دانشگاه تهران

دانشکده برق و کامپیوتر

گزارش پروژه درس شبکه اجتماعی

نگارش:

سپهر کریمی آرپناهی ۸۱۰۱۰۰۴۴۷

سید عرفان باقری دمنه ۸۱۰۱۰۰۲۹۸

پویا جمشیدی ۸۱۰۱۰۰۳۱۴

استاد درس:

دکتر اسدپور

زمستان ۱۴۰۰

۱. جمع آوری داده‌های اجتماعی

در مرحله اول برای استخراج داده‌های مربوط به هر نماد از سامانه هشتگ کمک گرفتیم.

نماد های گروه به شرح زیر می‌باشند:

- ابهام‌دار:
  - شیراز
  - بوعلی
  - پارس
- بدون ابهام:
  - حپترو
  - شپترو
  - وپترو
  - پترول
  - پارسان
  - شاروم
  - شخارک
  - شفن
  - شپدیس
  - شغدیر
  - تاپیکو
  - شاراک

- **استخراج داده های نماد های بدون ابهام:**

برای استخراج داده‌های نمادهای بدون ابهام، از آنجایی که این نمادها دارای ابهام خاصی نیستند و در زبان فارسی بار معنایی خاصی ندارند، به این صورت از سامانه گرفته شده است که اگر سامانه داده‌ای دارد که کلید واژه "اسم نماد" در آن وجود دارد آن داده در برگه مورد نظر آورده شود. برای مثال برگه نماد "حپترو" به این صورت است که تمام داده‌هایی که شامل کلیدواژه‌ی حپترو هستند در برگه مورد نظر آورده شوند، برای اکثر نمادهای بدون ابهام به این صورت عمل شد. اما در برخی نمادهای بدون ابهام نیز شاهد به وجود آمدن داده های اضافی بودیم. مثلا نماد "شاروم" با خواننده‌ای که اسم کوچک او شاروم بوده است مشترک بود. پس ما مجبور شدیم کلید واژه هایی راOR و کلید واژه‌هایی را Not کنیم تا بتوانیم به پست‌های بورسی مورد نظرمان در نماد شاروم برسیم.

- **استخراج داده های نماد های ابهام‌دار:**

برای نماد مبهم ***"شیراز"*** به این صورت عمل شده که از آنجایی که ما به‌دنبال داده‌هایی از جنس مطالب بورسی هستیم می‌بایست کلیدواژه‌ی ***"شیراز"*** را همراه با برخی کلمات مرتبط بیاوریم این کلمات در برخی موارد باعث ورود برخی داده‌ها می‌شوند که با هدف مد نظر ما سنخیتی ندارند بنابراین باید سعی کنیم که داده‌هایی که شامل برخی کلمات خاص هستند را حذف کنیم برای مثال یکی از کلماتی که ما برای استخراج داده با همراه کلیدواژه ***"شیراز"*** در نظر می‌گیریم کلمه‌ی ***"نماد"*** است.

بنابراین داده‌هایی که شامل کلمات کلیدی ذکر شده باشند باید در برگه مورد نظر قرار بگیرند ولی این کار باعث می‌شود بعضی داده‌ها ناخواسته وارد برگه شوند برای مثال شهر شیراز نماد شعر یا عشاق یا هر داده‌ای که به صورت ذکر شده باشد ولی در عمل برای ما مفید نباشد میتواند وارد داده‌ها شود در تحلیل ما تاثیرگذار باشد پس باید به‌نحوی تا حد امکان از ورود چنین داده‌هایی جلوگیری کنیم،

داده‌ها به صورت زیر جمع‌آوری شده‌اند،

داده‌هایی که شامل کلیدواژه‌ی ***"شیراز"*** باشند، همچنین یکی از کلمات زیر نیز همراه آن باشد:

نماد|فرابورس|بازارسرمایه|بازار سرمایه|سهم|سهام|سهمها|سیگنال خرید|سیگنال فروش

همچنین داده نباید دارای کلمات زیر باشد:

غزل|خواجه|حافظ|سعدی|نماینده|بازررسی|استانداری|اعتماد|ارسال|حضوری|سراسر|کشور|پوشاک|دیدنی|اردو|کنفدارسیون|نمایندگان|انتصاب|فامیلی|فوتبالی|شاهین|برق|آب|فساد|پهلوی|کرونا|قرمز|مسجد|نصیرالملک|تخت جمشید|نظامپزشکی|اتاق|زمین|متراژ|رهن|اجاره|سوئیت|استخدام|مارکت|بیمه|پزشک|بدمینتون|بین المللی|پرداخت|الکترونیک|دوزخ|جان|عزرائیل|کاشان|بورس کالا|مجاهدین|سرکرده|انتظامی|تامین اجتماعی|بازماندگان|مستمری|بازنشستگان|اشعار|شعر|آدرس|بازی|شهید|جانباز|مالکیت|مازاد|آرگون|حلال|برزن|کوی|بیابان|خیابان|کامیون|سروناز|سبز|تصمیم|عاشق|عشق|شراب|گلاب|خستگی|قمصر|دچار|برندینگ|جوجه|گنجشک|آشنا|صادق هدایت|قطار|مترو|اتوبان|تبریز|نماینده|نوازنده|نوآوری|دانش آموز|شرکت|خط|خطوط|ریلی|بوشهر|

راه آهن|مدینه|منتظر|دانشکده|دانشگاه شیراز|موکب|اسوه|ظلم|عادل آباد|عسلویه|باغوحش|شیربان|آرامش|آسمان|پنیرک|کرمان|دانشجو|کوه|خزانه|تور|گردشگری|آموزش|ادبیات|گل نرگس|شهلا|باغ|بوستان|میوه|آسیب|اسکان|اضطراری|سیلزده|سیلزده|هموطن|اندوه|باران|چتر|کمک|مردم|سیل زدگان|مدیریت|بحران|انسانیت|خیابون|سیلاب|طومار|خاطره|تعطیلات|ارگ|بم|

کریم خان|زند|نادرشاه|افشار|سفر|شهرداری|ژیان|نوستالژیک|دروازه|قرآن|گردشگر

بعد از ساخت برگه‌ها در سامانه برای هر نماد نوبت دریافت دیتا از سامانه می‌باشد البته در یک موضوع باید دقت شود که برای هر نماد دقت و recall بررسی شود برای نمادهای غیر مبهم که این موضوع به صورت پیشفرض برقرار است و فقط باید برای نماد مبهم بررسی شود برای اینکه بتوان بالا بودن دقت و recall را اثبات نمود دو راه وجود دارد، اول اینکه که همه‌ی داده‌های شامل کلمه کلیدی را بدست آورد سپس بررسی کرد که داده‌های حاصل در برگه با دقت بالا و همچنین recall بالا هستند راه دوم که راهی ساده‌تر است به‌صورت تجربی است که می‌توان تعدادی از داده‌ها را به‌صورت تصادفی بررسی کرد که شامل داده‌های خارج از موضوع بورس نباشند که در بررسی به عمل آمده این موضوع تا حد بسیار خوبی قابل بیان بود همچنین این را نیز باید در نظر گرفت که وقتی داریم تعداد زیادی کلمه را از جواب‌ها حذف می‌کنیم ممکن است داده‌های مرتبط نیز حذف شوند ولی کلماتی که در بخش حذف قرار داده شده‌اند تا حد امکان کمترین اشتراک را با داده‌های بورسی داشته‌اند و به دقت کافی انتخاب شده‌اند. بنابراین پس از اینکه متوجه داده‌های ما از دقت و RECALL مطلوبی برخوردار هستند وقت آن رسیده است که داده‌ها را خروجی گرفته و به سراغ تحلیل اجتماعی برویم.

۲. تحلیل اجتماعی داده ها

در روزهای اولیه‌ای که خروجی داده‌ها را دریافت کرده بودیم با توجه به زمان اولیه تحویل پروژه و در دسترس نبودن گراف نمادها باید برای این قسمت چاره‌ای اندیشیده می‌شد، از قضا راه‌حل اولیه ارائه شده برای این قسمت بعد از بررسی گراف‌ها نیز راه‌حل خوبی به نظر می‌رسد.

پس در ابتدا راه‌حل خودمان را که براساس فایل اکسل هر نماد بود توضیح می‌دهیم و سپس برای چندین سهم نشان می‌دهیم که این راه‌حل چرا مناسب است.

۲/۱. روش اول: بدست آوردن اهمیت کانال ها با استفاده از اکسل خروجی

در روزهای ابتدایی که خروجی را از سامانه دریافت کرده بودیم برای هر برگه یک خروجی از نوع فایل اکسل داشتیم بنابراین باید با استفاده از آن به نوعی یک تحلیل اجتماعی ارئه می‌شد. از چندین جنبه می‌توان تحلیل‌های اجتماعی برگه‌ها را مورد بررسی قرار داد برای مثال تعداد پست‌ها، تعداد بازدید مطالب، تعداد فورواردها و ....

ولی موضوعی که در اینجا مطرح هست تعداد بازدید یا تعداد پست‌های هر کانال می‌تواند تاثیر متفاوتی در خواننده بگذارد. به‌عبارت دیگر این موضوع به این معناست که هر کانال دارای یک اهمیت خاص است و این اهمیت نیز با توجه به داده‌هایی که از شبکه اجتماعی مورد نظر بدست آورده بودیم بر اساس دیدگاه‌های متفاوتی قابل محاسبه بود. بنابراین می‌بایست با توجه به داده‌های در دست داشته در فایل اکسل هر نماد برای هر کانال یک ضریب اهمیت محاسبه کرد.

مرحله اول: یک کد به زبان پایتون نوشته شد که در آن:

I. برای هر کانال تعداد کل پست‌ها و همچنین تعداد کل بازدیدها محسابه شد.

II. تاریخ‌های موجود در فایل اکسل که به صورت کاراکتر فارسی و تاریخ شمسی بودند به تاریخ میلادی معادل و کارکتر انگلیسی تبدیل شدند.

III. خروجی در یک فایل اکسل ذخیره شد.

مرحله دوم: حال که برای هر کانال تعداد کل پست‌ها و تعداد کل بازدیدها را داشتیم با توجه به فرمول زیر یک" ضریب اهمیت[1]" برای هر کانال در فایل اکسل محاسبه شد:

```
LOG10(10+( view-AVERAGE(views))/AVERAGE(views))*LOG10(10+( post -
AVERAGE(posts))/AVERAGE(posts))
```

در ابتدا این موضوع توضیح داده می‌شود که چرا در داخل هر لگاریتم یک عدد ثابت ۱۰ اضافه شده است این موضوع برای خنثی کردن اثر اعداد کمتر از یا مساوی ۹ می‌باشد و ما را خیلی درگیر اعداد کوچک همانند صفر نیز نخواهد کرد، سپس بعد از آن در داخل هر لگاریتم تعداد بازدید کانال مورد نظر منهای میانگین کل بازدیدها را داریم که بر میانگین کل بازدیدها تقسیم شده تا یک عدد نرمال شده بدست آید همین کار را برای تعداد پست‌ها انجام می‌دهیم و نتایج را در هم ضرب می‌کنیم عدد بدست آمده یک ضریب اهمیت برای کانال مورد نظر است.

از آنجایی که داده‌های اقتصادی دریافت شده از سایت و سامانه tsemc.com تاریخ را میلادی در نظر گرفته‌اند و ما به دنبال یک نوع تحلیل اجتماعی هستیم باید برای هر نماد بدانیم که در هر روز چه تعداد پست و مجموع بازدید روزانه پست‌های آن نماد چگونه است.

[1] importance

خروجی کد قبلی که یک فایل اکسل بود در آن نیز ما به محاسبه ضریب اهمیت هر کانال پرداختیم و همچنین تاریخ فارسی و شمسی را به میلادی و انگلیسی تبدیل کردیم یعنی در واقع آن فایل تمام داده‌های مورد نیاز ما را درون خود دارد.

مرحله سوم: بنابراین کد دومی نوشته شد که فایل ذکر شده را دریافت کند و براساس اینکه در هر روز و به تاریخ میلادی چه تعداد پست و چه تعداد بازدید برای آنها داشته‌ایم، سوالی که پیش می‌آید این است که از ضریب اهمیت هر کانال در کجا استفاده کرده‌ایم.

در بین دو معیار تعداد پست و تعداد بازدید در هر روز ما تصمیم گرفتیم که ضریب اهمیت را در تعداد بازدید ضرب کنیم و یک معیار جدید که **تعداد بازدید وزن‌دار شده** براساس ضریب اهمیت کانال می‌باشد را ایجاد کنیم. دلیل این کار هم این است هرچه تعداد بازدیدها بالاتر باشد ممکن است تاثیر بیشتری بر خواننده پیام بگذارد همچنین نمی‌توان به تعداغد بازدید به تنهایی نیز بسنده کرد بنابراین تعداد بازدید در ضریب اهمیت آن کانال ضرب شده و مجموع وزن‌دار تعداد بازدیدها برای هر روز محاسبه شده است.

به نوعی شاید بتوان گفت که یک معیار بازگشتی داریم چون ضریب اهمیت براساس تعداد بازدید و تعداد پست محاسبه می‌شود و سپس از خود آن ضریب برای وزن‌دار کردن تعداد بازدیدها استفاده می‌کنیم، البته همین عمل را بر روی تعداد پست‌ها نیز می‌توان انجام داد ولی از آنجایی که تعداد بازدیدها در دل خود تعداد پست‌ها و اهمیت هر نماد در روز مشخص شده را دارد، تصمیم بر آن شد که تعداد بازدیدها وزن‌دار شوند و در تحلیل‌های بعدی مورد استفاده قرار گیرند.

۲/۲. روش دوم: بدست آوردن اهمیت کانال ها با استفاده از گراف خروجی

پس از دریافت گراف نمادها از سامانه دو کار مد نظر قرار گرفتند،

در وهله اول اینکه ضریب اهمیت بدست آمده شده در توضیحات بالا را با مقدار معادل آن از گراف بررسی کنیم و در وهله دوم اینکه به دنبال تحلیل گراف‌های بدست آمده باشیم.

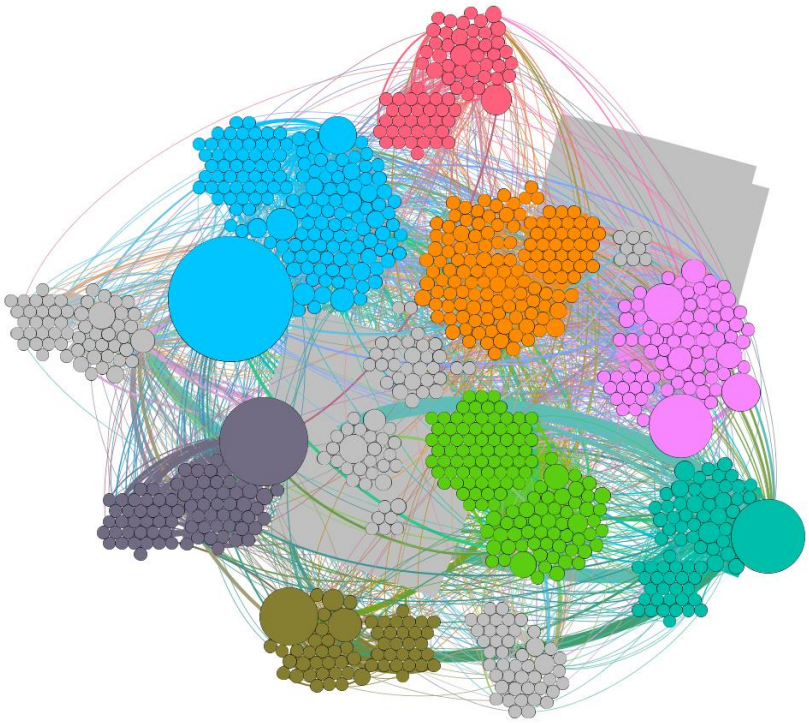
ضریب اهمیت بدست آمده از گراف با ضریب اهمیت محاسبه شده توسط کد ما اندکی تفاوت دارد که این تفاوت می‌تواند ناشی از آن باشد که برخی داده‌ها در گراف نیامده باشند و در محاسبات دخیل نشوند در حالیکه آنها در فایل اکسل موجود بوده باشند البته این اختلاف جزئی است و قابل صرف نظر است.

پس از سه معیار اجتماعی تعداد پست روزانه، تعداد بازدید روزانه و تعداد بازدید وزن‌دار شده با ضریب اهمیت هر کانال و دریافت گراف‌ها به تحلیل گراف‌ها مشغول شدیم که شاید بتوان معیاری دیگر را نیز ارائه داد تحلی تعدادی از گراف‌های گروه در ادامه آورده می‌شود.

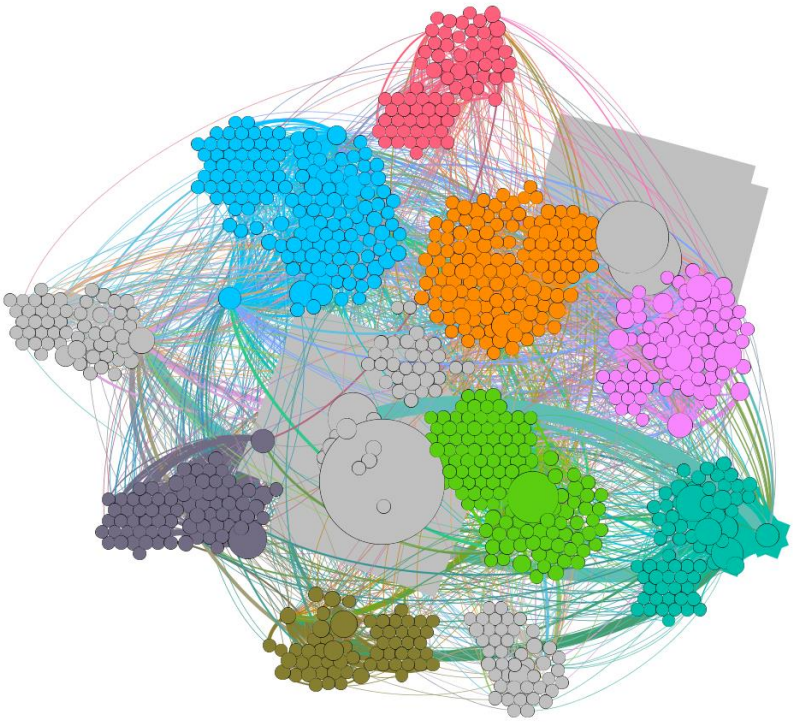
- **تحلیل گراف نماد شپترو :**

برای پیدا کردن یک تحلیل مناسب بهتر است با استفاده از الگوریتم ماژولاریتی موجود در نرم‌افزار گفی اجتماع‌های گراف را پیدا کنیم در شکل زیر اجتماعات گراف مربوط به نماد شپترو نمایش داده شده‌اند،

در یک شکل اندازه نودها بر اساس تعداد پست‌ها می‌باشد در حالیکه در شکل دوم اندازه نودها براساس تعداد بازدیدها می‌باشد.



شکل ۱: اجتماعات گراف نماد شپترو(اندازه نود ها براساس تعداد بازدید)



شکل ۲: اجتماعات گراف نماد شپترو(اندازه نود ها براساس تعداد پست)

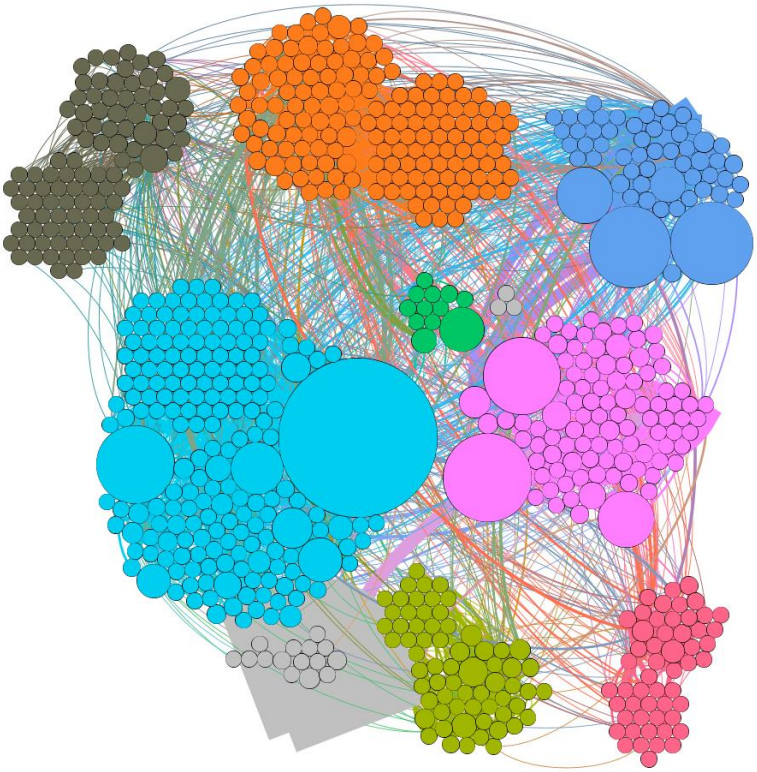
در شکل ۱، اندازه نودها بر اساس تعداد بازدیدها تعیین شده این موضوع نشان می‌دهد در اجتماعات بزرگتر تعداد نودها با تعداد بازید بالاتر موجود هستند. در حالیکه اگر به شکل ۲ توجه کنیم که اندازه نودها براساس تعداد پست‌ها تعیین شده است در اجتماعات کوچکتر نودهایی وجود دارند که تعداد پست بالایی دارند این تصاویر نشان می‌دهند که معیار تعداد بازدیدها می‌تواند معیار خوبی باشد و اگر موضوع وزن‌دار شدن آن همراه با ضریب اهمیت کانال را نیز در نظر بگیریم می‌تواند معیار بهتری از تعداد بازدیدها به تنهایی باشد.

همچنین یکی دیگر از ایده‌های ما برای تحلیل اجتماعی نحوه دیگری وزن دادن به کانال‌ها بود به این‌صورت که آنهایی که به‌طور تخصصی راجع‌به سهم شپترو صحبت کرده‌اند وزن بیشتری داشته باشند ولی با بررسی اجتماعات موجود در گراف تصاویر بالا اکثر کانال‌های موجود کانال‌هایی بودند که به‌صورت عمومی در مورد بورس صحبت کرده بودند حتی بعضی از کانال‌هایی که در تصاویر بالا بزرگتر رسم شده‌اند در حال حاضر وجود خارجی ندارند، که این به این علت می‌تواند باشد که در زمان گرفتن دیتا کانال‌ها عمومی بوده‌اند ولی هم‌اکنون دسترسی عمومی به آنها وجود ندارد.

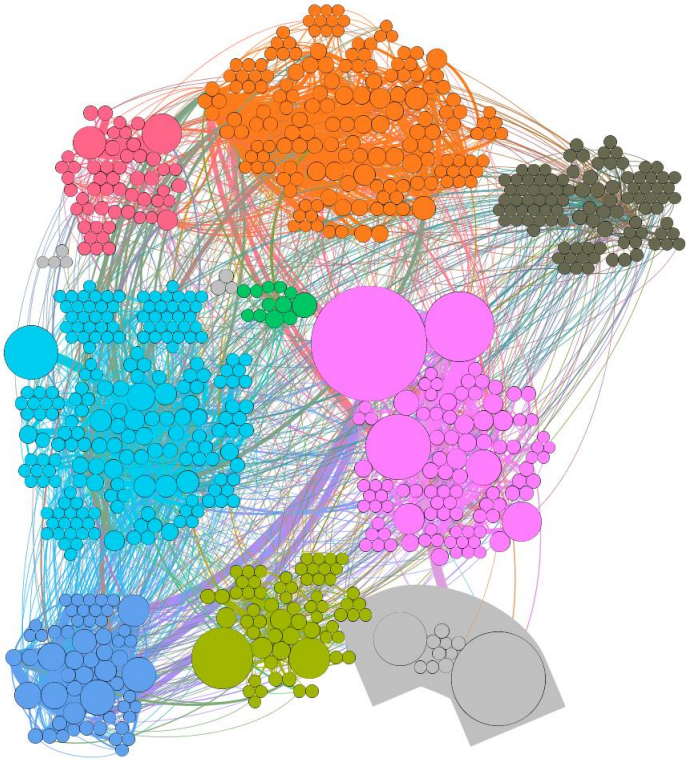
در مجموع توضیحات بالا نشان می‌دهد که معیارهای اولیه ارائه شده قبل از بررسی گراف‌ها معیارهای نسبتا خوبی بوده‌انده که بررسی گراف‌ها این مورد را تصدیق می‌کند اما بر اساس یک گراف از کل گراف‌ها نمی‌توان به‌سرعت همچین نتیجه‌ای گرفت بنابراین به‌سراغ برسی چند سهم دیگر نیز در ادامه خواهیم رفت.

- **تحلیل گراف نماد حپترو :**

حال به بررسی نماد حپترو می‌پردازیم:



شکل ۳: اجتماعات گراف نماد حیترو(اندازه نود ها براساس تعداد بازدید)



شکل٤: اجتماعات گراف نماد حیترو(اندازه نود ها براساس تعداد پست)

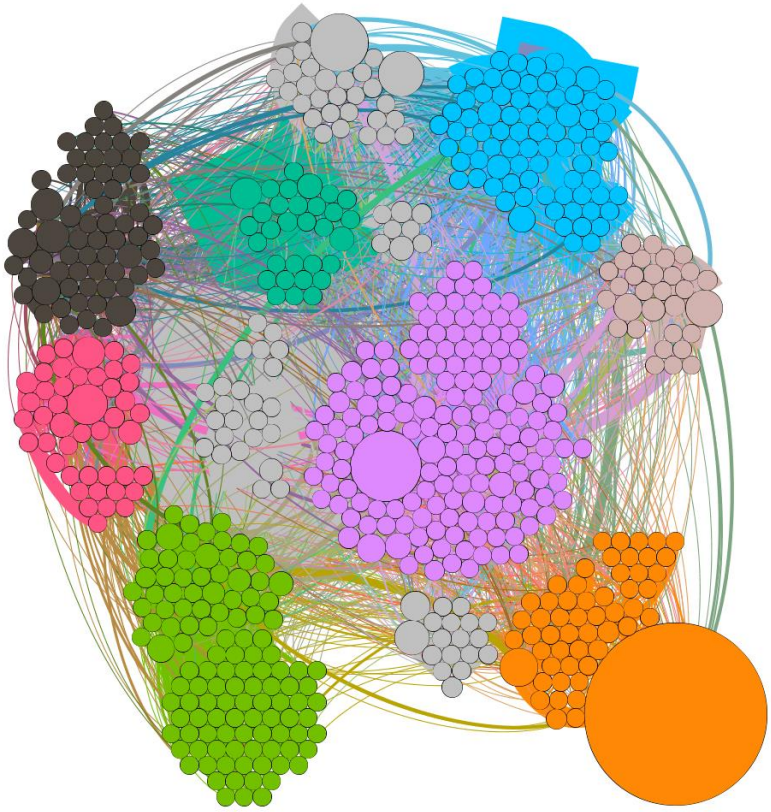
در شکل ۳ گراف نماد حپترو به‌صورت اجتماع‌های متفاوت و همچنین اندازه هر نود از تعداد بازدیدهای آن گرفته شده است و در شکل ۴ نیز همین گراف فقط اندازه نودها ناشی از تعدا پست‌ها می‌باشد.

همانند تحلیلی که در نماد شپترو داشتیم اینجا هم می‌توان نتیجه گرفت که با توجه به تصاویر بالا معیار تعداد بازدید وزن‌دار شده با ضریب اهمیت کانال می‌تواند معیار مناسبی باشد.

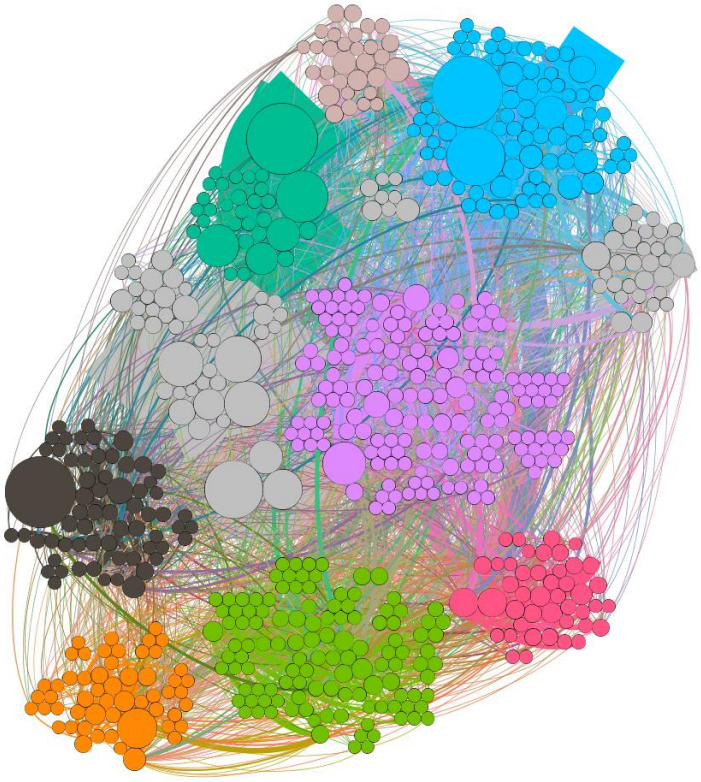
- تحلیل گراف نماد پارسان:

دو تصویر زیر برای نماد پارسان بدست آمده است، تصویر شماره ۵ گراف مربوط به نماد پارسان می‌باشد که به کمک الگوریتم ماژولاریتی اجتماع‌های آن تشخصی داده شده‌اند همچنین اندازه نودها متاثر از تعداد بازدید هر کانال می‌باشد در حالیکه در تصویر شماره ۶ اندازه نود متاثر از تعداد پست‌ها ولی همچنان اجتماعات پیدا شده همان اجتماع‌های موجود در تصویر شماره ۵ می‌باشد.

با توجه به اینکه در تصویر دوم تعداد نود بزرگ هستند ولی در اجتماع‌های بزرگ قرار ندارند و مانند تحلیل‌های گذشته می‌توان نتیجه گرفت که معیار اولیه‌ی همچنان می‌تواند به خوبی عمل کند



شکل۵: اجتماعات گراف نماد پارسان(اندازه نود ها براساس تعداد بازدید)



شکل ٦: اجتماعات گراف نماد پارسان(اندازه نود ها براساس تعداد پست)

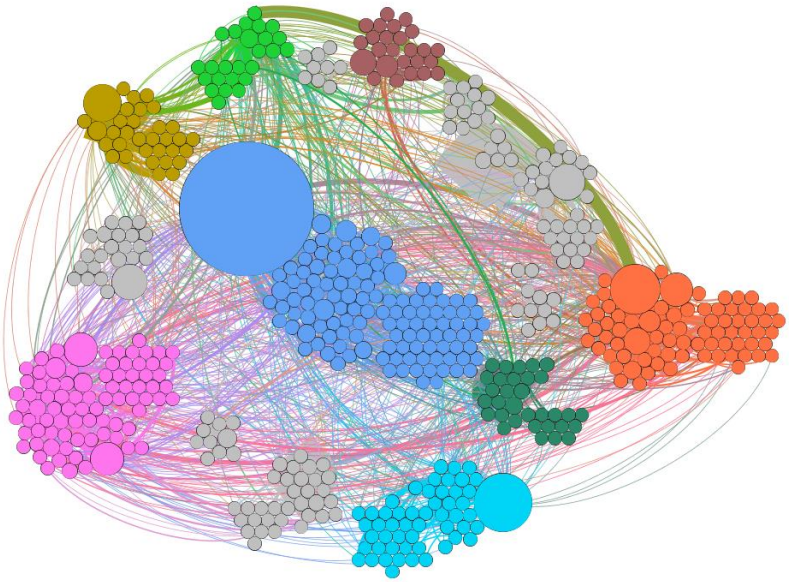
- تحلیل گراف نماد پارس:

دو تصویر زیر برای نماد پارس بدست آمده است، شکل ۷ گراف مربوط به نماد پارس می‌باشد که به کمک الگوریتم ماژولاریتی اجتماع‌های آن تشخیص داده شده‌اند همچنین اندازه نودها متاثر از تعداد بازدید هر کانال می‌باشد در حالیکه در تصویر شماره ۸ اندازه نود متاثر از تعداد پست‌ها ولی همچنان اجتماعات پیدا شده همان اجتماع‌های موجود در تصویر شماره ۷ می‌باشد.

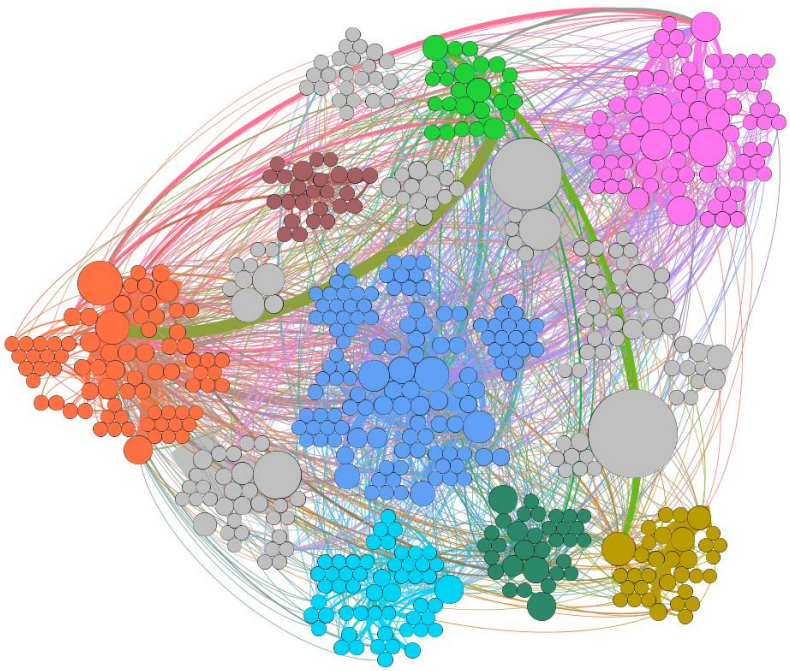
با توجه به اینکه در تصویر دوم تعداد نود بزرگ هستند ولی در اجتماع‌های بزرگ قرار ندارند و مانند تحلیل‌های گذشته می‌توان نتیجه گرفت که معیار اولیه‌ی همچنان می‌تواند به خوبی عمل کند.

البته ذکر یک نکته درباره‌ی تمام تصاویر گراف‌هایی که در بالا نمایش داده شد ضروری است و آن نکته این است که برای دید بهتر نسبت به گراف و تحلیل ساده‌تر الگوریتم k-core برای همه‌ی گراف‌های بالا اجرا شده است.

برای تصاویر ۵و۶ k-core برابر ۴ وبرای بقیه تصاویر این مقدار برابر ۲ در نظر گرفته شده‌است، همچنین برای نمایش گراف از circular pack layout استفاده شده است.



شکل۷: اجتماعات گراف نماد پارس(اندازه نود ها براساس تعداد بازدید)



شکل۸: اجتماعات گراف نماد پارس(اندازه نود ها براساس تعداد پست)

با توجه به گراف‌ها و اینکه اکثر کانال‌ها که همان نودها می‌باشند در تحلیل گراف ما بیشتر کانال‌های عمومی بورس می‌باشند و با توجه به بررسی تعدادی زیادی از این کانال‌ها که کانال‌های تخصصی در مورد کل بورس و نه فقط سهام پتروشیمی بودند می‌توان نتیجه گرفت که معیار معرفی شده ابتدایی یعنی تعداد بازدید وزن‌دار شده با ضریب اهمیت کانال می‌تواند معیار خوبی برای تحلیل باشد، البته نمی‌توان گفت که کانال‌های تخصصی برای سیگنال‌های تخصصی یا برای یک مجموعه سهام وجود ندارد بعضا دیده شده است که همچین کانال‌هایی وجود دارند ولی این کانال‌ها برای عضویت هزینه دریافت می‌کنند و به‌صورت خصوصی هستند وبه آنها دسترسی وجود ندارد.

با توجه به توضیحات داده شده و همچنین تحلیل گراف‌ها برای مراحل بعدی پروژه از سه معیار تعداد پست در هر روز، تعداد بازدید در هر روز و تعداد بازدید وزن‌دار شده با ضریب اهمیت کانال در هر روز استفاده می‌شود البته بیشتر از معیار سوم استفاده خواهد شد.

۳. پیشبینی قیمت با استفاده از یادگیری با نظارت[۲]

در این بخش ما می‌خواهیم خواسته پروژه که پیشبینی قیمت هر نماد است را بدست بیاوریم. برای این کار از یادگیری بردار پشتیبانی[۳] که یک روش یادگیری با نظارت می‌باشد استفاده کردیم. در این روش ما با استفاده از مجموعه ای از جفت های ورودی و خروجی به سیستم می‌دهیم. و سیستم تابعی از ورودی به خروجی را فرا گیرد.

یادگیری بردار پشتیبانی یا SVM ، یک روش یادگیری با نظارت است که از آن برای طبقه بندی داده ها و تشخیص الگو استفاده می شود.

در پیاده سازی ما از کتابخانه sklearn و ابزار support vector machine (svm) استفاده شده است. در این بخش در دو حالت لرنینگ انجام داده شده. یک بار به کمک داده های اقتصادی سایت tsetmc.com و یک بار به کمک این داده های اقتصادی و داده های اجتماعی که در مراحل قبلی برای هر روز با در نظرگیری ضریب اهمیت هر کانال بدست آمده است.

در مرحله اول دو داده ای که به عنوان ورودی های اقتصادی به ماشین لرنینگ میدهیم در دو تابع محاسبه میشوند. این داده ها میانگین مومنتوم قیمت و نوسان قیمت هستند که از روی قیمت بسته شدن سهم در روزهای گذشته محاسبه میشود.

- مومنتوم قیمت به صورت مقابل محاسبه میشود.

$$momentum = \begin{cases} 1 & closing\ price[i] > closing\ price\ [i-1] \\ -1 & closing\ price[i] < closing\ price\ [i-1] \end{cases}$$

که از این مقدار در سه روز گذشته میانگین گرفته شده است.

- برای محاسبه نوسان قیمت نیز از فرمول روبرو استفاده شده است.

$$volatility = \frac{closing\ price\ [i] - closing\ price\ [i-1]}{closing\ price\ [i-1]}$$

[2] Supervised Learning
[3] Support vector machines

که از این مقدار در سه روز گذشته میانگین گرفته شده است.

برای انجام لرنینگ روی داده های اقتصادی به تنهایی دو متغیر بالا در نظر گرفته شده اند. اما برای انجام لرنینگ روی داده های اقتصادی به همراه اجتماعی دو متغیر دیگر نیز به صورت زیر تعریف شده اند.

داده تعریف شده اول مومنتوم تعداد بازدید (با در نظر گرفتن ضریب اهمیت هر کانال) می باشد. که از طریق زیر محاسبه شده است.

$$social\ momentum = \begin{cases} 1 & imp.views[i] > imp.views\ [i-1] \\ -1 & imp.views[i] < imp.views\ [i-1] \end{cases}$$

که از این مقدار در سه روز گذشته میانگین گرفته شده است.

داده تعریف شده بعدی نوسان تعداد بازدید (با در نظر گرفتن ضریب اهمیت هر کانال) می باشد.

$$volatility = \frac{imp.views\ [i] - imp.views[i-1]}{imp.views\ [i-1]}$$

که از این مقدار در سه روز گذشته میانگین گرفته شده است.

در مرحله بعدی دیتا را آماده کرده و سپس خروجی طبقه بندی شده را برای یادگیری به صورت زیر تعریف میکنیم.

$$Y = \begin{cases} 1 & \frac{closing\ price[i+1] - closing\ price\ [i]}{closing\ price[i]} > 0.005 \\ 0 & -0.005 > \frac{closing\ price[i+1] - closing\ price\ [i]}{closing\ price[i]} > 0.005 \\ -1 & \frac{closing\ price[i+1] - closing\ price\ [i]}{closing\ price[i]} < -0.005 \end{cases}$$

در فرمول بالا اگر یک نماد در روز بیشتر از ۰/۵ درصد مثبت شود ۱ به عنوان خروجی ذخیره می‌شود. اگر بین -۰/۵ تا +۰/۵ تغییر کند، صفر به آن نسبت داده می‌شود و اگر کمتر از ۰/۵ درصد در روز سقوط کند، -۱ به آن نسبت می‌دهیم.

سپس به کمک دستور split داده های موجود را به دو بخش ترین (۷۵ درصد دیتا) و تست (۲۵ درصد دیتا) تقسیم می‌کنیم. در این بخش داده های ترین و تست را پشت سر هم گرفتیم. یعنی فرض بر این بوده که زمان گذشته کل داده های ترین باشد. و داده های تست همان پیشبینی ما از آینده هستند. سپس داده های بخش قبل را استاندارد سازی کرده و به ماشین میدهیم. پس از ترین و تست کردن درصد درستی پیشبینی نسبت به مقدار واقعی برای داده های تست، برای هر نماد در جدول زیر آمده است. [۴]

Financial + Social Score	Financial ML Score	نام نماد
59.39%	54.44	پارس
69.73%	56.79%	پترول
65.47%	66.69%	شاراک
73.91%	72.34%	شغدیر
70.90%	70.82%	وپترو
73.5	37.5	حپترو
71.64	70.15	شپترو
64.78	61.88	شفن
57.42	55.69	شیراز
52.87	50.32%	شخارک
57.44%	56.5%	بوعلی
63.3%	61.9%	پارسان
76.15%	74.1%	شاروم
67.8%	67.6%	شپدیس
67%	62.2%	تاپیکو

[۴] داده ها در پوشه 4_ML_results آمده است.

۳/۱. نتیجه گیری:

همانطور که می‌بینیم، داده های ترکیبی در اکثر نمادها خروجی بهتری از داده های صرفا اقتصادی داده‌اند. پس در پیش‌بینی قیمت داده‌های ترکیبی اجتماعی-اقتصادی می‌توانند اندیکاتور بهتری نسبت به اندیکاتور داده های اقتصادی بسازند.

البته در دو نماد مانند شاراک و تاپیکو، تحلیل اقتصادی بهتر عمل کرده است. که این می‌تواند ناشی از این باشد که داده‌های ما سنتیمنت ندارند و ممکن است در برخی نمادها داده های اجتماعی ما به خوبی عمل نکنند.

۴. نمایش اطلاعات و تحلیل نموداری داده ها

برای پلات کردن اطلاعات ما از کتابخانه mplfinance که در گذشته زیرمجموعه ای از MatPlotLib بوده است، استفاده می‌کنیم. این کتابخانه ابزار هایی برای نمایش اطلاعات اقتصادی و ساختن اندیکاتور ها و تحلیل های تکنیکال دارد.

در این قسمت ما ابتدا فایل‌های اطلاعات تحلیل اجتماعی و اطلاعات اقتصادی نماد را می‌خوانیم. سپس به ترتیب قسمت های زیر را بر روی پلات ترسیم می‌کنیم.

در این قسمت ما بر روی نمودار قسمت های مختلفی را نمایش می‌دهیم:

- **نمایش قیمت بسته شدن نماد بر روی پلات {Close Price}**

قیمت <close> هر نماد در بازه زمانی مشخص شده در هر روز بر روی پلات نمایش داده می‌شود.

قیمت روزانه نماد بر روی پلات با رنگ **قرمز** مشخص می‌شود.

- **نمایش مجموع وزن دار تعداد بازدید در روز (اندیکاتور اقتصادی)**

مجموع وزن دار تعداد بازدید را که با نماد <importace Views> مشخص کردیم در هر روز نمایش می‌دهیم. البته از آن‌جایی که تعداد بازدید در هر روز تغییرات زیادی داشت، میانگین متحرک ۱۵ روزه آن را محاسبه کرده و بر روی نمودار پلات کردیم.

مجموع وزن دار تعداد بازدید در روز با رنگ **آبی** مشخص می‌شود.

- **نمایش نقطه ورود های داده شده توسط یادگیری ماشین**

همانطور که گفته‌شد، در روزهایی که ماشین پیشبینی روز مثبت می‌کند خروجی آن روز را ۱ می-دهد. ما تمام خروجی های ۱ که پیشبینی روز های مثبت هستند را بر روی نمودار پلات می‌کنیم. نقاط ورود با فلش های **آبی**(مانند شکل ۱۰) نمایش داده می‌شود.

- **نمایش حجم[5] بازار در هر روز**

در قسمت زیرین نمودار <volume> آن سهم در هر روز نمایش داده می‌شود.

- **ترسیم میانگین های متحرک[6]**

اندیکاتور <moving average> را برای دو بازه زمانی ۵ روزه و ۲۰ روزه نمایش می‌دهیم.

خط میانگین متحرک ۵ روزه با رنگ **نارنجی** و میانگین متحرک ۲۰ روزه با رنگ **سبز** مشخص می‌شود.

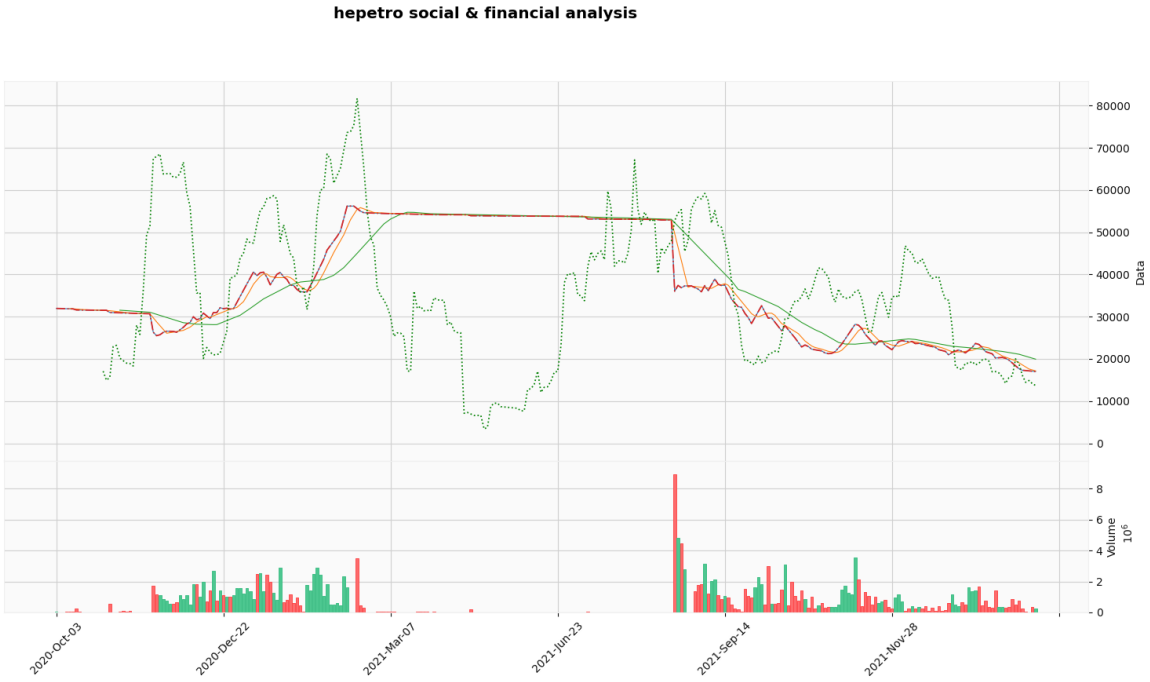
❖ تحلیل داده های حپترو:

همانطور که در قسمت قبل گفته شد. یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۷۳.۵ درصد موفقیت داشت. به همین علت ما در پلات، خروجی یادگیری ماشین در مرحله قبل را به عنوان نقطه های ورود را(که همان خواسته این پروژه است)، با فلش نمایش داده‌ایم.

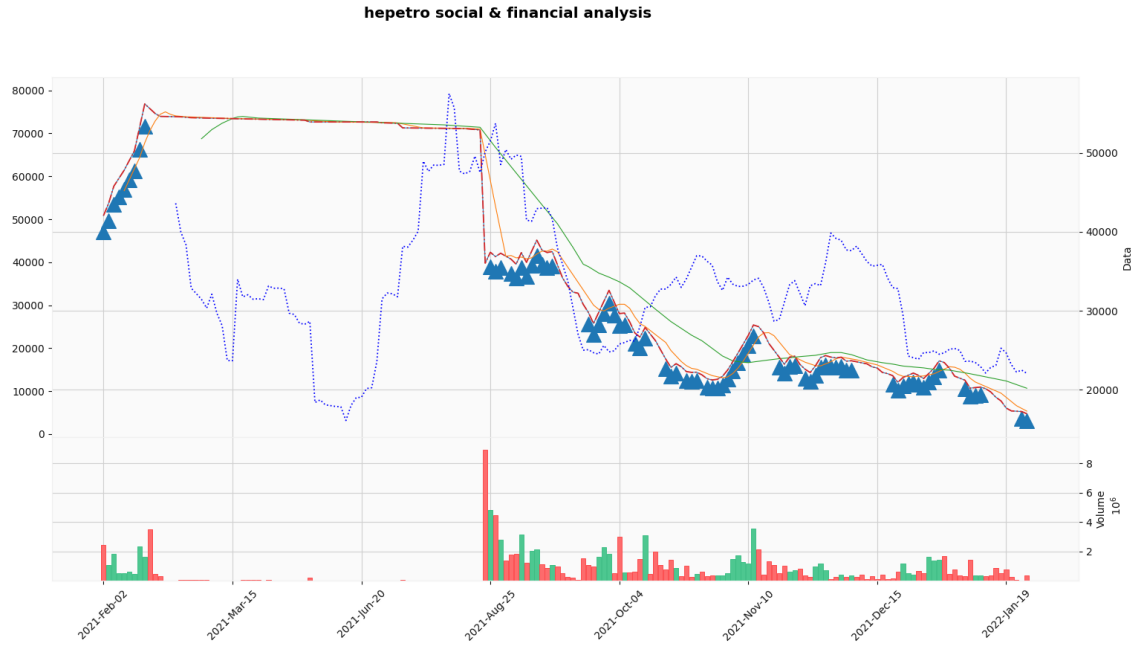
{تصویر تمام بررسی ها علاوه بر ورد در فولدر 5_Plots موجود است.}

در شکل۹ میبینیم که در قسمت پایین اندیکاتور حجم معاملات هر روز و در قسمت بالایی پلات، تغییرات نماد و میانگین های متحرک به همراه مجموع وزن‌دار بازدید ها اضافه شده است. همانطور که دیده می‌شود، در هنگامی که نماد حپترو بسته است، مجموع وزن‌دار بازدید ها نیز بسیار کاهشی شده است. و در بیشتر بازه های زمانی، سیر صعودی و نزول نماد را دنبال می‌کند. که به همین علت است که این داده ها در یادگیری درصد بالایی کسب کردند.

[5] Volume
[6] Moving average



شکل ۹: تحلیل داده های اقتصادی+ اجتمعا نماد حپترو{در اینجا اندیکاتور اقتصادی سبز است}



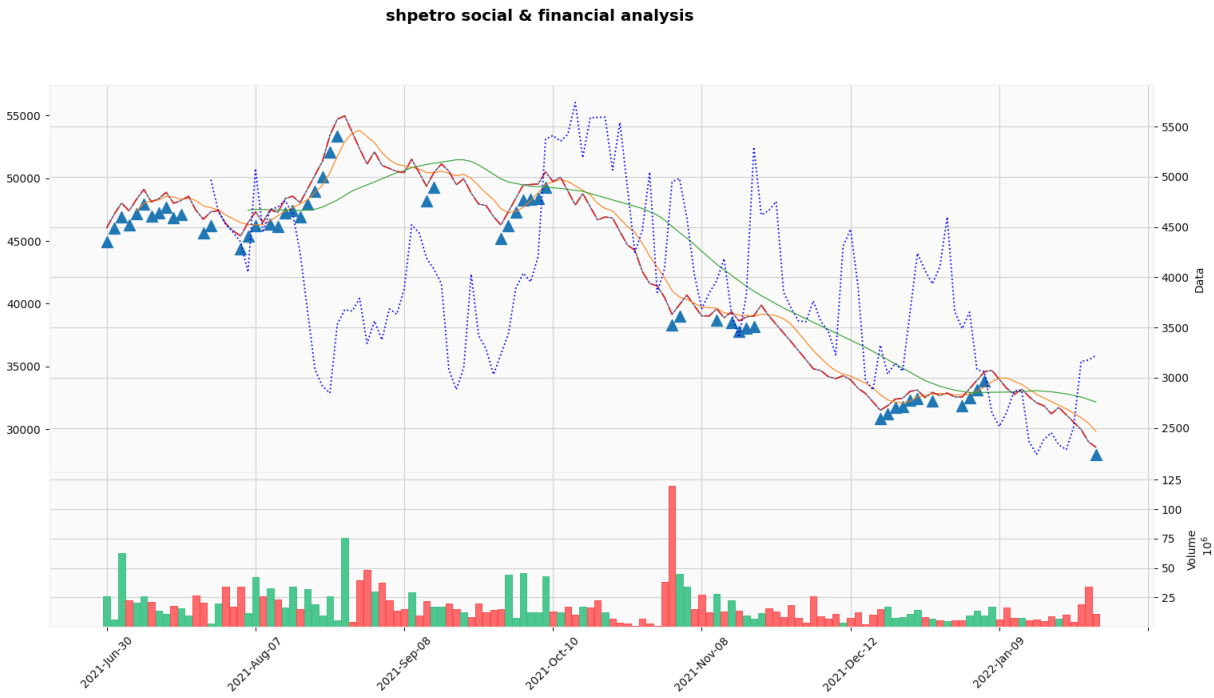
شکل ۱۰: تحلیل داده های اقتصادی+ اجتماعی نماد حپترو به همراه تعیین نقطه ورود

همانطور که در شکل ۱۰ دیده می‌شود، فلش های آبی نقاط پیشبینی یادگیری هستند که در آن روز مثبت خواهند شد.

❖ تحلیل داده های شپترو:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۷۱.۶ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.

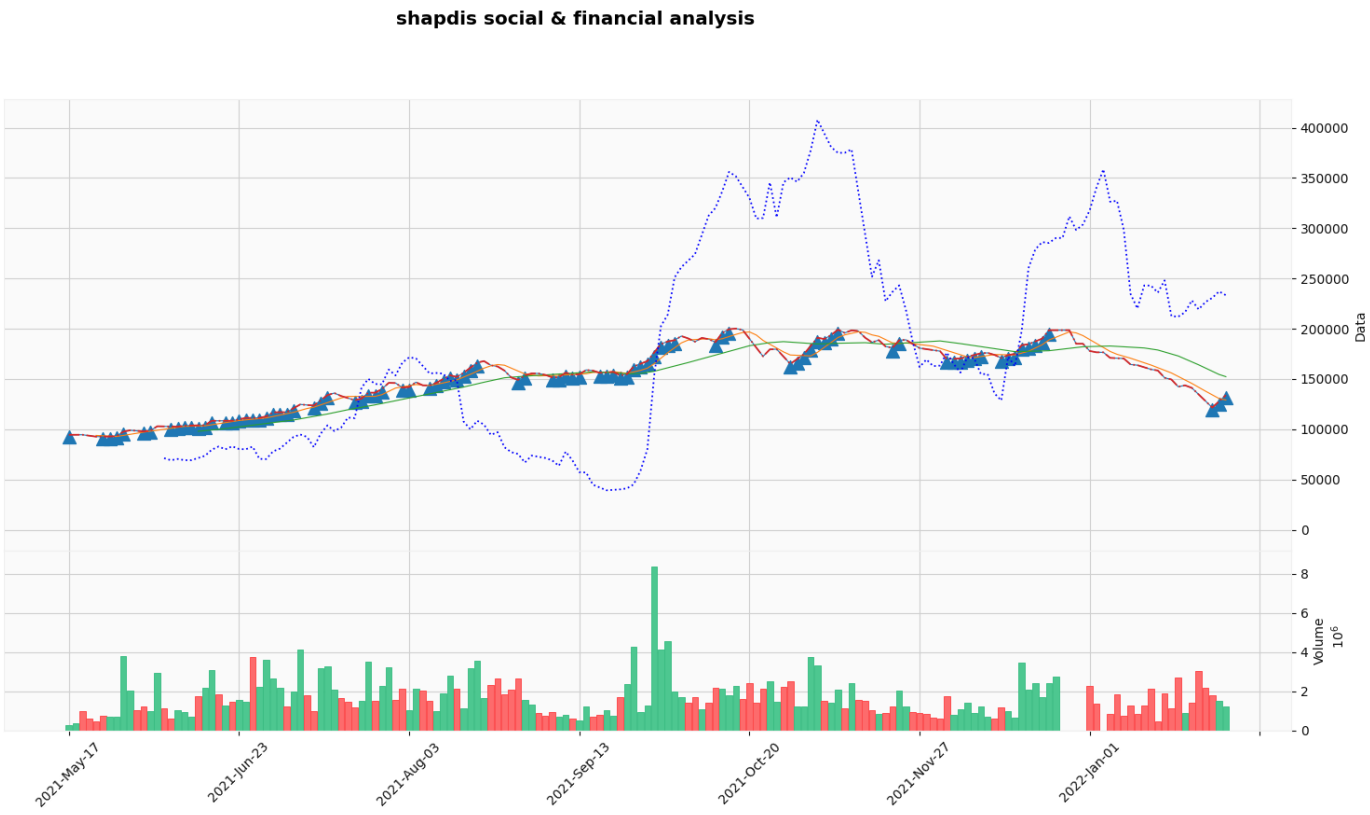


شکل۱۱: تحلیل داده های اقتصادی+ اجتماعی نماد شپترو به همراه تعیین نقطه ورود

❖ تحلیل داده های شپدیس:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۶۷.۸ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند. اگر دقت کنیم میبینیم که نمودار میانگین وزن‌دار بازدیدها نیز سیر صعودی و نزولی را به خوبی شناسایی می‌کند.



شکل۱۲: تحلیل داده های اقتصادی+ اجتماعی نماد شپدیس به همراه تعیین نقطه ورود

❖ تحلیل داده های پارسان:

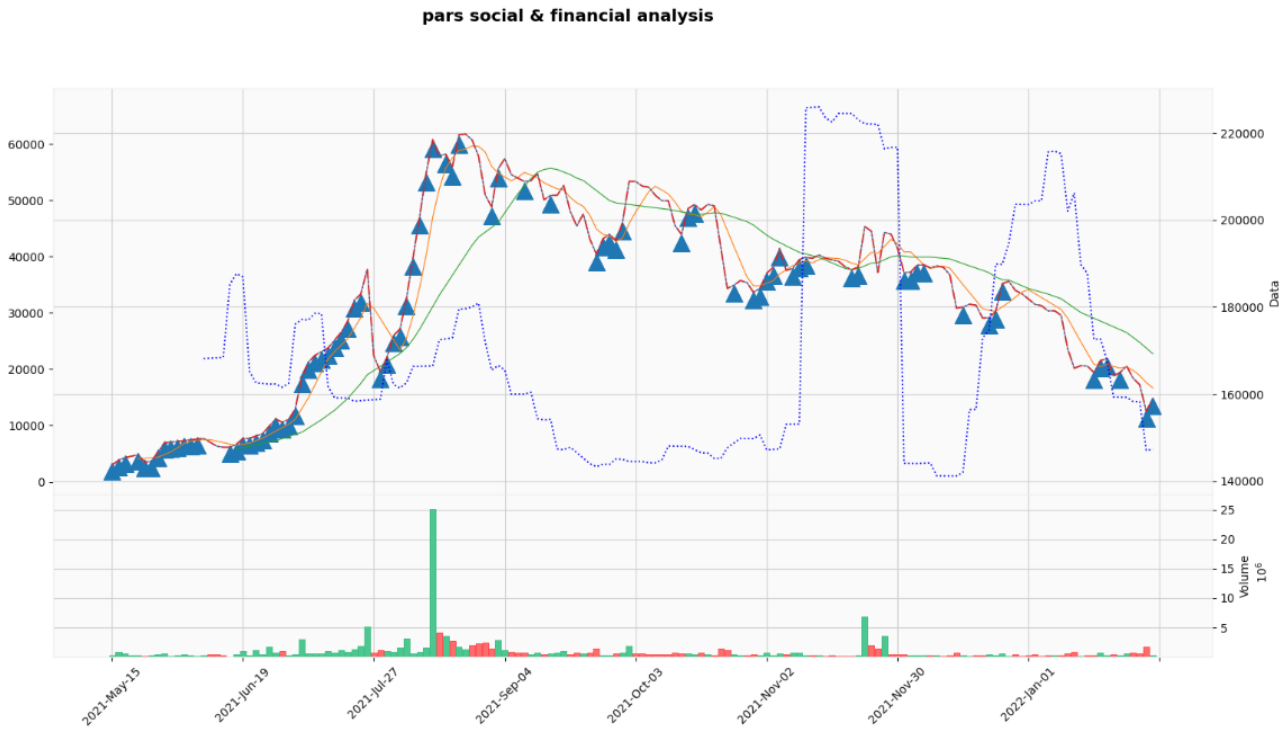
در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۶۷.۸ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند. در اینجا اندیکاتور(اقتصادی+اجتماعی) بسیار خوب عمل کرده و حتی در جاهایی که میانگین وزن‌دار شده بازدید ها نتوانسته سیر قیمت را نشان دهد، صعودی بودن نماد را تشخیص داده است.

شکل۱۲: تحلیل داده های اقتصادی+ اجتماعی نماد پارسان به همراه تعیین نقطه ورود

❖ تحلیل داده های پارس:

در این نماد ابهام‌دار، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۵۹.۸ درصد موفقیت داشت. همانطور که می‌بینیم، نقطه های ورود تقریبا به خوبی پیش‌بینی شده‌اند اما چون یک نماد ابهام دار است، در برخی موارد میانگین وزددار شده بازدید پست ها نتوانسته سیر تغییرات نماد را نشان دهد. اما اندیکاتور نقطه ورود را به خوبی شناسایی کرده است.

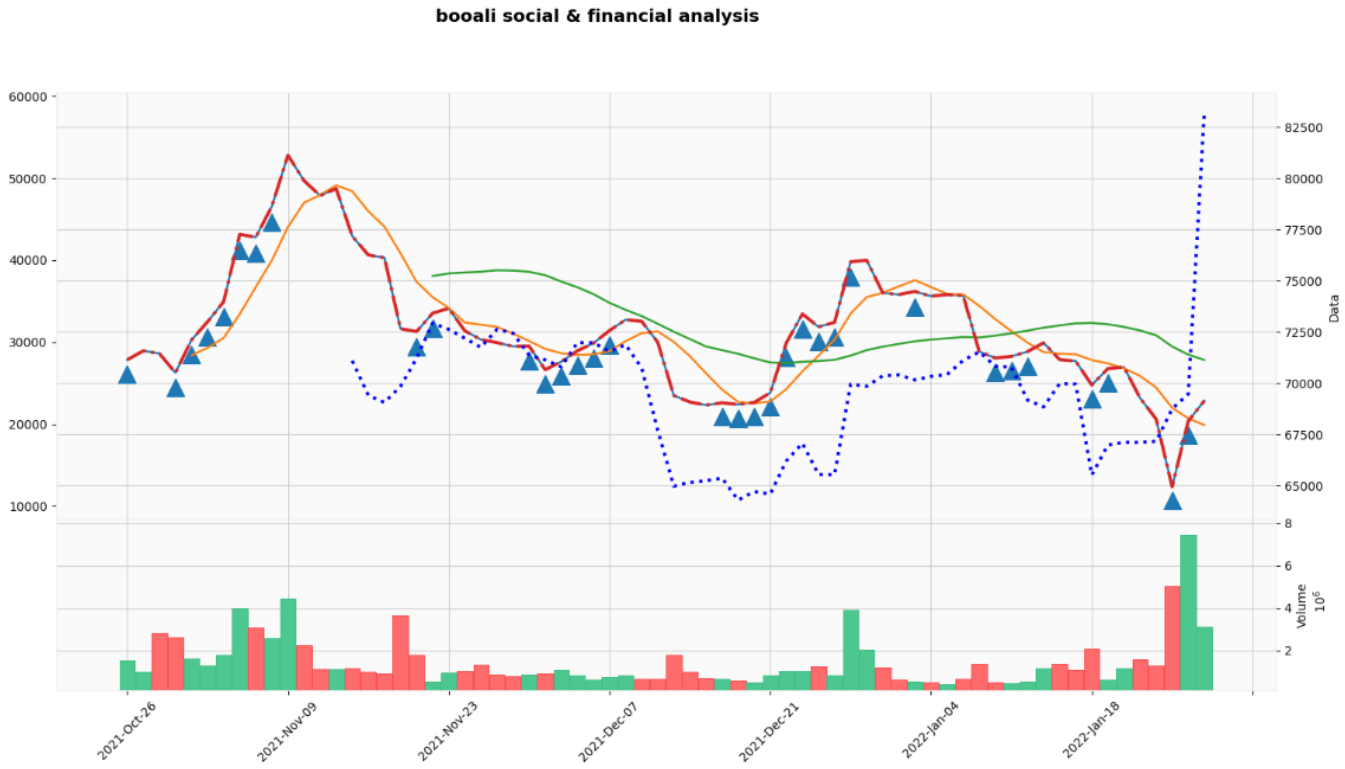


شکل۱۳: تحلیل داده های اقتصادی+ اجتماعی نماد پارس به همراه تعیین نقطه ورود

❖ تحلیل داده های بوعلی:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۵۷.۸ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.

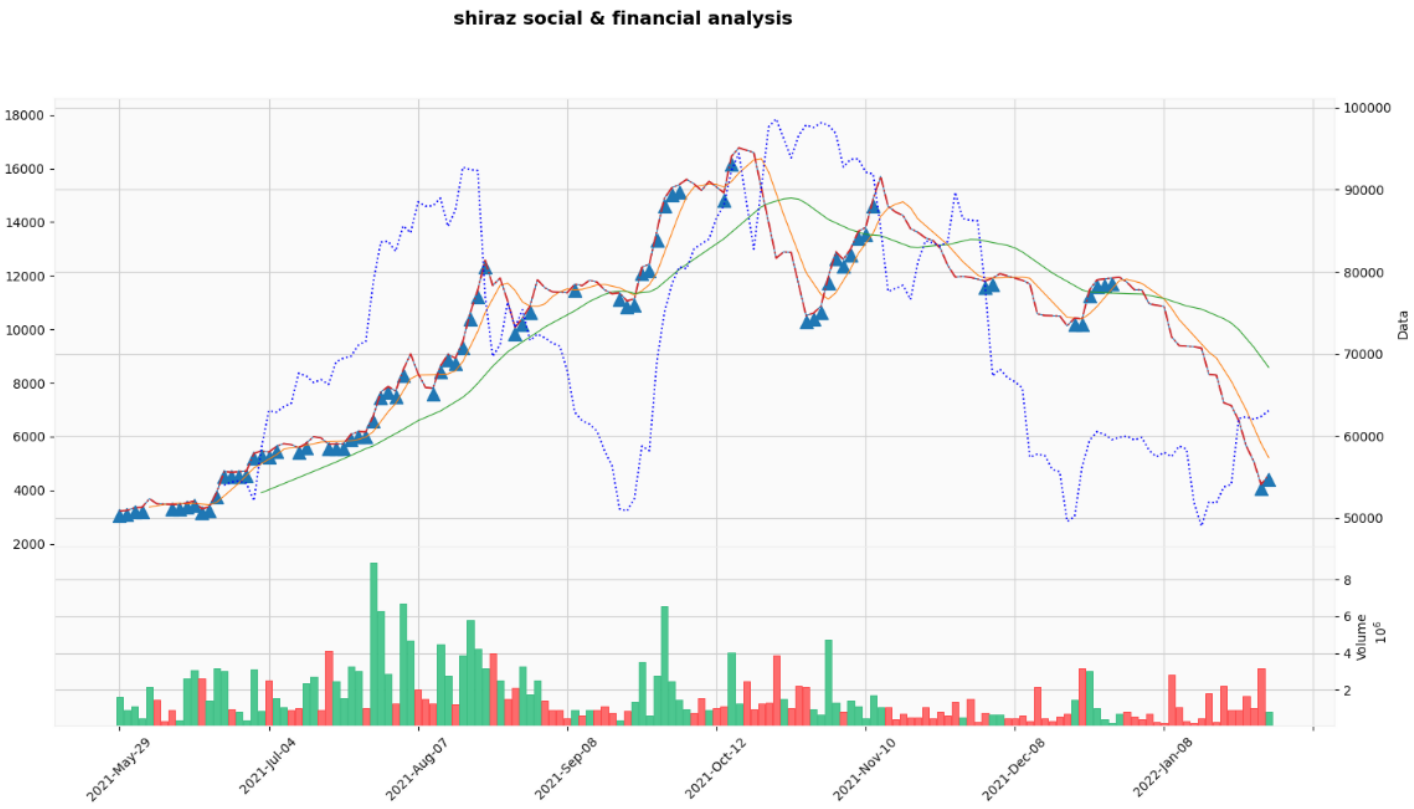


شکل۱۴: تحلیل داده های اقتصادی+ اجتماعی نماد بوعلی به همراه تعیین نقطه ورود

❖ تحلیل داده های شیراز:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۵۷.۴ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.



شکل۱۵: تحلیل داده های اقتصادی+ اجتماعی نماد شیراز به همراه تعیین نقطه ورود

❖ تحلیل داده های پترول:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۶۹.۸ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.

شکل۱۶: تحلیل داده های اقتصادی+ اجتماعی نماد پترول به همراه تعیین نقطه ورود

❖ تحلیل داده های وپترو:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۷۰.۹ درصد موفقیت داشت .

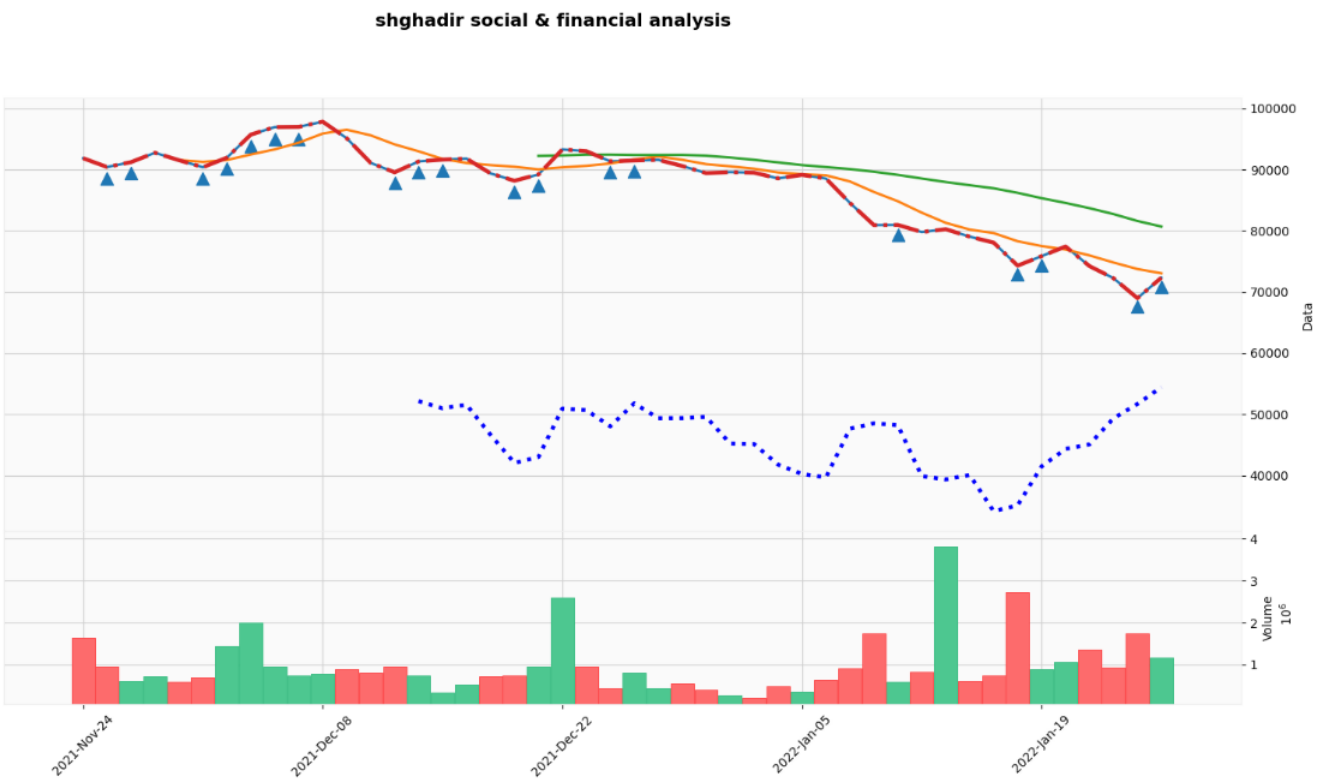
همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.

شکل۱۷: تحلیل داده های اقتصادی+ اجتماعی نماد وپترو به همراه تعیین نقطه ورود

❖ تحلیل داده های شغدیر:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۷۳.۹۱ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.



شکل۱۸: تحلیل داده های اقتصادی+ اجتماعی نماد شغدیر به همراه تعیین نقطه ورود

❖ تحلیل داده های شاراک:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۶۵.۴۷ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.

شکل۱۹: تحلیل داده های اقتصادی+ اجتماعی نماد شاراک به همراه تعیین نقطه ورود

❖ تحلیل داده های تاپیکو:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۶۷درصد موفقیت داشت .

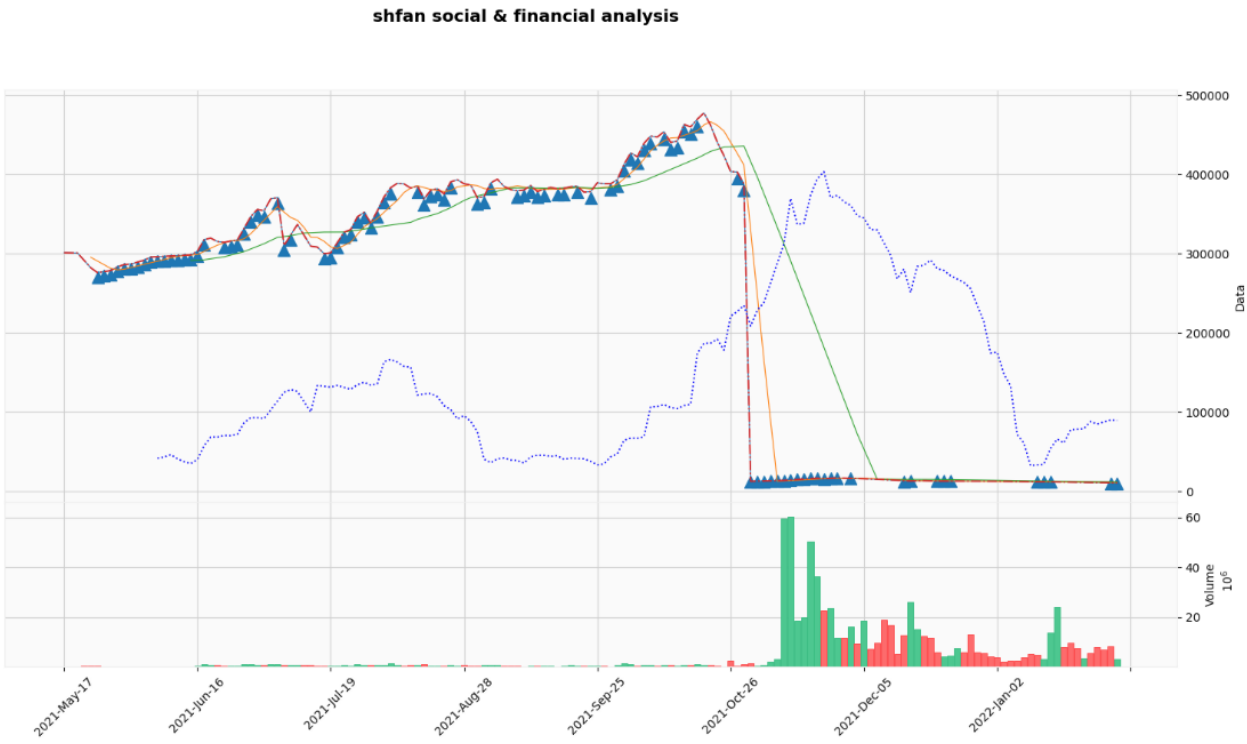
همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.

شکل۲۰: تحلیل داده های اقتصادی+ اجتماعی نماد تاپیکو به همراه تعیین نقطه ورود

❖ تحلیل داده های شفن:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۶۴.۸ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند.

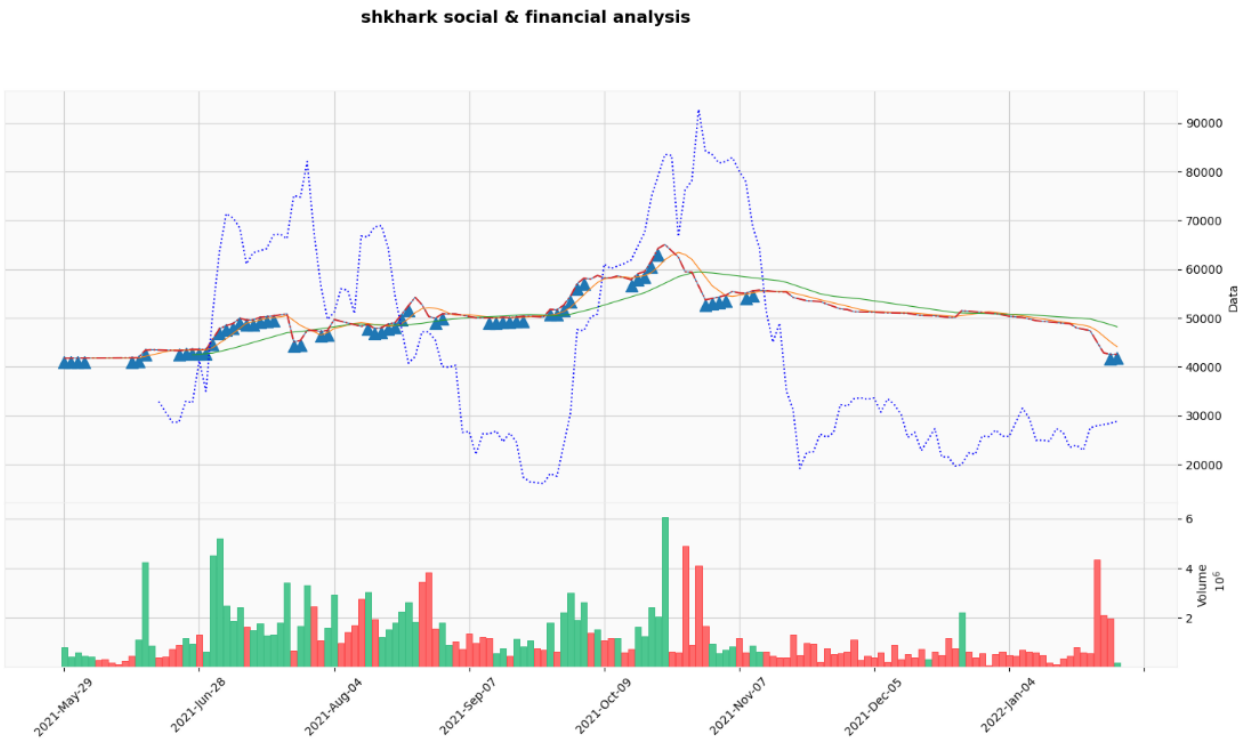


شکل۲۱: تحلیل داده های اقتصادی+ اجتماعی نماد شفن به همراه تعیین نقطه ورود

❖ تحلیل داده های شخارک:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۵۲.۸ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود به خوبی پیش‌بینی شده‌اند. و خط میانگین تعداد بازدیدهای وزن دار شده افزایش‌ها را تقریبا به خوبی نشان داده است.

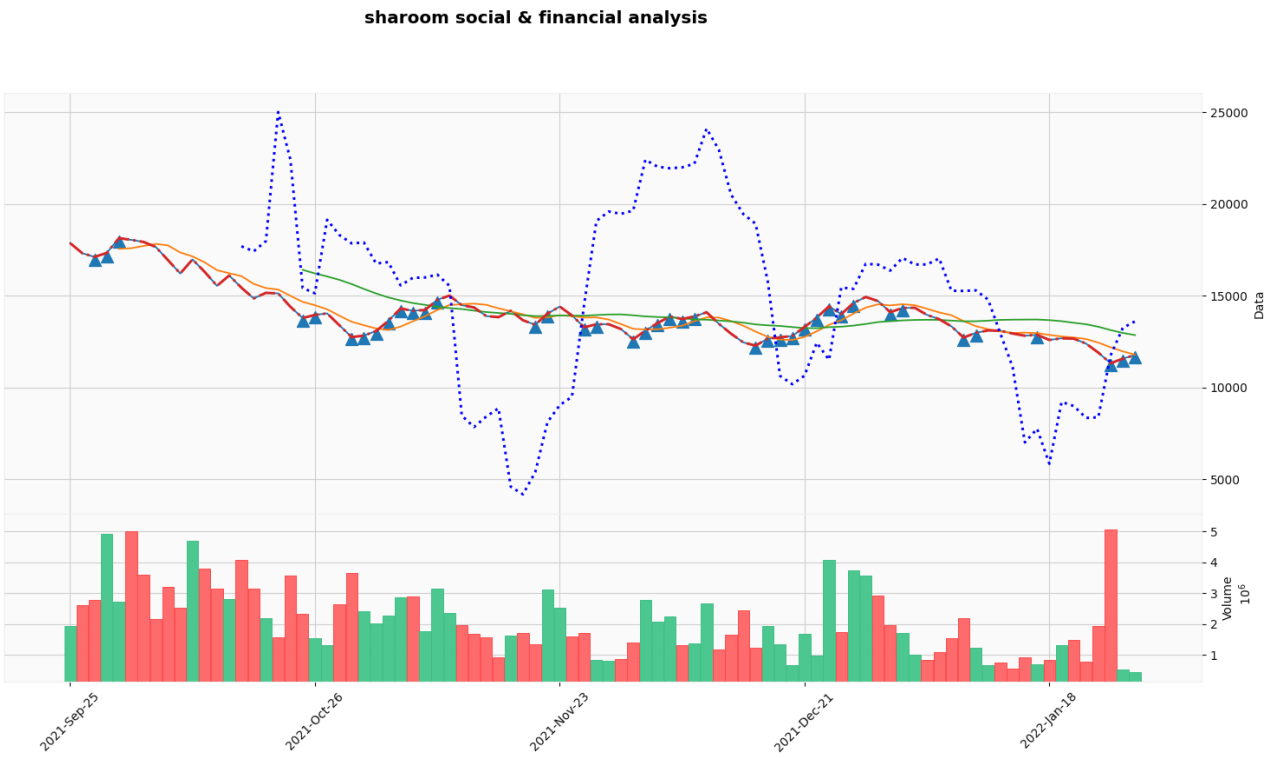


شکل۲۲: تحلیل داده های اقتصادی+ اجتماعی نماد شخارک به همراه تعیین نقطه ورود

❖ تحلیل داده های شاروم:

در این نماد، یادگیری با استفاده از داده های ترکیبی(اقتصادی+اجتماعی)، ۷۶.۱۵ درصد موفقیت داشت .

همانطور که می‌بینیم، نقطه های ورود با درصد پیشبینی بالایی درست پیدا شده‌اند. همینطور میانگین تعداد بازدیدها تقریبا سیر صعودی یا نزولی بودن نماد را خوب پیشبینی کرده است.



شکل۲۳: تحلیل داده های اقتصادی+ اجتماعی نماد شاروم به همراه تعیین نقطه ورود

پیوست: برگه‌های سامانه هشتگ:

پارس :

شامل همه ی کیلدواژه های : پارس

حداقل یکی از واژه های : #پارس ، بازار سرمایه ، بازارسرمایه ، سود ، ضرر ، پتروشیمی پارس ، نماد ، بورس

فاقد واژه های : خودرو، باستان ، کورش ، کوروش

پترول :

شامل همه ی کیلدواژه های : پترول

حداقل یکی از واژه های :

فاقد واژه های : لیترتیل ، افغان

شاراک :

شامل همه ی کیلدواژه های : شاراک

حداقل یکی از واژه های :

فاقد واژه های :

شغدیر :

شامل همه ی کیلدواژه های : شغدیر

حداقل یکی از واژه های :

فاقد واژه های :

وپترو :

شامل همه ی کیلدواژه های : وپترو

حداقل یکی از واژه های :

فاقد واژه های :

بوعلی :

شامل همه ی کیلدواژه های : بوعلی

حداقل یکی از واژه های : سهم ، سهام ، بورس ، فرا بورس ، فرابورس ، سیگنال خرید ، سیگنال فروش ، بازار سرمایه ، بازارسرمایه ، نماد ، سهمها

فاقد واژه های : شیخ ، وبوعلی ، طب ، ارامگاه ، گردشگری ، همدان ، روح ، جسم ، کتیبه ، بیمار ، بیماری ، پزشک ، پزشکی ، سرطان ، دیابت ، اختراع ، سلامت ، سلامتی ، سرمایه گذاری بوعلی ، ادبیات ، دانشمند ، دانشمندان ، مثلث ، الناس ، مشهد ، خادم ، نبش بوعلی ، میدان بوعلی ، حاج قاسم ، انجمن های علمی ، دانشگاه بوعلی ، دانشگاه های ، خیابان بوعلی ، ابن سینا ، حنانه ، بوس هنری ، عشاق ، موزیک ، جشن هزاره ، استکبار ، دانشگاه بوعلی ، دانشگاه بوعلی سینا ، انسانی ، خیام ، فرهنگی ، عطاری ، فلاسفه ، بیمارستان ، بازنشستگی

پارسان :

شامل همه ی کیلدواژه های : پارسان

حداقل یکی از واژه های : سهم ، سهام ، سهامها ، سیگنال خرید، سیگنال فروش ، فرابورس ، بورس ، نماد ، بازار سرمایه ، بازارسرمایه ، شاخص کل

فاقد واژه های :

شاروم :

شامل همه ی کیلدواژه های : شاروم

حداقل یکی از واژه های : سهم ، سهام ، بازار سرمایه ، بازارسرمایه ، بورس ، فرابورس ، سیگنال ، سهمها ، سهامها ، سهمهای ، پتروشیمی

فاقد واژه های : ترنس ، اهنگ ، سرود ، ملوبات ، جواهری ، دوجنسه ، خواننده ، LGBT ، موسیقی ، سریال ، فیلم ، شاروم شاهرخی ، MUSIC ، ترنسکشوال

شپدیس :

شامل همه ی کیلدواژه های : شپدیس

حداقل یکی از واژه های :

فاقد واژه های :

تاپیکو :

شامل همه ی کیلدواژه های : تاپیکو

حداقل یکی از واژه های :

فاقد واژه های :