پردازش سیگنال ها با استفاده از مدل های زبانی بزرگ

اعضا گروه: اریا زادابراهیمی : aryazady@gmail.com
اسرا کاشانی‌نیا : esra.ka170@gmail.com
محمدرضا عباس نیا : e.abbasniya@gmail.com
امیرتوکلی : Fifa19forfoli@gmail.com
محمدصالح کیانی : kiani.mohammadsaleh@gmail.com
سینا بهراملو : sinabahramloo@gmail.com

کلمات کلیدی: پردازش سیگنال، مدل های زبانی بزرگ

چکیده

با توجه به درآمدزا بودن بازارهای مالی افراد بسیاری در این بازارها سرمایه گذاری می‌کنند. چالش اصلی در این نوع سرمایه گذاری پیش‌بینی بازار است. برای این پیش‌بینی فاکتورهای زیادی دخیل هستند که برخی از آن‌ها شامل اخبار، کارایی سهم، صنعت، اقصاد، تمایلات سرمایه گذاران، اخبار شبکه‌های مجازی و ... می‌باشند. به طور کل برای پیش‌بینی می‌توان از تحلیل‌های بنیادی و یا تحلیل‌های فنی استفاده کرد. هدف این پروژه کار بر روی تحلیل فنی و استفاده بهتر از این نوع تحلیل است. تحلیل فنی خود می‌تواند به روش‌های پیش‌بینی روند کلی و یا چه پیش‌بینی تک کندل بعدی تقسیم شود. بیشتر افراد بصورت دستی به تحلیل این بازارها می‌پردازند. اما بهتر است برای تحلیل داده‌های زیادی همچون بازارهای مالی از روش‌های مبتنی بر یادگیری ماشین بهره گرفت. تا کنون روش‌هایی که وجود داشته مدل‌هایی همچون SVM، CNN و LSTM بوده است. در این پروژه با فرض اینکه سیگنال بازارهای مالی خود یک زیان هستند که با توجه به الگوهای تشکیل شده تغییر می‌کنند، تلاش کرده‌ایم تا با استفاده از مدل‌های زبانی بزرگ، ادامه سیگنال ورودی را پیش‌بینی کنیم. برای این کار از کندل‌ها بعنوان ورودی و از مدل GPT-2 برای پیش‌بینی استفاده شده.

مقدمه

بازارهای مالی محل مناسبی برای سرمایه گذاری است. این بازارها ریسک بالایی دارند ولی سود بسیار زیاد آن‌ها مخاطبان زیادی را به خود جلب کرده. با توجه به تاثیر فاکتورهای بسیار متنوع در تغییر قیمت این بازارها، روش‌های متفاوتی برای تحلیل آن‌ها وجود دارد. یکی از این روش‌ها تحلیل فنی می‌باشد. در تحلیل فنی با استفاده از خود نمودار تلاش می‌شود تا ادامه روند پیش‌بینی را پیش‌بینی کنیم. یعنی با استفاده از داده‌های تشکیل شده تا لحظه فعلی، ادامه روند چه بصورت روند یا بصورت تک کندل حدس بزنیم. برای این کار اکثر افراد بصورت دستی و با نگاه کردن به نمودار و تعیین مشخصه‌هایی سعی در پیش‌بینی بازار دارند. با توجه به اینکه انسان دارای خطا بالا در تحلیل داده‌های زیاد، بهتر است از توان پردازشی بیشتر که دارای خطای کمتری است استفاده کنیم. به همین منظور ربات‌هایی مرسوم برای خرید و فروش وجود دارند که کارشان تحلیل بازار حال با استفاده از الگوریتم‌های ساده و یا هوش مصنوعی است. برای کارایی بهتر بر روی داده‌های موجود، گزینه مناسب تر استفاده از یادگیری ماشین است. در این زمینه

روش‌هایی که تا کنون وجود داشته از SVM، CNN و LSTM استفاده شده. در روش‌های کانولوشنی سعی بر تشخیص وضعیت بازار با استفاده از تصاویر نمودار (کندل‌ها) می‌باشد و در روش‌های مبتنی بر LSTM نمودار را بعنوان یک توالی به مدل داده و انتظار آن است ادامه نمودار را با کمترین خطا تشکیل دهد. در جدول پایین به مقایسه این روش‌ها بصورت خلاصه می‌پردازیم.

منبع	روش	ورودی	خروجی	بررسی روزانه	هم‌بستگی توالی زمانی
۱	تحلیل کندل و مشخصه‌های آن‌ها با تمرکز بر روی بازنمایی‌های تصویری و ابزار تحلیل فنی	قیمت‌های باز، بسته، پایین و بالا	تحلیل تصویری	دارد	ندارد
۲	طراحی یک سیستم خبره برای تفسیر نمودار. این تفسیر با پنج الگو از قبل تعیین شده انجام می‌شود	الگوهای از قبل تعریف شده	پیش‌بینی حرکت قیمت سهم	ندارد	ندارد
۳	با استفاده از ویژگی‌های ورودی یک درخت تصمیم گیری CAD می‌سازد	رنگ، اندازه و موقیت نسبی کندل‌ها	پیش‌بینی انجام عملیات‌های خرید، فروش و نگه‌داشتن	ندارد	دارد
۴	از دو مدل ۱-رگرسیون تعمیم یافته ۲-شبکه کنترل منطق هیبرید استفاده می‌کند	رنگ، اندازه و موقیت نسبی کندل‌ها	پیش‌بینی حرکت قیمت سهم	ندارد	دارد
۵	استخراج ویژگی‌های احساسی شبکه‌های اجتماعی و یکپارچه کردن آن با اطلاعات کندل‌ها و تولید داده چند کاناله. استفاده از مدل کانولوشنی برای قیمت سهم	نمودار کندل	پیش‌بینی حرکت قیمت سهم	ندارد	ندارد
۶	تولید گراف نمودار کندل‌ها. استفاده از مدل Multiple attention GNN و پیش‌بینی نوسانات قیمت	رنگ، سایه بالا، سایه پایین و طول کندل	پیش‌بینی حرکت قیمت سهم	دارد	دارد

اما در این پروژه تلاش ما این بوده تا با فرض زبان بود سیگنال و استفاده از مدل‌های بزرگ زبانی به تحلیل فنی بازارهای مالی بپردازیم. یعنی با فرض اینکه کندل‌های تشکیل شده یک زبان خاص هستند، سعی می‌کنیم تا به درکی از این زبان برسیم و از آن برای پیش‌بینی استفاده کنیم. برای این کار ابتدا لازم است ورودی که کندل‌ها می‌باشند را به ورودی مناسب تبدیل کرده و سپس با استفاده از این مدل‌های زبانی بزرگ الگویی بین ورودی‌ها پیدا کنیم. در نهایت پس از آموزش مدل می‌توانیم با استفاده از روش تولید متن به پیش‌بینی نمودار چه بصورت روند و چه بصورت پیش‌بینی توکن بعدی عمل کنیم. در ادامه به روش‌های استفاده شده در این پروژه می‌پردازیم.

روش‌ها و نتایج

همانطور که دیدیم برای استفاده از داده‌های بازار مالی روش‌های مختلفی وجود دارد. داده‌های مورد استفاده در این پروژه بصورت ۴ مقدار هر کندل (باز، بسته، بالا و پایین) می‌باشد. این داده‌ها از وب سایت Forexsb دانلود شده‌اند. داده‌های مرتبط با نمودار در پنجره‌های زمانی مختلف موجود هستند، یعنی هر کندل می‌توان بیان‌گر تغییرات قیمت در بازه‌های مختلف (روزانه، ساعتی، دقیقه‌ای و ...) باشد. برای بهره‌مندی از این داده‌ها می‌توان به صورت‌های مختلف عمل کرد.

مدل اصلی مورد استفاده برای پیش‌بینی ، مدل GPT-2 می‌باشد. ورودی و خروجی این مدل در حالت استاندارد متن است در صورتی که داده‌های ورودی در مسئله ما عددی است. یعنی یک فاصله بین داده و مدل وجود دارد که برای برطرف کردن این مشکل از رویکردهای متنوعی استفاده کرده‌ایم که در ادامه به توضیح هر یک از روش‌ها می‌پردازیم.

روش اول – پردازش سیگنال به کمک زبان طبیعی

در سال‌های اخیر، با پیشرفت فناوری و رشد روزافزون حجم داده‌ها، استفاده از روش‌های یادگیری ماشین و هوش مصنوعی در پیش‌بینی بازارهای مالی به‌طور چشمگیری افزایش یافته است. ابزارهای یادگیری ماشین، از جمله الگوریتم‌های شبکه‌های عصبی عمیق (Deep Learning) ودر دیتهای Time Series مانند LSTM و GRU، به عنوان ابزارهایی قدرتمند برای مدل‌سازی و پیش‌بینی نوسانات بازارهای مالی شناخته می‌شوند.

مزایای استفاده از ابزارهای یادگیری ماشین مانند قدرت پردازش بالا و قابلیت یادگیری از داده‌های بزرگ، قادر به تجزیه و تحلیل پیچیدگی‌های بازارهای مالی هستند که قبلاً توسط روش‌های سنتی دست‌نیافتنی بود. همچنین پیش بینی ابزارهای یادگیری ماشین معمولاً موثرتر از روش‌های سنتی می‌باشند. این امر باعث می‌شود تصمیم‌گیری‌های مالی بهتری انجام شود. استفاده از داده‌های زمان‌دنباله‌ای: ابزارهای یادگیری ماشین مانند LSTM و GRU قادر به مدل‌سازی دقیق داده‌های زمانی و نوسانات قیمت‌ها در بازارهای مالی می‌باشند که این امر می‌تواند به بهبود پیش‌بینی‌ها کمک کند.

در این تحقیق، از ابزار های مبتنی برPrompt engineeringبرای تولید قیمت های جدید بر اساس مدل ترین شده و فاین توین شده بر روی دیتای تاریخی یک سهام استفاده شد. در مدل اول پرامپت ها بصورت قیمت روزانه مستقیم پایانی تعیین میشود و مدل ما توسط این پرامپت و همچنین فاین توین دیتای تاریخی اقدام به تولید قیمت جدید بصورت auto-regressive میکند. ابزار ارزیابی ماRMSE است. مدل دقیق استفاده شده GPT2 است و نتایج پیش بینی شده دقت 1توسطی دارند. مدل بعدی استفاده شده مبتنی بر LSTM through time است که مطابقا در نمودار دارای دقت بهتری بر روی توزیع داده های ما دارد و همچنین در معیار ما، ارزیابی بهتری میشود و در نهایت از ابزارهای کار با دیتاهای سری زمانی مانند Prophet استفاده شد که علی رغم وجود دقت مناسب بشدت ساده انگارانه به مساله نگاه میکنند و در عمل غیر قابل استفاده هستند.

دیتا و تقسیم

دیتای استفاده شده از پکیج yfinance و قیمت سهام بصورت روزانه است. این دیتا بصورت یک دیتافریم Pandas است و در ابزارKeras استفاده میشود. ستون های اطلاعاتی ما شامل close open high low volume است که ما اکثرا از قیمت close

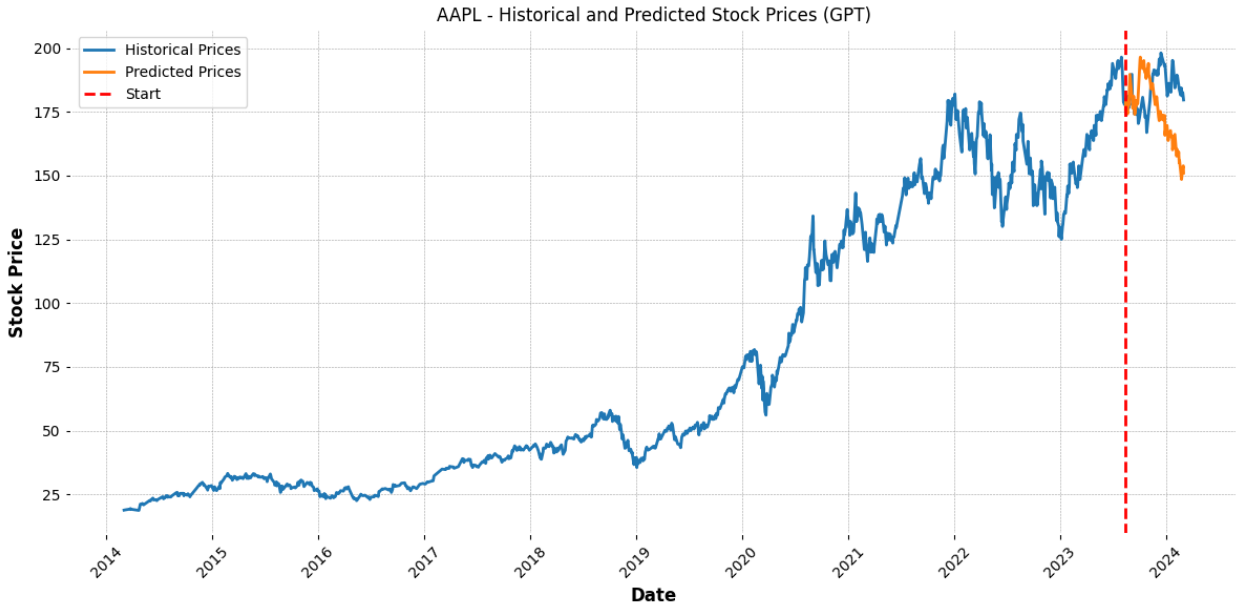
استفاده میکنیم. این دیتا چون بصورت زمانی است تقسیم آن به دیتای تمرینی و تست بصورت قراردادن نقطه ای در میان زمان است. تسک ما بصورت Self-supervised می باشد.

معیار ارزیابی

معیار ما استفاده از RMSE برای مقایسه نقطه به نقطه قیمت های اصلی تست و پیش بینی شده است.

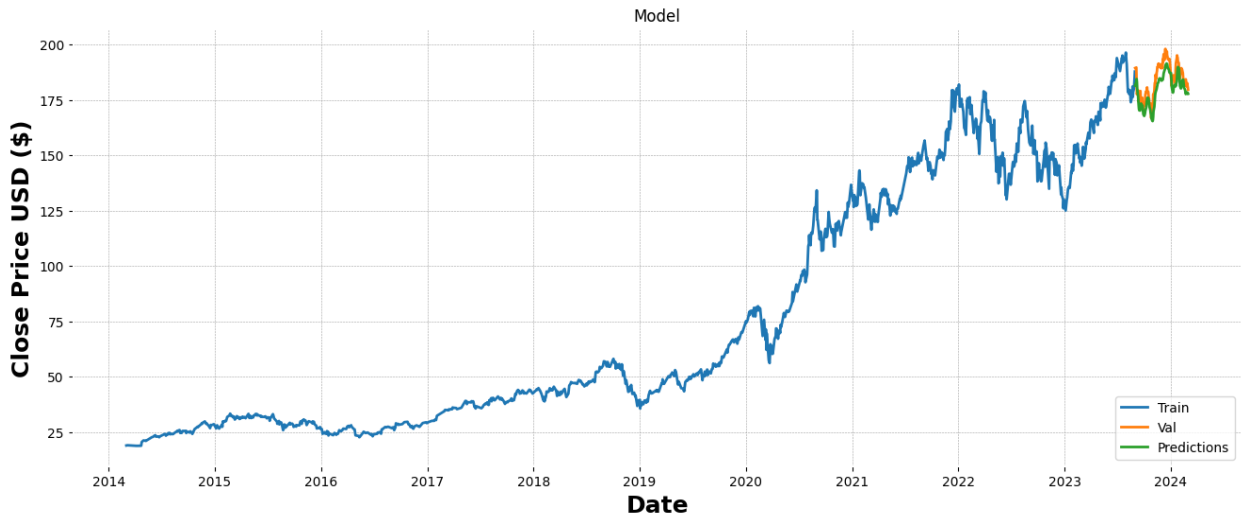
روش های پیاده سازی

در روش اول ما مدل GPT2 را بر روی دیتا ها عددی تمرین بصورت کلی و جلورونده فاین توین میکنیم. این عمل بصورت auto-regressive انجام میشود. نتایج حاصله بصورت متتعیر در دسترس است و میتوانیم ان را برروی نمودار ببینیم:



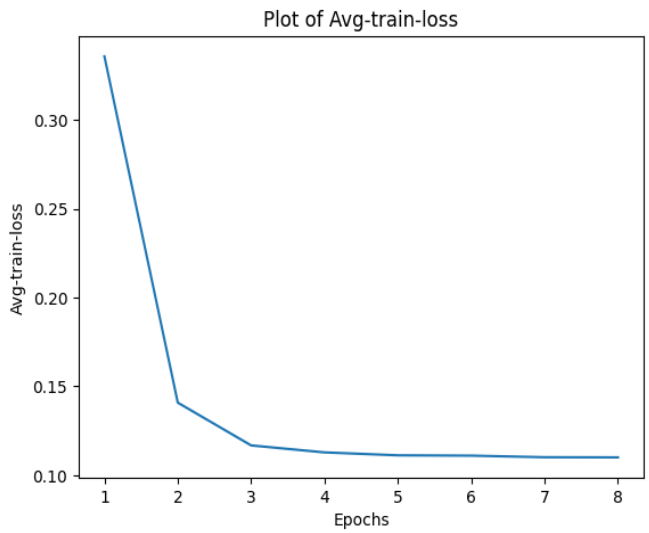
همین طور که مشخص است. ترند های مارکت در ابتدا به درستی تشخیص داده شده اند اما در ادامه مسیر کلی تغییر کرده است. از نظر معیار ارزیابی مان دقت کمی پایین است و خوب با توجه مدل ما که کمی ضعیف است و دیتا های ما که جنبه تقریبا رندومی در خود دارند به نسبت خوب عمل شده است.

روش دوم استفاده از مدل های LSTM through time برای تولید قیمت ها در طی زمان بر روی دتیای ترین و مقایسه ان با قیمت آینده است. معماری ساخته شده واسطه استفاده از لایه های LSTM بر روی هم بصورت استک شده است.



قیمت پیش بینی شده نسبتا دقت خوبی دارد و به قیمت اصلی شباهت زیادی دارد. از نظر معیار ارزیابی هم وضعیت بهتر است که نشان دهنده معماری خوب مدل است.

روش سوم استفاده از ابزار های پیش بینی داده های زمانی مانند است که بسیار ساده انگارانه به قضیه نگاه کردند و علی رقم وجود دقت خوب پیش بینی ان ها بصورت چند جمله ای با یک نویز گوسین است که مناسب این تسک نیستند.



نکته: فاین توین کردن اصلی مدل بر روی دیتای ۱۰ ساله استاک مارکت شرکت Apple و بر روی پلتفرم های مختلف مانند Kaggle انجام شده است. نمودار بدست امده Loss train را مشاهده میکنیم:

روش دوم – پردازش سیگنال با فرض داشتن زبان مخصوص به خود

رویکرد ۱

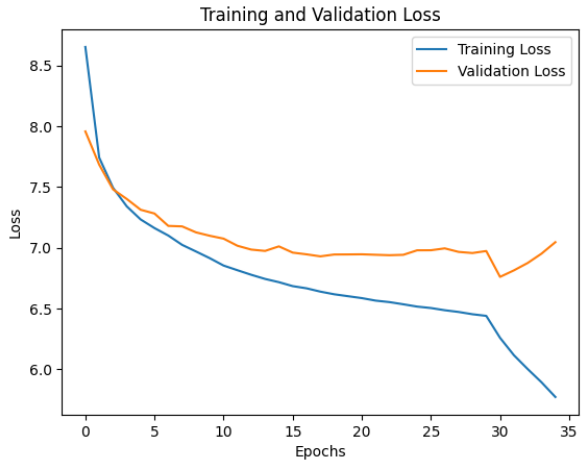
در این روش هر یک از کندل‌ها را توکن زبان فرض کرده‌ایم و هدف پیش‌بینی توکن بعدی است. در این رویکرد تبدیل کندل‌ها به داده‌های گسسته لغات ما را می‌سازند.

در این رویکرد ابتدا داده‌ها را نرمال سازی کرده. سپس آن‌ها را دهک بندی کردیم. در هر دهک، مقدار وسط بعنوان نماینده دهک انتخاب شده و از آن استفاده می‌کنیم. با توجه به اینکه بخشی از اطلاعات دور ریخته می‌شود ولی پس از بازسازی کندل‌ها اطلاعات زیادی از دست نمی‌رود. حال که ورودی بصورت گسسته می‌باشد می‌توان بر اساس قیمت‌های کندل آن‌ها را گروه بندی کردی که در نهایت هر گروه تبدیل به یک کلمه می‌شود. برای ساخت جمله از پنجره با طول ۱۰ استفاده شده. یعنی هر ۱۰ کندل را یک جمله در نظر می‌گیریم. در نهایت ورودی ما شامل یک لیست با ۱۰ آیدی می‌باشد که نمایندگی کندل‌های آن پنجره را به عهده دارند. در این مرحله، از مدل تعلیم دیده شده GPT-2 استفاده می‌کنیم. تنها لایه بازنمایی باید تغییر کند و متناسب با تعداد لغات (گروه‌ها) شود. پس از تغییر لایه بازنمایی، ابتدا تمامی وزن‌های مدل را ثابت می‌کنیم و فقط لایه بازنمایی را تعلیم می‌دهیم. هدف این است که بازنمایی مناسبی برای ورودی (گروه‌ها) پیدا شود. این مرحله در ۲۰ ایپاک انجام می‌شود. سپس تمامی وزن‌های مدل را فعال می‌کنیم و ۱۰ ایپاک دیگر مدل را تعلیم می‌دهیم که تنظیم ریز شود. حال که مدل تعلیم دیده فقط کافیست پنجره به طول ۱۰ از کندل‌ها را به مدل بدهیم تا کندل ۱۱ام را پیش‌بینی کند و آیدی مربوط به آن را برگرداند. از این آیدی می‌تواند برای پیدا کردن کندل مورد نظر استفاده و آن را باز تولید کرد.

رویکرد ۲

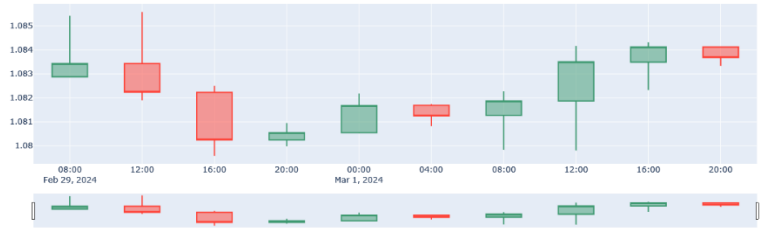
در رویکرد دوم ورودی را به همان شکل پیوسته در نظر می‌گیریم تا اطلاعات کامل آن را داشته باشیم. همانند مدل قبل عمل می‌کنیم با این تفاوت که قبل و بعد از مدل GPT-2 دو لایه قرار می‌دهیم تا ورودی ۴تایی ما را به ۷۶۸ بعد ببرد. از بردار جدید بعنوان ورودی برای GPT استفاده می‌کنیم. تابع خطا در نظر گرفته شده در این قسمت MSE است. مدل ابتدا فقط با وزن‌های دو لایه اضافه شده تعلیم می‌دهیم و در نهایت کل مدل را. برای تولید کندل جدید فقط لازم است یه پنجره از کندل‌ها (۴ تایی) را به مدل بدهیم تا کندل ۱۱ام را پیش‌بینی کند.

در رویکرد اول برخلاف اینکه تابع خطا به میزان خوبی کاهش پیدا کرده و تعمیم پذیری بالایی نیز دارد ولی مدل نتوانسته در پیش‌بینی کندل بعدی خوب عمل کند. با توجه به تابع خطا استفاده شده، مدل طی ۳۰ ایپاک تعلیم می‌بیند و بعد از آن overfit می‌شود.

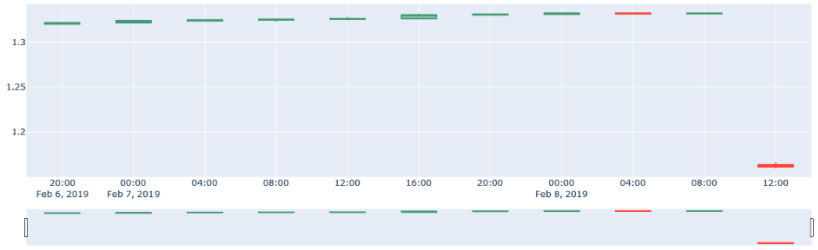


نتایج بدست آمده در دو رویکرد به شکل زیر است.

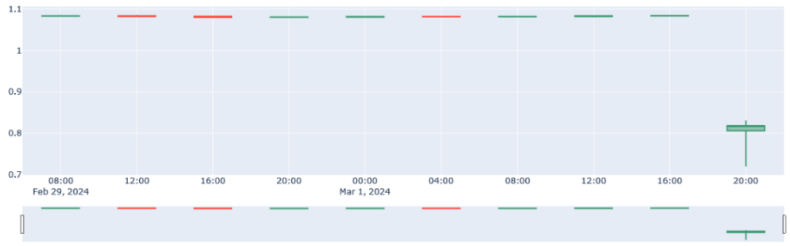
نمودار اصلی



پیش‌بینی با رویکرد اول



پیش‌بینی با رویکرد دوم



رویکرد ۳

در این رویکرد ما فرض میکنیم که سیگنال اولیه یک سیگنال متشکل از کنار هم قرار گرفتن تعدادی سیگنال محدود بوده است و بر همین اساس زبانی برای سیگنال اولیه متصور میشویم که توکن های این زبان همان سیگنال های محدودی است که از کنار هم قرار گرفتنشان سیگنال اولیه ساخته میشود.

برای پیاده سازی این روش ما فقط بر روی داده ستون close تمرکز میکنیم چرا که اولا ستون open از روی آن (با از دست دادن اولین کندل) قابل محاسبه است و همچنین داده های high – low را برای ساده سازی استفاده نکردیم و به این ترتیب با یک سیگنال یک بعدی گسسته در طول زمان سر و کار داریم.

مراحل کار در این رویکرد به این ترتیب است:

۱- داده ها جوری تغییر دادیم که هر داده فقط به قبلی خود وابسته باشد و به جای یک لیست از اعداد حقیقی که هر کدام مقدار close در یک زمان مشخص را دارند یک لیست از اعداد حقیقی که هر کدام میزان تغییر نسبت به کندل قبلی را نگهداری میکنند.

```
[1.3502 ,
 1.35002,
 1.35013,
 ...,
 0.9995 ,
 1.00015,
 0.9998 ]
```

```
[ 1.3502e+00,
 -1.1000e-04,
  2.2200e-03,
  ...,
  -6.5000e-04,
  3.5000e-04,
  -3.5040e-01]
```

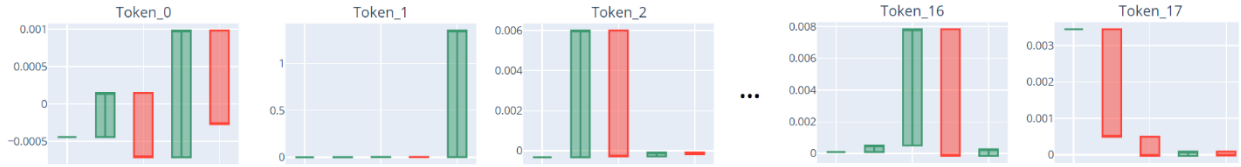
۲- به ازای یک متغیر q و یک متغیر sq داده های سیگنال را quantize کردیم و سیگنال که یک لیست از اعداد بود را به لیستی از وکتور های به طول q+2sq تبدیل کردیم.

```
                    sq                    q                    sq
... 1.3502e+00, -1.1000e-04,  2.2200e-03,  2.6000e-04, -5.5000e-04, 2.6000e-04, -5.5000e-04, -5.0000e-05, -5.8000e-04, -1.1200e-03 ...
                                                                  sq                    q                    sq
... 1.3502e+00, -1.1000e-04,  2.2200e-03,  2.6000e-04, -5.5000e-04, 2.6000e-04, -5.5000e-04, -5.0000e-05, -5.8000e-04, -1.1200e-03 ...
```

```
[ 1.3502e+00,
 -1.1000e-04,
  2.2200e-03,
  ...,
  -6.5000e-04,
  3.5000e-04,
  -3.5040e-01]
```

```
[[ 1.3502e+00, -1.1000e-04,  2.2200e-03,  2.6000e-04, -5.5000e-04],
 [ 2.6000e-04, -5.5000e-04, -5.0000e-05, -5.8000e-04, -1.1200e-03],
 [-5.8000e-04, -1.1200e-03,  4.6700e-03, -1.0000e-04, -3.2700e-03],
 ...,
 [-4.0000e-04, -6.0000e-04,  2.5000e-03, -1.9500e-03, -9.0000e-04],
 [-1.9500e-03, -9.0000e-04,  2.4100e-03, -3.1600e-03,  6.8000e-03],
 [-3.1600e-03,  6.8000e-03,  6.4500e-03, -1.3500e-03, -2.3000e-03]]
```

۳- در ادامه این وکتور ها را به کمک kmean-clustring به ۱۸ خوشه، خوشه بندی کردیم و این یعنی فرض کردیم کل زبان سیگنال دارای ۱۸ توکن است که برخی از آنها را در زیر مشاهده میکنید. دقت شود این توکن ها مستقل از مقدار های اولیه هستند و فقط میزان تغییرات هر کندل نسبت به قبلی را دارند.



۴- حال بر حساب نحوه خوشه بندی به هر وکتور یک لیبل اختصاص میدهیم که متناظر با توکن آن در نظر گرفته میشود.

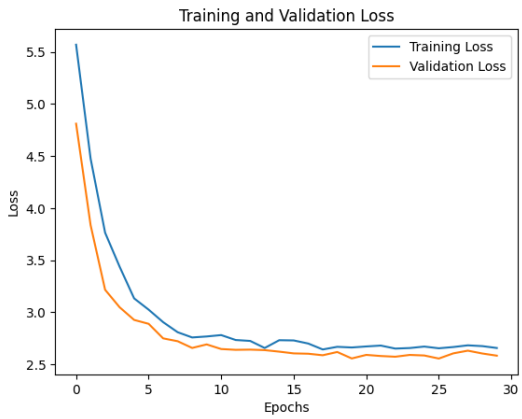
```
[[ 1.3502e+00, -1.1000e-04,  2.2200e-03,  2.6000e-04, -5.5000e-04],
 [ 2.6000e-04, -5.5000e-04, -5.0000e-05, -5.8000e-04, -1.1200e-03],
 [-5.8000e-04, -1.1200e-03,  4.6700e-03, -1.0000e-04, -3.2700e-03],
 ...,
 [-4.0000e-04, -6.0000e-04,  2.5000e-03, -1.9500e-03, -9.0000e-04],
 [-1.9500e-03, -9.0000e-04,  2.4100e-03, -3.1600e-03,  6.8000e-03],
 [-3.1600e-03,  6.8000e-03,  6.4500e-03, -1.3500e-03, -2.3000e-03]]
```

```
['USDCAD_1',
 'USDCAD_0',
 'USDCAD_7',
 ...,
 'USDCAD_7',
 'USDCAD_4',
 'USDCAD_13']
```

۵- در مرحله بعدی فرض میکنیم این لیست توکن های یک زبان هستند و آنها را با window گرفتن تبدیل به یک سری sample میکنیم که مدل قرار است ترتیب آنها را یادبگیرد. نهایتا داده ما آماده برای یادگیری توسط یک مدل زبانی است.

	x	y
0	[USDCAD_0, USDCAD_12, USDCAD_12, USDCAD_7, USD...	USDCAD_1
1	[USDCAD_0, USDCAD_7, USDCAD_7, USDCAD_7, USDCA...	USDCAD_3
2	[USDCAD_14, USDCAD_2, USDCAD_12, USDCAD_12, US...	USDCAD_3
3	[USDCAD_7, USDCAD_0, USDCAD_0, USDCAD_0, USDCA...	USDCAD_3
4	[USDCAD_13, USDCAD_17, USDCAD_5, USDCAD_3, USD...	USDCAD_12
...	...	...
1430	[USDCAD_11, USDCAD_0, USDCAD_5, USDCAD_16, USD...	USDCAD_13
1431	[USDCAD_17, USDCAD_16, USDCAD_5, USDCAD_16, US...	USDCAD_12
1432	[USDCAD_0, USDCAD_17, USDCAD_5, USDCAD_3, USDC...	USDCAD_17
1433	[USDCAD_6, USDCAD_17, USDCAD_5, USDCAD_15, USD...	USDCAD_8
1434	[USDCAD_13, USDCAD_17, USDCAD_5, USDCAD_3, USD...	USDCAD_0

توکن را به یک سری integer id تبدیل کردیم و مدل زبانی GPT2 را برای train کردم استفاده کردیم که نمودار loss آن برای ۳۰ ایپاک به شکل زیر است.



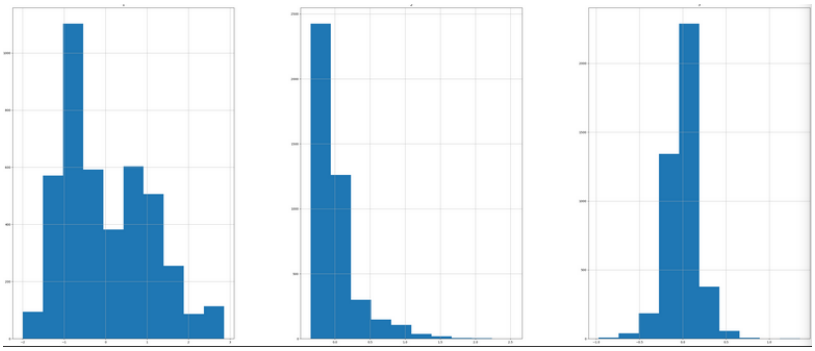
رویکرد ۴

در این روش ابتدا دسته‌های ۸تایی از کندل‌ها را کنار هم می‌گذاریم تا توکن‌های ورودی GPT-۲ را تشکیل بدهیم. هر توکن ۸کندلی با توکنی که از ۸ کندل بعد آن تشکیل شده در حد ۲ کندل اورلپ دارد.

سپس در هر توکن، بردار هر کندل را از بردار کندل بعدی‌اش کم می‌کنیم تا به جز کندل اول، برای بقیه‌ی کندل‌ها، کندل iام دارای مقدار اختلاف کندل اصلی iام با کندل i-۱ام باشد. همینطور میانگین و واریانس مقادیر ۸ کندل (شامل اعداد open, close, high, low, volume) را روی سطر (یعنی این پنج عدد در یک بازه ۴ ساعته) و روی ستون (یعنی هر کدام از این مقادیر به تنهایی در یک

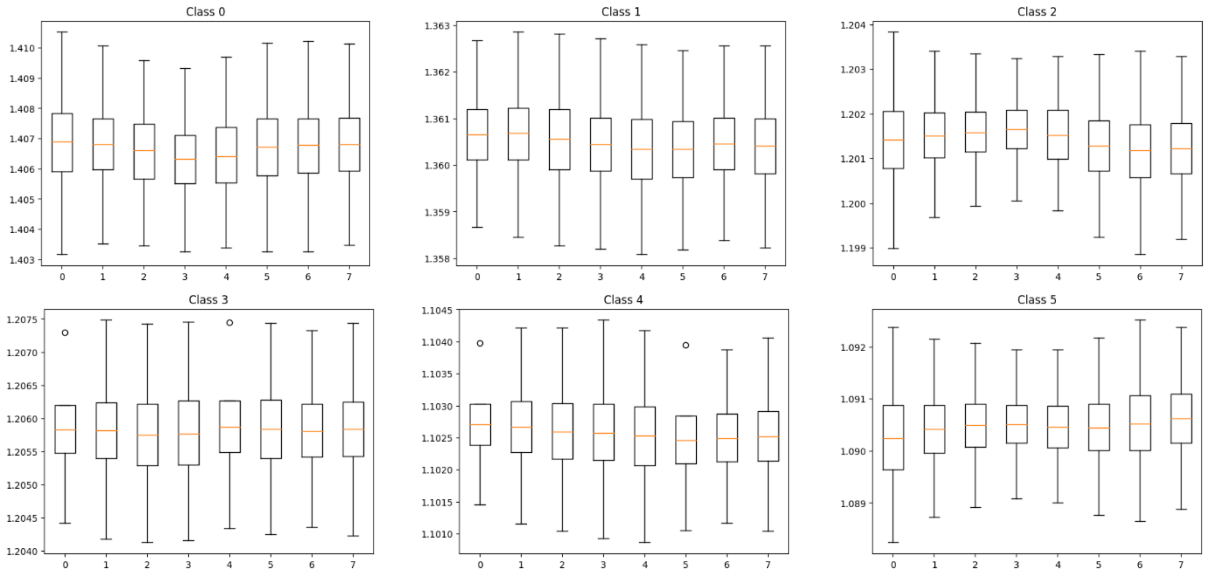
بازه ۸*۴ ساعته) را حساب می‌کنیم و به بردار flattenشده‌ی توکن ۸کندلی، concat می‌کنیم. سپس با PCA بردارهای حاصل برای توکن‌ها را به ۳ بعد کاهش می‌دهیم.

توزیع مقادیر سه بعد به شکل زیر است:

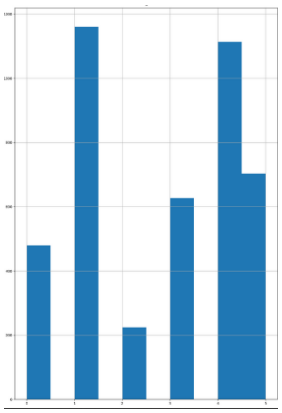


بعد مقادیر هر کدام از سه بعد را بازه‌بندی می‌کنیم و به بعد اول سه بازه و بعد دوم دو بازه اختصاص می‌دهیم و بنابراین توکن‌ها در ۳*۲=۶ کلاس طبقه‌بندی می شوند (تعداد ابعاد حاصل از PCA و بازه‌های ابعاد هایپرپارامتر است و با آزمون و خطا این مقادیر بدست آمد).

سپس به هر توکن یک عدد از ۰ تا ۵ اختصاص می‌دهیم که همان شماره کلاس است که بر اساس اینکه مقادیر ابعاد بردارهای آنها در کدام بازه‌ها قرار می‌گیرد تعیین می‌شود. در زیر میانگین توکن‌های ۸ کندلی بر اساس کندل‌های داده خام برای هر کلاس نشان داده شده است.



توزیع فراوانی توکن‌ها در هر کلاس بصورت زیر است:

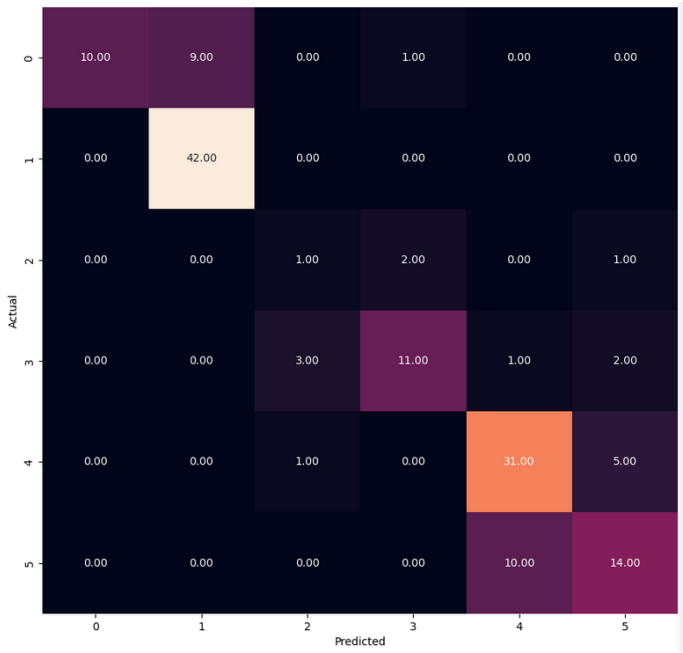


ادامه ی کار مانند قبل است. از توکن‌های ۸کندلی، دنباله‌های ۶تایی درست می‌کنیم که این ۶ توکن از نظر زمانی پشت سر هم اتفاق افتاده‌اند. قرار است مدل GPT به صورتی ترین شود که وقتی کلاس ۵ توکن را به آن می دهیم، کلاس توکن بعدی را پیش‌بینی کند.

داده را به سه قسمت ترین و ولیدیشن و داده تست تقسیم می‌کنیم و مدل GPT-۲ را ترین می‌کنیم. Loss برای داده‌های ترین و ولیدیشن در طول ۳۰ ایپاک در هنگام train به صورت زیر است:

البته چون داده‌های batchها ثابت نیست Loss برای ولیدیشن کمی تغییر می‌کند اما معمولا بعد از ۳۰ ایپاک بین ۷۵/۰ و ۲/۱ است.

در نهایت هم با داده‌ی تست پیش‌بینی را انجام می‌دهیم. Confusion matrix برای داده‌ی تست به صورت زیر است.



مشاهده می‌کنیم که تعداد داده در دقت مدل بسیار تعیین‌کننده است. به صورتی که کلاس ۲ که تعداد داده‌هایش بطور قابل‌توجهی کمتر از بقیه بود بیشتر با کلاس ۳ اشتباه گرفته می‌شود.

هم‌چنین مدل وقتی بین کلاس ۰ و ۱ مردد است برچسب ۱ را ترجیح می‌دهد چون فراوانی کلاس ۱ بیشتر است. اگر داده‌ی بیشتری داشته باشیم balanced کردن داده‌ها برای ادامه کار پیشنهاد می‌شود.

نتیجه گیری

نتیجه گیری روش اول

بصورت کلی جنس دیتای سری زمانی و پیش بینی آن، تسک سختی برای مدل های زبانی است و نیاز است که از مدل های پیچیده با ساختار ها و معماری های مختلف استفاده شود. همچنین پرامپت دهی در این روش برای مدل های زبانی بزرگ و دریافت اطلاعات مناسب از انها کمی مشکل است و نیازمند بررسی بیشتر می باشد. با توجه به نکات گفته شده این روش ها دقت متوسطی در ارایه روند کلی بازار های مالی دارند و دیتاهای بسیار بیشتر به همراه قدرت پردازشی بیشتر نیاز است.

نتیجه گیری روش دوم

رویکرد ۱و ۲

در این پروژه فرض ابتدایی این بوده که سیگنال‌های بازار مالی خود یه زبان هستند و می‌توان با مدل‌های زبان بزرگ الگوهای موجود را پیدا کرد و با توجه به الگوهای بدست آمده به پیش‌بینی بازار دست پیدا کرد. در روش مدل GPT-2 نتوانست الگو خاصی تشخیص دهد و پیش‌بینی به پیش‌بینی دقیقی برسد. اما همچنان می‌تواند با تغییر معماری مدل استفاده شده انتظار داشت مدلی تعلیم داد تا بتواند سیگنال‌های بازار مالی را درک کند و از آن در پیش‌بینی استفاده کرد.

رویکرد ۳

با فرض هایی که در این روش کردیم متوجه شدیم که داد های سیگنالی میتوانند زبانی قابل ارزیابی و یادگیری داشته باشند. تاثیر نوع خوشه بندی و دقت آن در نتیجه نهایی بخاطر ماهیت مسئله (یادگیری زبان) تا حدی کاهش یافته بود ولی قطعا با بررسی بیشتر میتوان کار های بهتری انجام داد. ما معتقدیم در این ایده میتوان با بررسی مدل های بیشتر دقت های بهتری را به دست اورد.

منابع

۱. M. F. Dicle, “Candle charts for financial technical analysis”, 2019

۲. K. H. Lee and G. S. Jo, “Expert system for predicting stock market timing using a candlestick chart”, 1999.

۳. S. ammakesorn and O. Sornil, “Generating trading strategies based on candlestick chart pattern characteristics”, 2019.

۴. T. Kamo and C. Dagli, "Hybrid approach to the Japanese candlestick method for financial forecasting with Applications”,2009. Expert Systems

۵. T. T. Ho and Y. Huang, “Stock price movement prediction using sentiment analysis and CandleStick chart representation” 2021.

۶. Jun Wang, “Predicting Stock Market Volatility from Candlestick Charts: A Multiple Attention Mechanism Graph Neural Network Approach” 2022

۷. pecial reference for the helpful information and plots

۸. https://www.kaggle.com/iamleonie/intro-to-time-series-forecasting

۹. ARIMA Docs

۱۰. T. T. Ho and Y. Huang, “Stock price movement prediction using sentiment analysis and CandleStick chart representation” 2021.

۱۱. Jun Wang, “Predicting Stock Market Volatility from Candlestick Charts: A Multiple Attention Mechanism Graph Neural Network Approach” 2022

۱۲. https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06