press
State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations …
☆ Išsaugoti 99 Cituoti Cituoja 5765 Susiję straipsniai Visos 14 versijos »

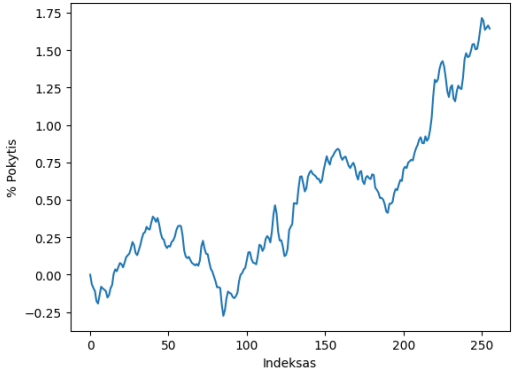
Query: one dog sitting on the grass

Hipotezė

Hipotezė

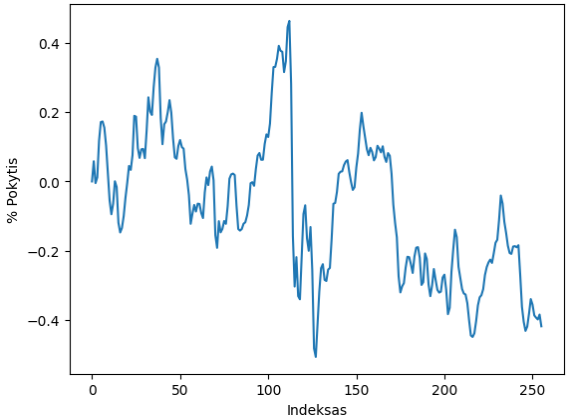
Ar galima apjungti tekstą ir laiko eilutes tiesiogiai su mašininio mokymosi modeliu?

"$BTC will go higher, I'm feeling positive!"



Hipotezė

"After my massive losses, I am selling my $BTC"



Tikslas ir uždaviniai

Tikslas: atlikti finansinių laiko eilučių ir su jomis susijusių Twitter įrašų modeliavimą, identifikuojant galimybę apjungti finansines laiko eilutes ir teksto įrašų duomenis.

Uždaviniai:

1. Kriptovaliutų kainų ir Twiter įrašų duomenų surinkimas, apdorojimas.
2. Duomenų jungimas paruošiant mokymo ir testavimo duomenų aibes.
3. Mokslinės literatūros skaitymas, analizavimas, esamo viešo kodo analizė.
4. Originalaus CLIP modelio kūrimas, parametrų vertinimas.
5. Modifikuoto CLIP modelio realizavimas, parametrų vertinimas.
6. Rezultatų palyginimas.
7. Išvadų formulavimas.

Įrankiai

Naudotos programavimo kalbos

Python - beveik visas programinis kodas
Shell - parametrų vertinimo kodas pateikiamas VU HPC

Naudotos bibliotekos

Pandas - duomenų analizei
numpy - modelio operacijoms
tqdm - progreso komandos eilutėje sekimui
librosa - konvertuoti iš laiko eilutės į paveiksliuką
matplotlib - grafikams kurti ir saugoti
sklearn - duomenų paruošimui

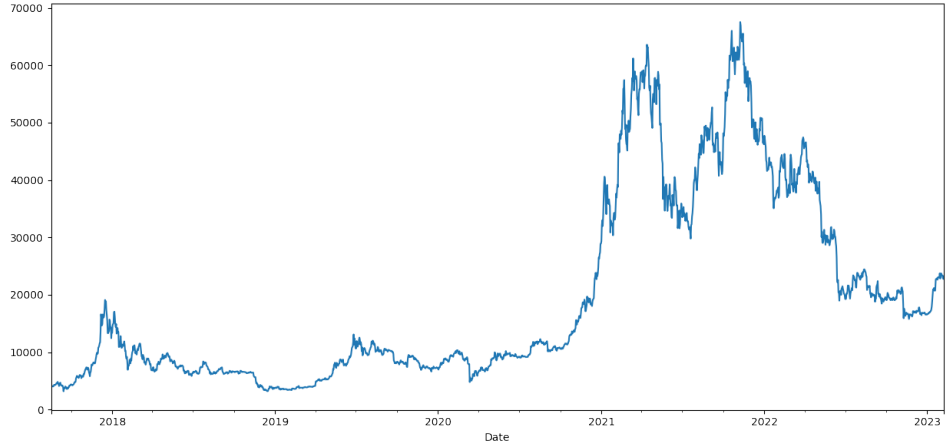
Laiko eilučių duomenys

Kainų duomenys

Kriptovaliutų laiko eilučių duomenys apie Bitcoin, Ether, Ripple kriptovaliutas dieniniais, valandiniais, minutiniais intervalais.

Duomenų kiekis

Dieniniai - 2001 stebėjimai
Valandiniai - 47'894 stebėjimai
Minutiniai - 520'541 stebėjimai


70000
60000
50000
40000
30000
20000
10000
0
2018
2019
2020
2021
2022
2023
Date

Laiko eilučių duomenų šaltinis

CryptoDataDownload

CryptoDataDownload yra duomenų šaltinis, kuris siūlo nemokamus kripto valiutų istorinius duomenis, skirtus moksliniams tyrimams, analizei ir strategijų kūrimui. Ši platforma suteikia prieigą prie didelio kiekio kriptovaliutų prekybos duomenų, apimančių įvairias prekybos platformas, valiutų poras, intervalus ir laikotarpius.

Tai yra svarbus šaltinis tiek pradedantiesiems, tiek pažengusiems kriptovaliutų tyrėjams ir prekiautojams, nes ji leidžia greitai ir lengvai gauti reikalingus duomenis įvairioms analizės užduotims.

Twitter duomenys

Twitter įrašų duomenys

Twitter įrašai, kuriuose buvo paminėti raktažodžiai BTC, ETH, XRP.

Duomenų kiekis

Bitcoin - 1'685'865 stebėjimai
Ether - 787'021 stebėjimai
Ripple - 363'441 stebėjimai

Twitter duomenų šaltinis

Twint

Twint yra atviro kodo Python biblioteka, skirta be jokios autentifikacijos gauti ir analizuoti „Twitter“ duomenis. Ji leidžia naudotojams lengvai rasti ir surinkti informaciją apie „Twitter“ paskyras, pranešimus, raktinius žodžius ir kitus duomenis be jokios autentifikacijos, kas reiškia, kad naudotojams nereikia prisijungti prie „Twitter“ API ir gauti raktą.

Twint privalumai

1) Pranešimų paieška pagal naudotoją, raktinį žodį, datos ribas ir kitus parametrus.
2) „Twitter“ naudotojų paieška pagal raktinius žodžius, vietą, šaltinius ir kitus kriterijus.
3) Gautos informacijos išsaugojimas įvairiais formatais, pvz. CSV, JSON, SQLite.
4) Gautos informacijos analizė ir vizualizavimas.
5) Paralelizavimo galimybės leidžia greitai surinkti didelius duomenų kiekius.

Twitter įrašų duomenys

	created_at	time	username	name	tweet	replies_count	retweets_count	likes_count	cashtags	retweet
0	2021-01-02 01:59:37 EET	01:59:37	jeremysleeks	Jeremy Sleeks	$BTC $ETH $LINK $UNI Last ticket check before...	0	0	0	['btc', 'eth', 'link', 'uni']	False
1	2021-01-02 01:59:28 EET	01:59:28	pablo_algoboss	Don Pablo	Not long now IMO.. big impulse moves coming fo...	1	3	7	['spi', 'btc', 'eth', 'xrp', 'xlm', 'rfi', 'ada']	False
2	2021-01-02 01:59:23 EET	01:59:23	newscryptobot	DoodBot	Jan 1, 2021 23:59:00 UTC The price of $BTC cur...	0	0	0	['btc']	False
3	2021-01-02 01:59:20 EET	01:59:20	realcryptosheet	CRYPTOSHEET	Track your portfolio in Excel instead of an ap...	0	0	0	['btc', 'eth', 'neo', 'tusd', 'ltc', 'dot', 'u...	False
4	2021-01-02 01:58:47 EET	01:58:47	ultrahdr6	UltraHDR6	$BTC #BTC 30K in play! https://t.co/zfE5rbubqO	0	0	1	['btc']	False
...	...	...	...	...	...	...	...	...	...	...
5054	2021-01-01 02:00:03 EET	02:00:03	whaletrades	WhaleTrades🐋	Bitmex: OI: 570,187,093 Funding: 0.01% 24H Vol...	0	0	2	['btc', 'btc']	False
5055	2021-01-01 02:00:02 EET	02:00:02	indacoin	Indacoin	🎄Happy holidays and a happy new year to all ou...	2	6	8	['btc', 'inda']	False
5056	2021-01-01 02:00:02 EET	02:00:02	btchourlyprice	Bitcoin Hourly Price	1 $BTC = $28,949.40 $USD 🟢 +0.28% chg/24h #Bi...	0	0	0	['btc', 'usd']	False
5057	2021-01-01 02:00:00 EET	02:00:00	binancerekts	MyAlgo MONITOR	🔊 ⚠️ $BTC 4 Hours Update ⚠️ 🔊 Price @ $28950 (...	0	0	0	['btc']	False
5058	2021-01-01 02:00:00 EET	02:00:00	binancerekts	MyAlgo MONITOR	🔊 ⚠️ $BTC Daily Update ⚠️ 🔊 Price @ $28950 (Ch...	0	0	0	['btc']	False

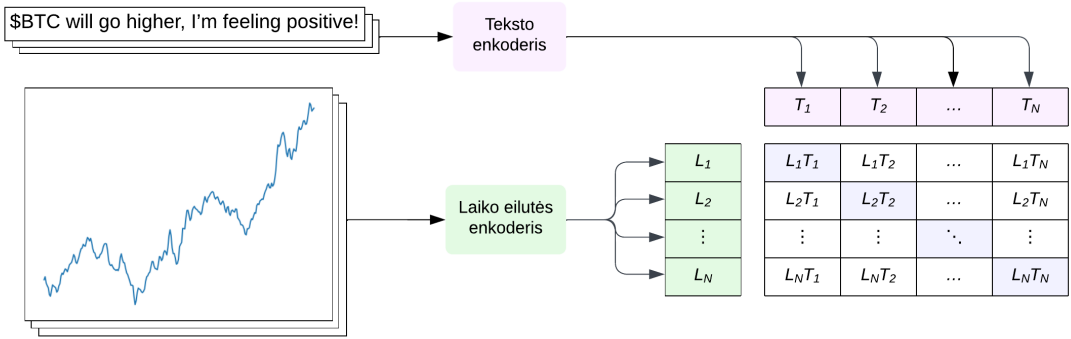
5059 rows × 10 columns

Twitter įrašų duomenys

	tweet	username	timestamp
0	$BTC Price: $47933 $ETH Price: $3577 $LTC Pric...	coincapitan	2021-09-17 02:59:22
1	@mikealfred @philipmak Zero chance $BTC will b...	tallseller	2021-09-17 02:58:54
2	$BTC is going to $80,000 by October 31st and $...	0xgoodies	2021-09-17 02:58:45
3	Premium Membership Update! Made a few updates...	charts_bitcoin	2021-09-17 02:57:41
4	Kind of agree. $TSLA $BTC. That's about it.	umbisam	2021-09-17 02:57:28
...	...	...	...
1685860	$ANY here is the DD link... No excuse on missi...	boondockqueens	2021-07-22 19:28:35
1685861	$clv @clover_finance @Polkadot They want your ...	cspratt15139	2021-07-22 19:28:28
1685862	Jul 22, 2021: The current Mayer Multiple is 0....	tipmayermultple	2021-07-22 19:28:21
1685863	@ARKInvest @CathieDWood Ok, it's now Bitcoin t...	marian55276549	2021-07-22 19:28:08
1685864	$AMB #AMB going to 2X from here from 90 Satosh...	dcryptoservice	2021-07-22 19:28:02

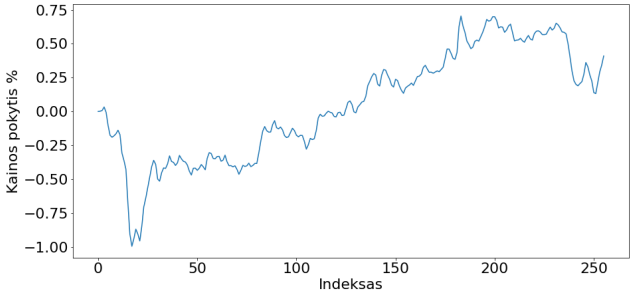
1685865 rows × 3 columns

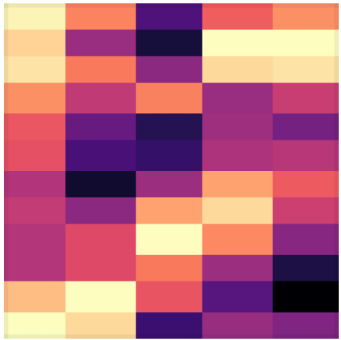
CLIP modelio architektūra



Įvestis originalaus modelio

Modifikuotas modelis naudoja įvestį kairėje, originalus - dešinėje

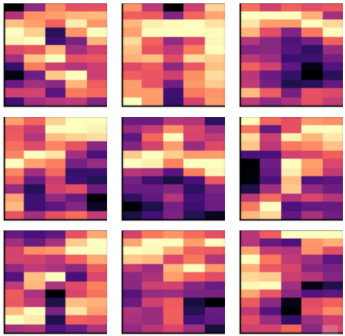




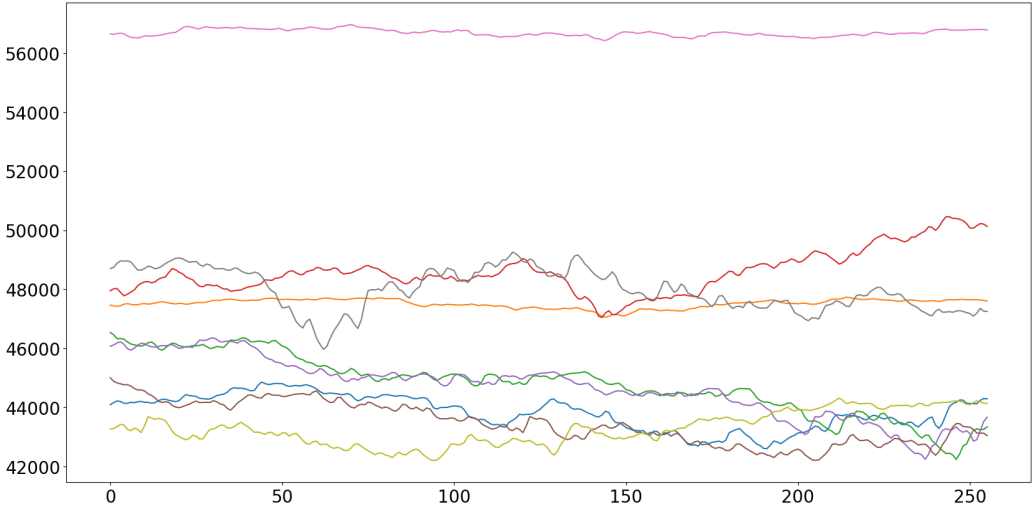
Prognozavimas originalaus modelio

Užklausa

The steady rise in BTC price not only reflects its immense value as a digital asset, but also offers a promising outlook for investors, fueling excitement and confidence in its future trajectory



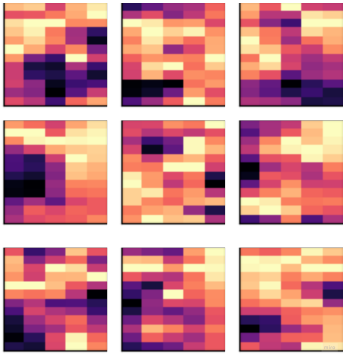
Prognozavimas originalaus modelio



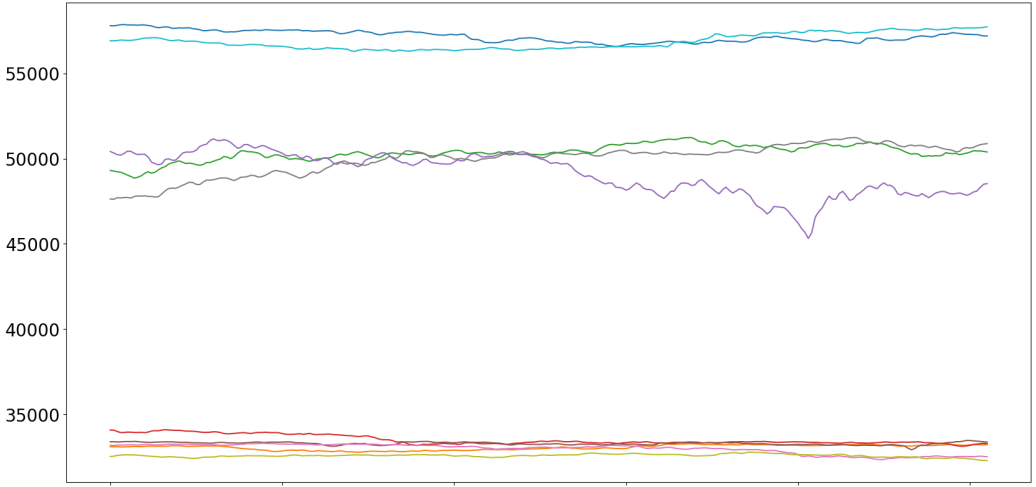
Prognozavimas originalaus modelio

Užklausa

The volatile nature of BTC price can make it challenging for investors to predict and navigate, causing anxiety and uncertainty in the market.



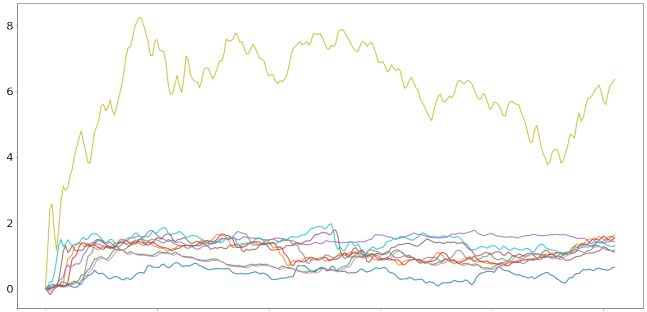
Prognozavimas originalaus modelio



Prognozavimas modifikuoto modelio

Užklausa

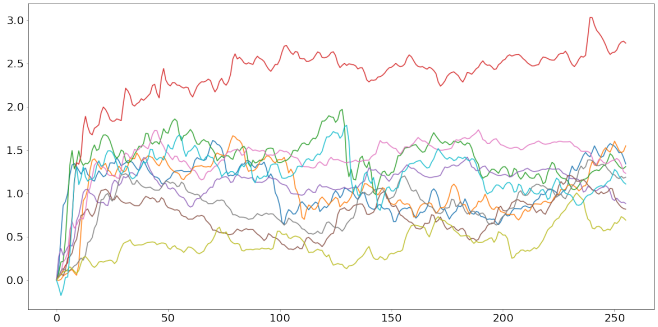
The steady rise in BTC price not only reflects its immense value as a digital asset, but also offers a promising outlook for investors, fueling excitement and confidence in its future trajectory



Prognozavimas modifikuoto modelio

Užklausa

The volatile nature of BTC price can make it challenging for investors to predict and navigate, causing anxiety and uncertainty in the market.



Iššūkiai

1) Ilgai užtruko duomenų gavimas
2) Modifikuoti CLIP modelį laiko eilutėms
3) Neužteko VU HPC duotų valandų
4) Ilgai truko mokymas

Rezultatai

1) Įvertinti parametrai 6 modeliams, nes turėjau 3 kriptovaliutas: BTC, ETH, XRP ir kiekvienam modelio ir kriptovaliutos porai įvertinau parametrus.

2) Modelis daug informacijos iki galo neišmoko

3) Prognozuotų laiko eilučių vidurkiai žymiai skyrėsi

Išvados

1) Buvo atlikti 2 eksperimentai, pirmasis susijęs su originalaus CLIP modelio parametrų vertinimu, kitas su modifikuoto CLIP modelio parametrų vertinimu.

2) Gauti modeliai, jų svoriai, ir duomenys išsaugoti

3) Modifikuotas modelis priima ir spėja laiko eilutes, o ne paveiksliukus kaip originalus CLIP modelis

4) Šis uždavinys buvo ir vis dar yra sunkus šiuolaikiniams modeliams

5) Iš padarytų eksperimentų galima šiek tiek suprasti kad CLIP modelis nėra geras būdas modeliuoti laiko eilutes ir Twitter įrašus apie jas.

Ačiū už dėmesį

Recenzija

Kas yra supervizija?

Supervizija - tiesioginis vertinys iš anglų kalbos, sakinyje norima pasakyti kad natūralios kalbos tekstai gali būti panaudoti kaip nepriklausomas kintamasis modeliuoti vaizdus arba laiko eilutes

Autorius formuluoja įvade „mano tikslas“ bei „bakalauro darbo tikslas“, kurie iš esmės sutampa. Ar autorius turėjo skirtingą tikslą nei buvo bakalauro darbo tikslas?

Norėjau pabrėžti tikslą dar kartą, tikslai pilnai sutampa.

Darbe prie 5 ir 6 paveiksliukų nenurodyti šaltiniai. Ar tai paties autoriaus paveiksliukai? Jei taip, kodėl jie anglų kalba?

Tai yra iš CLIP modelio publikacijos, kuria remiasi visas darbas. Ši publikacija darbe cituota daug kartų.

Recenzija

4.4 skyrelyje minimos teigiamos ir neigiamos poros. Kas tai yra?

Pavyzdys: ten kur nuotraukose yra katė ir tekstas yra apie katę, tai vaizdo kintamieji ir teksto kintamieji vektorinėje erdvėje bus panašus (artimi/teigiami), o paėmus šuns ir katės porą, bus erdvėje nepanašūs (tolima/neigiama).

Paaiškinkite epochos sąvoką

Parametriniame modelyje vykdant parametrų vertinimą jis vykdomas iteraciniais metodais, tada viena iteracija yra vadinama epocha. Per ją modelis pamato visus iteracijos duomenis

Recenzija

23 pav. parodytos neigiamų sentimentų laiko eilutės panašios į atsitiktinė klaidžiojimą. Ar sentimentų laiko eilutėms buvo taikomi laiko eilučių testai, tiriantys jų stacionarumą?

32 psl. pateikiami teigiamų ir neigiamų sentimentų vidurkiai. Ar skirtumas yra statistiškai reikšmingas? Jei tai nebuvo nagrinėta, paaiškinkite kodėl?

Atsakymas

Geros pastabps, prie tokių ir dar gilesnių palyginimų nespėjau prieti, nes labai didelė laiko dalis buvo praleista gaunant duomenis ir vertinant parametrus