



Amirkabir University of Technology
(Tehran Polytechnic)

Department of information technology

Tracking social media events related to
Iranian stock market

By
Mehrzaad Ahmadian

Professor
Meisam Nazariani

June 2022

چکیده

تحلیل احساسات افراد به عنوان بازیگران بازار بزرگ بورس اوراق بهادار و کشف نظرات آن‌ها در مورد افزایش و کاهش قیمت سهام و ارزش شرکت‌ها، کمک شایانی به معامله‌گران در این بازار می‌نماید. استفاده از ابزارهای بروز و مدرن یادگیری ماشین از جمله کارسازترین روش‌های حل این مسئله می‌باشد. در این پژوهش، به بررسی روش‌ها و چالش‌های استخراج پیام‌های افراد و ذخیره آن‌ها پرداخته و روش‌هایی جهت ایجاد مدل‌هایی به منظور کشف نمادها، تحلیل احساسات و همچنین تحلیل سیگنال‌های خرید و فروش در پیام‌ها معرفی می‌نماییم. این روش‌ها با استفاده از دو دیتاست نظرات وبسایت فروش کتاب طاقچه و همچنین دیتاستی که به صورت محدود توسط ما تولید شده است، مورد ارزیابی قرار گرفته و نتایج آن گزارش شده است. مدل‌های ایجاد شده در این پژوهش، در گستره مدل‌های کلاسیک مبتنی بر بیز و همچنین مدل‌های مدرن برت مبتنی بر شبکه مبذل می‌باشد.

کلمات کلیدی: تحلیل احساسات، کشف سیگنال، بورس اوراق بهادار، یادگیری ماشین، یادگیری عمیق.

فهرست مطالب

آ.....	چکیده
1.....	فصل اول مقدمه
2.....	1-1- پیش گفتار
2.....	2-1- کاربردها
2.....	3-1- تعریف مسئله
3.....	فصل دوم مرور ادبیات
4.....	مرور ادبیات
4.....	1-2- شبکه‌های عصبی
5.....	2-2- شبکه‌های عصبی عمیق
5.....	3-2- پیوند آموزش و یادگیری انتقالی
6.....	4-2- شبکه مبدل
8.....	5-2- شبکه برت
9.....	فصل سوم استخراج داده
10.....	استخراج داده
10.....	1-3- عضویت در تلگرام و دریافت api_id
10.....	1-1-3- عضویت در تلگرام
10.....	2-1-3- ورود به پنل توسعه تلگرام
12.....	3-1-3- ساخت اپلیکیشن جدید
12.....	4-1-3- چالش‌ها
13.....	2-3- اتصال به رابط برنامه‌نویسی تلگرام
14.....	1-2-3- چالش‌ها
15.....	3-3- استخراج و بررسی ویژگی‌های پیام‌های تلگرام
16.....	1-3-3- ویژگی‌های موجود در پیام‌های استخراجی از تلگرام
17.....	4-3- ذخیره داده‌ها در پایگاه داده
17.....	1-4-3- ایجاد پایگاه داده
19.....	2-4-3- اتصال به mysql از طریق پایتون
19.....	3-4-3- ایجاد جدول

21	3-4-4- ذخیره پیام‌ها.....
21	3-5- تولید داشبورد هوش تجاری.....
24	فصل چهارم مدل سازی.....
25	مدل سازی.....
25	4-1- استخراج نماد از متن پیام‌ها.....
25	4-2- تحلیل احساسات موجود در متن.....
25	4-2-1- آموزش مدل با استخراج ویژگی خودکار.....
26	4-2-2- آموزش مدل با استخراج ویژگی به صورت دستی.....
27	4-3- تشخیص سیگنال.....
27	4-3-1- آموزش مدل با استخراج ویژگی خودکار.....
27	4-3-2- آموزش مدل با استخراج ویژگی به صورت دستی.....
28	فصل پنجم بررسی نتایج و تحلیل آن‌ها.....
29	بررسی نتایج و تحلیل آن‌ها.....
29	5-1- تحلیل احساسات.....
29	5-1-1- آموزش مدل با استخراج ویژگی خودکار.....
31	5-1-2- آموزش مدل با استخراج ویژگی دستی.....
34	5-2- تشخیص سیگنال.....
34	5-2-1- آموزش مدل با استخراج ویژگی خودکار.....
35	5-2-2- آموزش مدل با استخراج ویژگی دستی.....
38	فصل ششم جمع‌بندی و نتیجه‌گیری و پیشنهادات.....
39	جمع‌بندی و نتیجه‌گیری.....
39	پیشنهادهات.....

فهرست اشکال

- شکل 1- ترتیب علوم هوش مصنوعی.....4
- شکل 2- شمای عملیاتی ریاضی داخل یک نورون - نمونه‌ای از یک شبکه عصبی چندلایه.....5
- شکل 3- نمای کلی ساختار یک شبکه مبدا.....7
- شکل 4- صفحه ورود به پنل توسعه تلگرام.....11
- شکل 5- پیام ارسالی حاوی کد تایید برای ورود به پنل توسعه تلگرام.....11
- شکل 6- صفحه ورود به پنل توسعه تلگرام در حال انتظار برای ورود به کد تایید.....11
- شکل 7- صفحه ساخت اپلیکیشن جدید در پنل توسعه تلگرام.....12
- شکل 8- صفحه اطلاعات اپلیکیشن ساخته شده در پنل توسعه تلگرام.....12
- شکل 9- تنظیمات مورد نیاز جهت اتصال به تلگرام از طریق telethon.....13
- شکل 10- قطعه کد لازم برای اتصال به تلگرام از طریق telethon.....14
- شکل 11- استفاده از ابزار copyTables جهت انتخاب آسان ستون نماد در مروگر فایرفاکس.....16
- شکل 12- صفحه ایجاد پایگاه داده جدید در ابزار phpMyadmin.....18
- شکل 13- قطعه کد لازم برای اتصال به mysql در پایتون.....19
- شکل 14- کوئری استفاده شده برای ایجاد جدول messages.....20
- شکل 15- کوئری استفاده شده برای ایجاد پیام جدید در جدول messages.....21
- شکل 16- صفحه اول از داشبورد هوش تجاری تولید شده.....22
- شکل 17- صفحه دوم از داشبورد هوش تجاری تولید شده.....22
- شکل 18- صفحه سوم از داشبورد هوش تجاری تولید شده.....23
- شکل 19- نمودار خطای آموزش و ارزیابی مدل.....27
- شکل 20- نمودار معیارهای ارزیابی مدل بر روی داده‌های تست.....28
- شکل 21- نمودار خطای آموزش و ارزیابی مدل.....28
- شکل 22- نمودار معیارهای ارزیابی مدل بر روی داده‌های تست تولید شده توسط خودمان.....29
- شکل 23- ماتریس آشفتگی مدل.....30
- شکل 24- ماتریس آشفتگی مدل با ضرایب ۱ برای کلمات.....31
- شکل 25- ماتریس آشفتگی مدل با ضرایب ۱۰ برای کلمات.....32
- شکل 26- ماتریس آشفتگی مدل با ضرایب ۲۰ برای کلمات.....32
- شکل 27- نمودار خطای آموزش و ارزیابی مدل.....32

- شکل 28- نمودار معیارهای ارزیابی مدل بر روی داده‌های تست تولید شده..... 33
- شکل 29- ماتریس آشفتگی مدل با ضرایب ۱ برای کلمات..... 34
- شکل 30- ماتریس آشفتگی مدل با ضرایب ۱۰ برای کلمات..... 35
- شکل 31- ماتریس آشفتگی مدل با ضرایب ۲۰ برای کلمات..... 35

فهرست جداول

- جدول 1- لیست ویژگی‌های موجود در پیام‌های استخراجی تلگرام.....16
- جدول 2- مشخصات ستون‌های جدول messages.....20
- جدول 3- مقدار معیار ارزیابی f_1 مدل در دوره‌های آموزشی و دیتاست‌های گوناگون.....28
- جدول 4- مقدار معیار ارزیابی f_1 مدل در دوره‌های آموزشی و دیتاست‌های گوناگون.....29
- جدول 5- نتایج ارزیابی مدل.....30
- جدول 6- نمونه‌ای از کلمات مثبت و منفی تاثیرگذار.....30
- جدول 7- نتایج ارزیابی مدل.....31
- جدول 8- مقدار معیار ارزیابی f_1 مدل در دوره‌های آموزشی گوناگون.....33
- جدول 9- نمونه‌ای از کلمات مثبت و منفی تاثیرگذار.....34
- جدول 10- نتایج ارزیابی مدل.....34

فصل اول

مقدمه

1-1- پیش‌گفتار

طی چند سال گذشته پیشرفت علوم مرتبط با هوش مصنوعی امکانات و تغییرات بسیار زیادی را در زندگی بشر به وجود آورده است. از جمله مهم‌ترین و پرتعدادترین زمینه‌های نمود هوش مصنوعی، حوزه یادگیری ماشین و شبکه‌های عصبی می‌باشد. یکی از وظایف کاربردی ایجاد شده توسط پژوهشگران در چند سال اخیر پایش رویدادهای فضای مجازی برای پاسخ‌گویی به نیازهای اطلاعاتی و تحلیل اطلاعات آن بوده است.

1-2- کاربردها

از جمله بارزترین نمونه‌های استفاده از این حوزه در صنعت، پایش اطلاعات و رویدادهای فضای مجازی در حوزه بورسی می‌باشد. بسیاری از سازمان‌ها از جمله سازمان بورس، صندوق‌های سرمایه‌گذاری و یا اشخاص حقیقی و حقوقی به منظور آنالیز، زیر نظر گرفتن، تحلیل بهتر بازار و ... به سمت استفاده از سیستم‌های پایشگر هوشمند فضای مجازی رفته‌اند.

1-3- تعریف مسئله

در این پروژه قرار است ابتدا برای استخراج پیام‌های مرتبط با بورس از چند کانال مطرح تلگرامی¹ اقدام شود و بانکی از پیام‌های بورسی تشکیل دهیم. سپس با استفاده از مباحث مرتبط با یادگیری برای سه هدف ذیل تلاش خواهیم کرد:

- شناسایی نمادهای بورسی در پیام‌ها
- تحلیل احساسی پیام‌ها و دسته‌بندی مثبت یا منفی بودن آن‌ها از منظر احساسی
- تحلیل پیام‌ها و دسته‌بندی آن‌ها از منظر سیگنال خرید یا فروش بودن

¹ Telegram.org

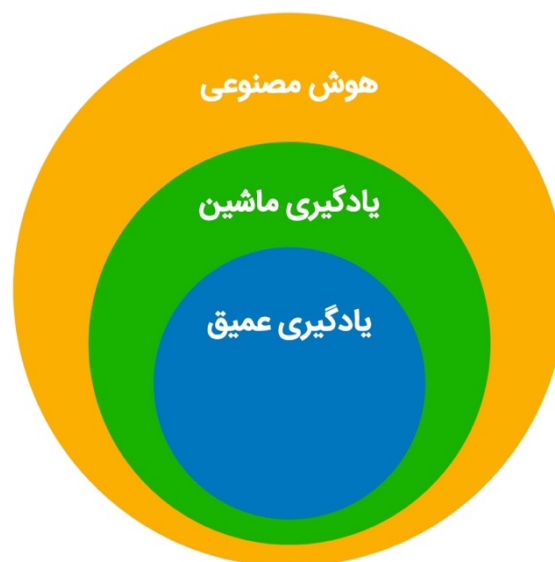
فصل دوم مرور ادبیات

مرور ادبیات

در این فصل به بررسی به بررسی مفاهیم اولیه و بررسی برخی از مدل‌ها و ابزارهای مورد استفاده در این پژوهش می‌پردازیم.

2-1- شبکه‌های عصبی

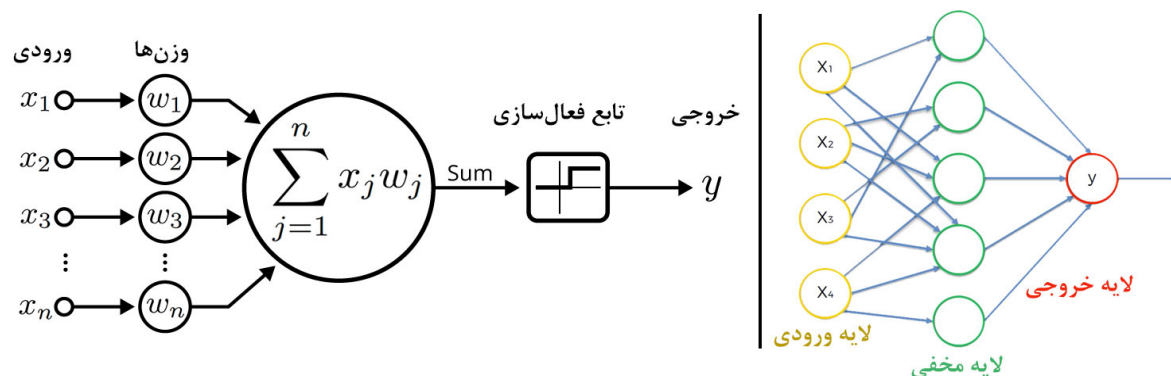
از جمله روش‌های یادگیری ماشین استفاده از شبکه‌های عصبی مصنوعی و یادگیری عمیق است که می‌توان با استفاده از آن مسائل پویایی که با روش‌های برنامه‌نویسی معمول در رایانه قابل انجام نیست را با الگوبرداری از روش پردازش اطلاعات در مغز انسان حل نمود. یادگیری عمیق از زیرمجموعه‌های یادگیری ماشین می‌باشد که خود از علوم زیرمجموعه هوش مصنوعی است.



شکل ۱. ترتیب علوم هوش مصنوعی.

به صورت کلی یک شبکه عصبی از مجموعه‌ای از نورون‌ها تشکیل شده و چیدمان این نورون‌ها در کنار یکدیگر یک لایه از شبکه را تشکیل می‌دهند. هر شبکه عصبی از چندین لایه تشکیل شده است که به لایه‌های قبل و بعد خود متصل است. شبکه عصبی می‌تواند با دریافت اطلاعات در لایه اول که لایه ورودی نامیده می‌شود و گذر آن از لایه‌های میانی (لایه‌های مخفی)، در لایه خروجی به نتیجه دلخواه برسد و مسائل پویا که نیاز به یادگیری دارند را حل نماید. هر نورون شامل وزن‌هایی است که به صورت جداگانه در مقادیر ورودی نظیر آن ضرب می‌شود. سپس مقادیر حاصل با یکدیگر و پس از آن با مقداری که بایاس نامیده می‌شود جمع بسته می‌شوند و به عنوان ورودی به تابعی که از آن با عنوان تابع فعال‌سازی یاد

می‌شود داده می‌شوند. این تابع در خروجی خود مقادیر ورودی را به فضایی مشخص برده و به نورون‌های متصل بعدی خود ارسال می‌نماید. در یک شبکه عصبی هر نورون و هر لایه مسئول انجام عملیات و تشخیص مفهومی خاص در راستای حل مسئله کلی است. اینکه هر نورون در شبکه نسبت به چه ترکیب مقادیری از داده‌ها حساس باشد و دریافت سایر مفاهیم را به نورون‌های دیگر واگذار کند بستگی به وزن‌های ورودی و مقدار بایاس آن دارد. یک شبکه عصبی در فرآیند آموزش خود از طریق تنظیم این مقادیر حل مسئله‌ی تعریف شده را می‌آموزد.



شکل ۲. (چپ) شمای عملیات ریاضی داخل یک نورون (راست) نمونه‌ای از یک شبکه عصبی چندلایه.

2-2- شبکه‌های عصبی عمیق

هر نورون در شبکه عصبی در واقع یک تابع است که با دریافت داده‌های ورودی و اعمال عملیات ریاضی مربوط به خود خروجی خاصی تولید می‌کند. هرچه شبکه عصبی دارای تعداد لایه‌های مخفی بیشتری باشد، به دلیل بالاتر رفتن مرتبه آن قادر است عملیات ریاضی پیچیده‌تری را انجام دهد. از این‌گونه شبکه‌های عصبی با تعداد لایه‌ها و نورون‌های بیشتر با عنوان شبکه‌های عصبی عمیق یاد می‌شود. همچنین برای حل بسیاری از مسائل پیچیده‌تر از ترکیب چندین شبکه عصبی که هر کدام توانایی انجام عملیات و تشخیص مفاهیم مخصوص به خود را دارند استفاده می‌شود. شبکه‌های عصبی عمیق به دلیل برخورداری از تعداد زیادی پارامترهای قابل آموزش نیازمند آموزش با تعداد زیادی داده برچسب خورده هستند که این امر با ظهور سخت‌افزارهای پیشرفته میسر شده است.

2-3- پیش‌آموزش و یادگیری انتقالی

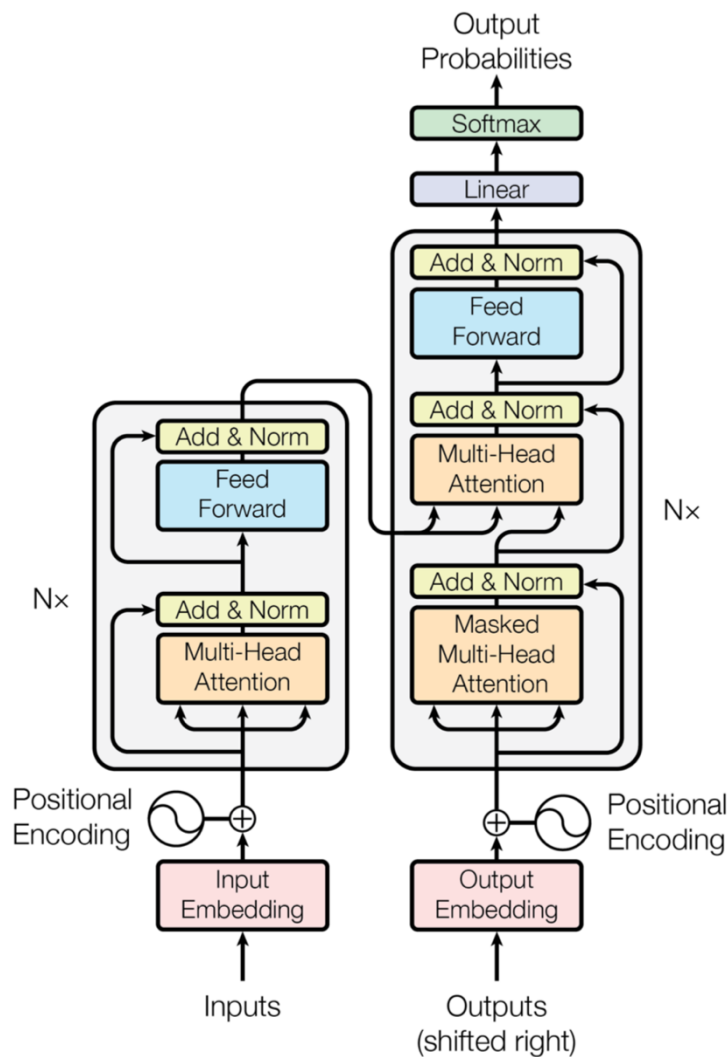
شبکه‌های عصبی عمیق به دلیل خودکار بودن پروسه مهندسی ویژگی‌ها در آنان به تعداد زیادی داده جهت آموزش با کیفیت نیازمندند. در بسیاری از موارد تعداد دادگان محدود است و نمی‌توان از طریق آموزش شبکه بر روی تعداد محدود داده‌ها به دقت بالایی رسید. از جمله راه‌کارهای مورد استفاده پژوهشگران در

چنین مواردی این است که ابتدا کل شبکه یا بخش‌هایی از آن توسط داده‌های جمع‌آوری شده برای مسئله نزدیکی با مسئله اصلی آموزش دیده و سپس شبکه پیش‌آموزش دیده با آموزش بر روی داده‌های اصلی بهینه‌سازی می‌شود.

دانشمندان در دو دهه گذشته اقدام به آموزش شبکه‌هایی با اهداف کلی جهت استفاده در مسائل دیگر بدون اینکه نیاز به آموزش مجدد آن‌ها باشد نموده‌اند. برای مثال انواعی از شبکه‌های عصبی وجود دارد که از آن‌ها جهت استخراج ویژگی‌های تصاویر استفاده می‌شود. این شبکه‌ها بر روی داده‌های زیادی آموزش دیده و عموماً به دقت بالایی دست می‌یابند. به پروسه آموزش یک شبکه عصبی و انتقال آن به ساختار شبکه دیگر یادگیری انتقالی گفته می‌شود. عموماً در زمان آموزش شبکه اصلی وزن‌های شبکه منتقل شده به حالت ثابت در آمده و آموزش نمی‌بینند.

2-4- شبکه مبدل

در سال ۲۰۱۷ محققین با اشاره به این نکته که بهترین شبکه‌های پردازش داده‌ها به صورت گام به گام با استفاده از اتصال یک شبکه کدگذار به شبکه کدگشای بازگردنده که در آن از مکانیزم توجه استفاده شده است، شبکه جدیدی را معرفی نمودند که در آن شبکه بازگردنده حذف شده و تنها به نوع خاصی از مکانیزم توجه اکتفا شده است. این شبکه دارای توانایی پردازش داده‌های دارای توالی زمانی به صورت یکجا هستند. برای مثال در ترجمه ماشینی می‌توان به جای پردازش گام به گام ورودی به صورت کلمه به کلمه، کل جمله را به صورت یکجا وارد شبکه نمود. در این مقاله عنوان شده است که این شبکه‌ها علاوه بر کیفیت و دقت بالاتر در نتایج، به دلیل پردازش داده‌ها در یک گام نیاز به زمان کمتری جهت آموزش نیز دارند.



شکل ۳. نمای کلی ساختار یک شبکه مبدل.

شبکه مبدل دارای دو بخش کدگذار و کدگشا بوده و مانند سایر شبکه‌هایی که از این معماری بهره می‌برند ابتدا ورودی در بخش کدگذار پردازش شده و به داده‌های مورد نیاز جهت تولید خروجی در شبکه کدگشا تبدیل می‌شوند. این شبکه مشکلاتی که شبکه‌های بازگردنده با آن روبرو هستند از جمله مسئله گرادیان محو شونده و انفجاری و همچنین مشکل فراموشی داده‌های گام‌های دور در گذشته را برطرف ساخته است. شبکه مبدل به گونه‌ای طراحی شده است که می‌توان بخش کدگذار و کدگشای آن را به صورت پشته روی یکدیگر چید و از مزایای عمیق‌تر شدن شبکه در حل مسائل پیچیده‌تر نیز بهره جست. در شکل زیر مشاهده می‌شود که با قرار دادن چند شبکه کدگذار تبدیل کننده در امتداد یکدیگر، اطلاعات ورودی توسط لایه اول پردازش شده و به ترتیب به عنوان ورودی به شبکه لایه‌های بعدی ارسال می‌شود. سپس اطلاعات

پردازش شده توسط آخرین لایه کدگذار به عنوان ورودی به تک تک شبکه‌های کدگشا که پشت سر یکدیگر چیده شده‌اند داده می‌شود.

2-5- شبکه برت

در سال ۲۰۱۸ پژوهشگران شرکت گوگل^۱ مدلی با نام برت^۲ را معرفی نمودند. این مدل از پشته‌ای از شبکه‌های کدگذار مبدل (در مقاله اصلی ۱۲ عدد) ساخته شده است. این مدل با هدف فراهم آوردن نمایه‌های زبانی غنی‌تر نسبت به گذشته طراحی شده است. برت برخلاف شبکه‌های مبدل که به صورت خودگردان کلمات را تولید می‌کند، به صورت خودکدگذار^۳ طراحی شده است. به این معنی که برخلاف شبکه‌های مبدل پایه که در هنگام تولید کلمات در زمان آموزش فقط امکان استفاده از اطلاعات گام‌های قبلی را دارند، این امکان را دارد که به صورت دو طرفه به اطلاعات دسترسی داشته باشد و بتواند در زمان حدس یک کلمه علاوه بر اطلاعات گذشته، اطلاعات آینده را هم در نظر بگیرد. همچنین شبکه برت با هدف بکارگیری در مسائل و شبکه‌های دیگر طراحی شده است و پس از پیش‌آموزش قابلیت بکارگیری در ساختار شبکه‌های دیگر را دارد.

¹ Google

² BERT (Bidirectional Encoder Representations from Transformers)

³ AutoEncoding

فصل سوم استخراج داده

استخراج داده

استخراج داده و تشکیل دیتاست از جمله اولین و مهم‌ترین مراحل کار در این پروژه است. منابع مختلفی برای دستیابی به داده‌ها و پیام‌های بورسی وجود دارد. یکی از این منابع کانال‌های خبری و اطلاع‌رسانی در نرم‌افزارهای پیام‌رسان هستند. پیام‌رسان تلگرام به عنوان یکی از پرمخاطب‌ترین پیام‌رسان‌های حال حاضر در کشور بوده و در این پروژه به عنوان منبع استخراج داده از آن استفاده خواهد شد.

روند کلی کار به این خواهد بود که تعداد صد پیام آخر از چند کانال مطرح بورسی را در تلگرام استخراج کرده و سپس در جدولی از نوع پایگاه داده ¹mysql ذخیره خواهیم کرد. مراحل کار با جزئیات بیشتر در ادامه کار شرح داده شده است.

3-1- عضویت در تلگرام و دریافت api_id

جهت اتصال به رابط برنامه‌نویسی تلگرام بایستی پس از عضویت در آن، اقدام به ساخت اپلیکیشن تلگرام و دریافت api_id نماییم. مراحل کار جهت ثبت نام و دریافت api_id در ادامه آورده خواهد شد.

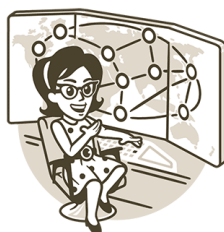
3-1-1- عضویت در تلگرام:

در اولین مرحله می‌بایست پس از نصب اپلیکیشن تلگرام بر روی تلفن همراه خود، اقدام به عضویت در آن نمایید.

3-1-2- ورود به پنل توسعه تلگرام

در مرحله بعد می‌بایست از طریق مرورگر به آدرس <https://my.telegram.org/apps> مراجعه و با استفاده از شماره تلفن همراه استفاده شده در مرحله قبل وارد پنل توسعه پیام‌رسان تلگرام شویم. پس از ورود شماره همراه، پیامی حاوی کد تایید ورودی به آن شماره در اپلیکیشن تلگرام ارسال خواهد شد که به منظور ورود می‌بایست در سایت وارد شود.

¹ Mysql.org



Delete Account or Manage Apps

Log in here to **manage your apps** using Telegram API or **delete your account**. Enter your number and we will send you a confirmation code via Telegram (not SMS).

Your Phone Number

+12223334455

Please enter your number in [international format](#)

Next

شکل ۴. صفحه ورود به پنل توسعه تلگرام.

Web login code. Dear Mehrzad, we received a request from your account to log in on my.telegram.org. This is your login code:

[Redacted]

Do **not** give this code to anyone, even if they say they're from Telegram! **This code can be used to delete your Telegram account.** We never ask to send it anywhere.

If you didn't request this code by trying to log in on my.telegram.org, simply ignore this message.

21:56

شکل ۵. پیام ارسالی حاوی کد تایید برای ورود به پنل توسعه تلگرام.



Delete Account or Manage Apps

Log in here to **manage your apps** using Telegram API or **delete your account**. Enter your number and we will send you a confirmation code via Telegram (not SMS).

Your Phone Number

+989303301830 (Incorrect?)

Confirmation code

Confirmation code

☐ Remember Me

Sign In

شکل ۶. صفحه ورود به پنل توسعه تلگرام در حال انتظار برای ورود کد تایید.

3-1-3 ساخت اپلیکیشن جدید

پس از ورود به پنل توسعه تلگرام می‌بایست اقدام به ساخت اپلیکیشن جدید نمود.

Create new application

App title:

Short name:
alphanumeric, 5-32 characters

URL:


Platform: ☒ Android
☐ iOS
☐ Windows Phone
☐ BlackBerry
☐ Desktop
☐ Web
☐ Ubuntu phone
☐ Other (specify in description)


Description:

شکل ۷. صفحه ساخت اپلیکیشن جدید در پنل توسعه تلگرام.

پس از ساخت اپلیکیشن جدید، اطلاعاتی نظیر `api_id` و `api_hash` که جهت اتصال به رابط برنامه‌نویسی تلگرام مورد نیاز هستند، در اختیار ما قرار داده خواهد شد.

App configuration

App api_id: 

App api_hash: 

App title:

Short name:
alphanumeric, 5-32 characters

شکل ۸. صفحه اطلاعات اپلیکیشن ساخته شده در پنل توسعه تلگرام.

4-1-3 چالش‌ها

در انجام مراحل بالا با چالش‌های ذیل روبه‌رو شدیم:

• مسدود بودن تلگرام

در حال حاضر بستر تلگرام در کشور مسدود می‌باشد و برای اتصال به آن می‌بایست از VPN استفاده نمود.

• خطا به هنگام ساخت اپلیکیشن جدید در پنل توسعه تلگرام

پس از ورود موفق به پنل توسعه تلگرام در مرحله ساخت اپلیکیشن جدید با خطا نامعلومی مواجه شدیم. پس از بررسی‌های صورت گرفته مشخص شد که علت خطا استفاده از VPN عمومی بود که با آن وارد پنل توسعه تلگرام شده بودیم و امکان ساخت اپلیکیشن از طریق آدرس IP آن فراهم نبود.

راه حل: این مشکل پس از استفاده از VPN خصوصی رفع گردید.

3-2- اتصال به رابط برنامه‌نویسی تلگرام

به منظور استفاده از رابط برنامه‌نویسی^۱ تلگرام و استخراج پیام‌ها می‌توان از کتابخانه‌های^۲ آماده موجود در زبان پایتون استفاده کرد. یکی از معروف‌ترین این کتابخانه‌ها Telethon^۳ می‌باشد که برای اتصال به رابط برنامه‌نویسی تلگرام از آن استفاده کردیم.

به منظور نصب این کتابخانه در پایتون می‌توان از قطعه کد زیر استفاده نمود:

pip install telethon

پس از نصب telethon می‌توان با تنظیم چهار مورد شامل api_id، api_hash، شماره تماس و نام کاربری به رابط برنامه‌نویسی تلگرام متصل شد.

```
api_id = 15916079
api_hash = 'cc474a2ff96f96f3db2276c896bf6bd7'
phone = '+989303301830'
username = 'Mehrzadahmadian'
```

شکل ۹. تنظیمات مورد نیاز جهت اتصال به تلگرام از طریق telethon.

^۱ Api

^۲ Library

^۳ <https://docs.telethon.dev/en/stable/>

برای آموزش و کسب اطلاعات بیشتر در مورد نحوه کار با کتابخانه telethon می‌توان به وبسایت آن به آدرس ذیل مراجعه نمود:

[/https://docs.telethon.dev/en/stable](https://docs.telethon.dev/en/stable)

جهت اتصال به تلگرام از طریق telethon از قطعه کد ذیل استفاده شده است.

```
client = TelegramClient(username, api_id, api_hash)
client.start()
print("Telegram Client Created")

if not client.is_user_authorized():
    client.send_code_request(phone)
    try:
        client.sign_in(phone, input('Enter the code: '))
    except SessionPasswordNeededError:
        client.sign_in(password=input('Password: '))
```

شکل ۱۰. قطعه کد لازم برای اتصال به تلگرام از طریق telethon.

زمانی که برای اولین بار قطعه کد بالا را اجرا می‌کنید، گذرواژه‌ای از طرف تلگرام برای شما ارسال خواهد شد که می‌بایست آن را در محیط کنسول وارد نمایید. پس از ورود موفق به تلگرام فایلی با پسوند session. در محل اجرای برنامه ایجاد خواهد شد که حاوی اطلاعات ورود شما به تلگرام خواهد بود. در صورتی که این فایل حذف گردد، بهنگام اتصال به تلگرام می‌بایست مجدد گذواژه ورودی را دریافت و وارد نمایید.

3-2-1- چالش‌ها

• دریافت پیغام خطا بهنگام استفاده مکرر از telethon

پس از چندین مرتبه استفاده از کتابخانه telethon برای دریافت پیام‌های بورسی با پیغامی مواجه شدیم که در آن ذکر شده بود:

سقف استفاده رایگان شما از telethon پر شده است. لطفا جهت استفاده مجدد ۴۸ ساعت صبر نمایید.

راه حل: برای حل این مشکل ناچار شدیم تا موقتا از شماره همراه دیگری که سقف استفاده آن پر نشده بود اقدام به اتصال کنیم.

3-3- استخراج و بررسی ویژگی‌های پیام‌های تلگرام

پیام‌های استفاده شده در این پروژه از ده کانال مطرح بورسی به آدرس‌های زیر استخراج شده‌اند:

- https://t.me/Codal360_ir
- <https://t.me/asiasarmayeh>
- <https://t.me/NoavaranAmin>
- <https://t.me/snipersahamyab>
- <https://t.me/ChanelVIP20>
- <https://t.me/groupeprezabourse>
- https://t.me/signalle_bartare_bours
- <https://t.me/sahamyab>
- <https://t.me/topsignalexchange>
- <https://t.me/vipmosbat5>

به منظور جستجو نمادها در پیام‌های بورسی اقدام به جستجو لیست نمادها در متن هر پیام کردیم. برای دریافت لیست کل نمادها به صفحه لیست نمادها در وبسایت شرکت مدیریت فناوری بورس تهران^۱ مراجعه کردیم. آدرس این صفحه به شرح ذیل می‌باشد:

<http://www.tsetmc.com/Loader.aspx?ParTree=111C1417>

پس از مراجعه به این صفحه با جدولی حاوی اطلاعات تمامی نمادهای بورسی مواجهه خواهید شد. به منظور دریافت آسان‌تر ستون نمادها از این جدول می‌توانید از ابزارهایی نظیر CopyTables در مرورگرهای فایرفاکس یا کروم استفاده نمایید.

¹ tsetmc.com

--	--

-	legacy
-	edit_hide
وضعیت پین بودن پیام	pinned
-	from_id
شناسه پیامی که این پیام از آن فوروارد شده است	fwd_from
شناسه ربات ارسال کننده پیام	via_bot_id
شناسه پیامی که این پیام در پاسخ به آن داده شده است	reply_to
حاوی اطلاعات چندرسانه‌ای پیام	media
-	reply_markup
-	entities
تعداد بازدیدهای پیام	views
تعداد فورواردهای پیام	forwards
تعداد پاسخ‌هایی که به این پیام داده شده است	replies
تاریخ ویرایش پیام	edit_date
-	post_author
-	grouped_id
-	restriction_reason
-	ttl_period

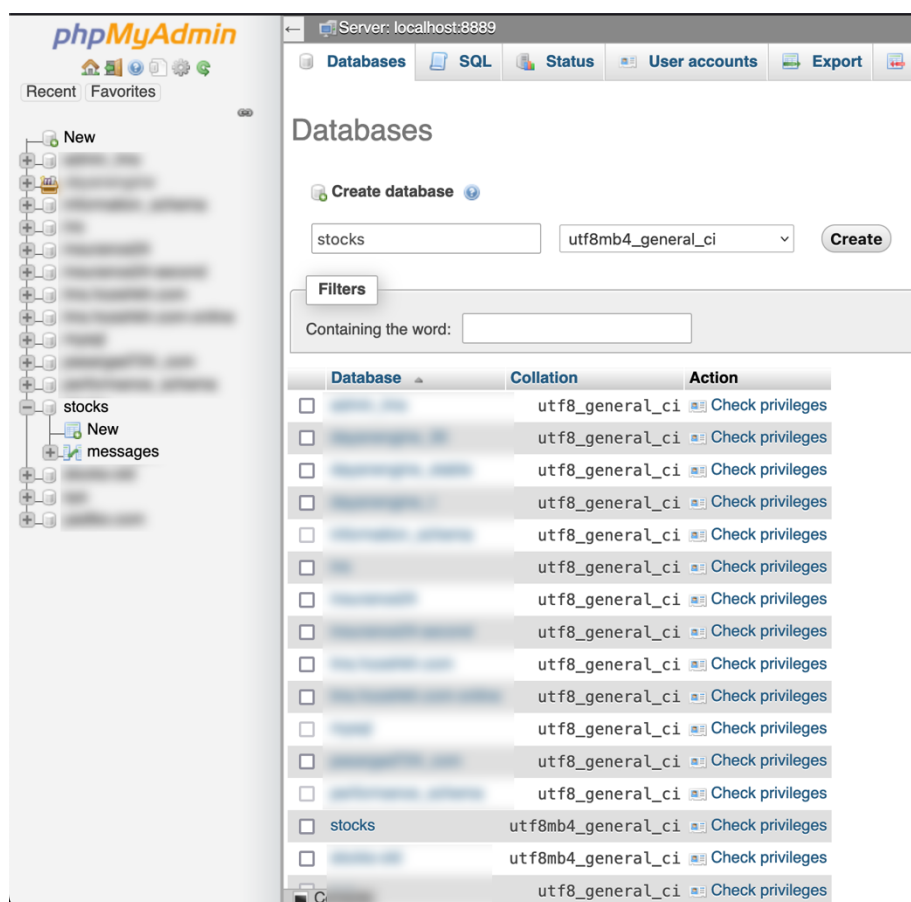
3-4- ذخیره داده‌ها در پایگاه داده

در این پروژه برای ذخیره پیام‌های استخراج شده و ویژگی‌های آن‌ها از پایگاه داده mysql استفاده شده است که بخش‌های کار در قسمت‌های ذیل توضیح داده خواهد شد.

3-4-1 ایجاد پایگاه داده

به منظور ساخت پایگاه داده ابتدا نرم افزار phpMyadmin را نصب و آن را از طریق مرورگر خود باز

می‌کنیم. برای دسترسی به phpMyadmin در سیستم عامل macOS می‌توان به آدرس `http://localhost:8888/phpmyadmin` در مرورگر مراجعه کرد. پس از ورود به صفحه اصلی phpMyadmin، به پنجره Databases رفته و دیتابیس مورد نظر خود را تحت عنوان 'stocks' ایجاد می‌کنیم. همچنین برای پشتیبانی بهتر از نمادها و کاراکترهای خاص مقدار collation را بر روی حالت `utf8mb4_general_ci` قرار دادیم.



شکل ۱۲. صفحه ایجاد پایگاه داده جدید در ابزار phpMyadmin.

¹ phpmyadmin.net

3-4-2- اتصال به mysql از طریق پایتون

برای اتصال به پایگاه داده mysql از کتابخانه Mysql connector استفاده شده است. برای نصب این کتابخانه می توان از قطعه کد زیر استفاده کرد:

```
pip install mysql-connector-python
```

پس از نصب این کتابخانه می توان بوسیله قطعه کد زیر و ورود تنظیمات لازم به پایگاه داده mysql خود متصل شد.

```
connection = mysql.connector.connect(  
    host='localhost',  
    port='8889',  
    database='stocks',  
    user='root',  
    password='root',  
)
```

شکل ۱۳. قطعه کد لازم برای اتصال به پایگاه داده mysql در پایتون.

3-4-3- ایجاد جدول

برای ذخیره پیام ها در پایگاه داده ایجاد شده احتیاج به ساخت جدولی در آن داریم که فرمت لازم برای ذخیره اطلاعات و ویژگی های پیام های استخراجی را داشته باشد. بدین منظور از کوئری زیر برای ایجاد جدول 'messages' در پایگاه داده 'stocks' استفاده کردیم.

```
CREATE TABLE messages (
  id int(11) NOT NULL AUTO_INCREMENT,
  message_id INT(11) NULL DEFAULT NULL,
  reply_to_message_id INT(11) NULL DEFAULT NULL,
  channel_id INT(11) NULL DEFAULT NULL,
  channel_name varchar(250),
  message TEXT CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci NULL DEFAULT NULL,
  found_namads TEXT CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci NULL DEFAULT NULL,
  found_companies TEXT CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci NULL DEFAULT NULL,
  sentiment_label TEXT CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci NULL DEFAULT NULL,
  signal_label TEXT CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci NULL DEFAULT NULL,
  views INT(11) NULL DEFAULT NULL,
  forwards INT(11) NULL DEFAULT NULL,
  replies INT(11) NULL DEFAULT NULL,
  published_at TIMESTAMP NULL DEFAULT NULL,
  created_at TIMESTAMP NULL DEFAULT NULL,
  params MEDIUMTEXT CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci NULL DEFAULT NULL,
  PRIMARY KEY (id))
```

شکل ۱۴. کوئری استفاده شده برای ایجاد جدول messages.

مشخصات ستون‌های جدول messages و توضیحات مرتبط با هر ستون در جدول ذیل ذکر شده است.

جدول ۲. مشخصات ستون‌های جدول messages.

ستون	نوع داده	توضیحات
id	int	شناسه پیام در سیستم ما
message_id	int	شناسه پیام در تلگرام
reply_to_message_id	int	شناسه پیامی که این پیام در پاسخ به آن داده شده است
channel_id	int	شناسه کانال
channel_name	varchar	نام کانال
message	text	متن پیام
found_namads	text	لیست نمادهای یافت شده در متن پیام به فرمت json
found_companies	text	نام شرکت‌های یافت شده در متن پیام به فرمت json
sentiment_label	text	برچسب احساسی زده شده
signal_label	text	برچسب سیگنالی زده شده
views	int	تعداد دفعات بازدید پیام
forwards	int	تعداد دفعات فوروارد پیام
replies	int	تعداد دفعات پاسخ به پیام

تاریخ انتشار پیام در تلگرام	timestamp	published_at
تاریخ ثبت پیام در سیستم ما	timestamp	created_at
لیست تمامی ویژگی‌های پیام به فرمت json	mediumtext	params

3-4-4- ذخیره پیام‌ها

با استفاده از کوئری زیر می‌توان پیام‌های استخراج شده را در جدول 'messages' ذخیره کرد.

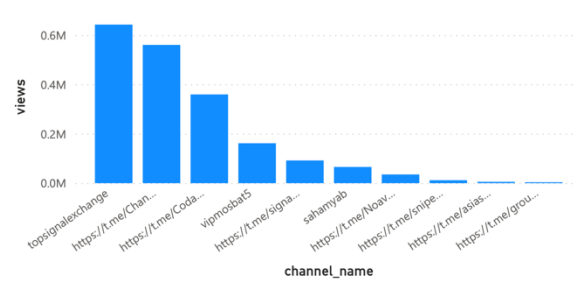
```
INSERT INTO messages (
  channel_name,
  channel_id,
  message_id,
  reply_to_message_id,
  message,
  found_namads,
  found_companies,
  sentiment_label,
  signal_label,
  views,
  forwards,
  replies,
  published_at,
  created_at,
  params
) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s)
```

شکل ۱۵. کوئری استفاده شده برای ذخیره پیام جدید در جدول messages.

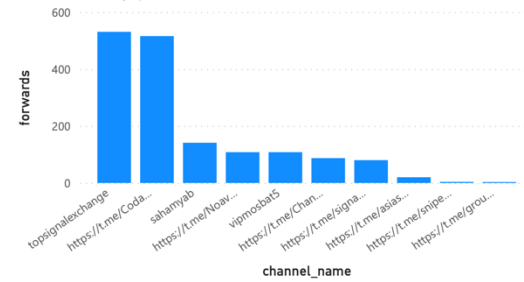
3-5- تولید داشبورد هوش تجاری

به منظور لمس بهتر و درک داده‌های جمع‌آوری شده مرحله اقدام به ساخت داشبورد هوش تجاری در نرم‌افزار PowerBI کردیم. تصاویر صفحات داشبورد تولید شده را می‌توانید در قسمت ذیل مشاهده بفرمایید:

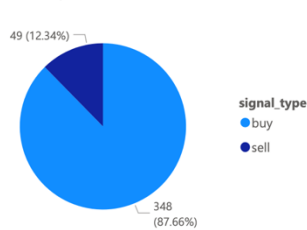
تعداد کل بازدید پیام ها به ازای هر کانال



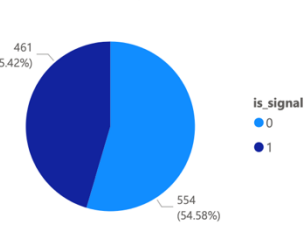
تعداد کل فرورارد پیام ها به ازای هر کانال



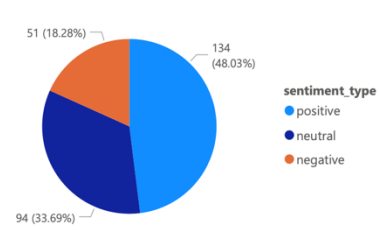
نسبت سیگنال های نوع خرید به فروش



نسبت پیام های سیگنالی به غیر سیگنالی

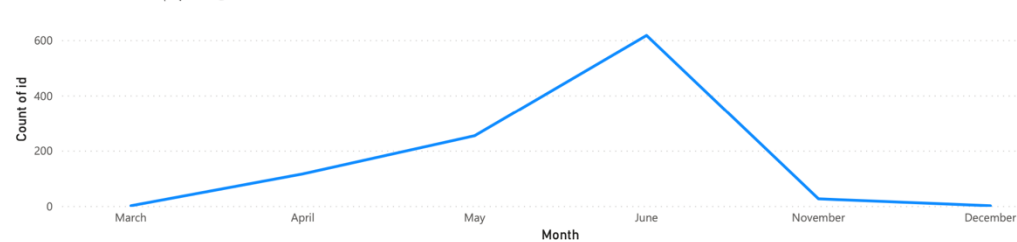


نسبت پیام هایی با احساس منفی به مثبت به خنثی

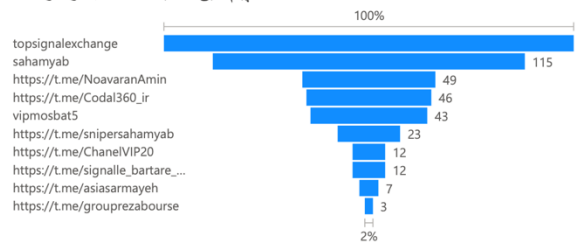


شکل ۱۶. صفحه اول از داشبورد هوش تجاری تولید شده.

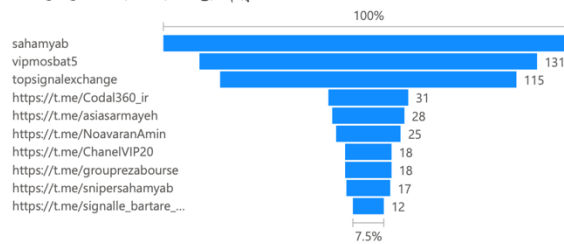
نمودار خطی تعداد پیام های منتشر شده در هر ماه



تعداد پیام هایی که سیگنال هستند به ازای هر کانال

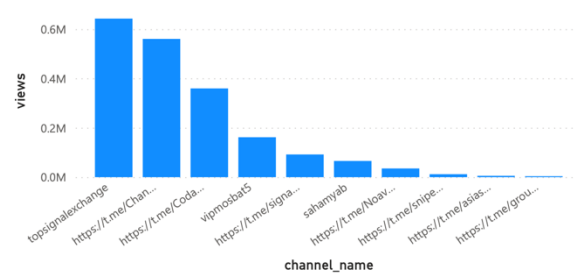


تعداد پیام هایی که سیگنال نیستند به ازای هر کانال

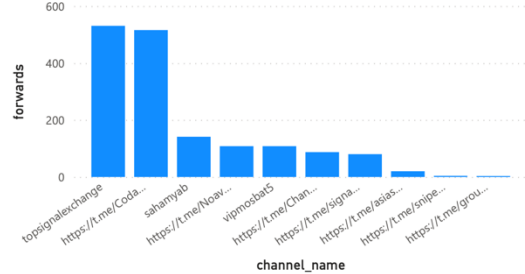


شکل ۱۷. صفحه دوم از داشبورد هوش تجاری تولید شده.

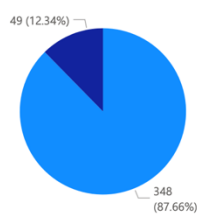
تعداد کل بازدید پیام ها به ازای هر کانال



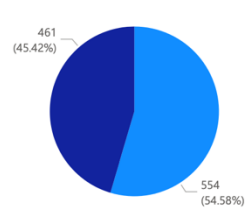
تعداد کل فرورارد پیام ها به ازای هر کانال



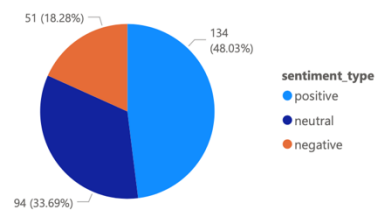
نسبت سیگنال های نوع خرید به فروش



نسبت پیام های سیگنالی به غیر سیگنالی



نسبت پیام هایی با احساس منفی به مثبت به خنثی



شکل ۱۸. صفحه سوم از داشبورد هوش تجاری تولید شده.

فصل چهارم

مدل سازی

مدل سازی

در این فصل به بررسی مدل های ساخته شده و روش های انجام وظایف این تمرین پرداخته شده است. این وظایف شامل استخراج نماد از متن پیام ها، تحلیل احساسات موجود در متن و تشخیص سیگنال در پیام ها می باشد.

4-1- استخراج نماد از متن پیام ها

اولین وظیفه مورد نظر در این پژوهش، استخراج نمادهای بورسی از متن پیام ها است. لیست تمامی نمادهای شرکت های بورسی از سایت شرکت مدیریت فناوری بورس تهران جمع آوری شده و در قالب یک لیست تعریف شده است. این لیست در برنامه ای به زبان پایتون قرار گرفته است که با دریافت یک متن، در صورتی که نام یکی از نمادهای بورسی در آن یافت شود، آن را مشخص نموده و نمایش می دهد.

4-2- تحلیل احساسات موجود در متن

دومین وظیفه مورد بررسی در این پژوهش، تحلیل احساسات موجود در متون است. برای بررسی این وظیفه تعداد چهار عدد مدل یادگیری ماشین ایجاد نمودیم. در دو عدد از این مدل ها که یکی با استفاده از دیتاست طاقچه و دیگری با استفاده از ایجاد شده توسط ما آموزش دیده اند، ویژگی های متن به صورت خودکار استخراج می گردد. این دو مدل با استفاده از شبکه پیش آموزش دیده برت فارسی (پارس برت¹) ایجاد شده است. در دو شبکه طراحی شده دیگر، جهت استخراج ویژگی ها از روشی نیمه دستی استفاده شده است که در ادامه به توضیح دقیق تر این موارد پرداخته شده است. این وظیفه یک طبقه بندی سه کلاسه شامل کلاس های مثبت (positive)، منفی (negative) و خنثی (neutral) است.

4-2-1- آموزش مدل با استخراج ویژگی خودکار

در این بخش به ایجاد دو عدد مدل پرداخته شده است. یکی از مدل ها با استفاده از دیتاست طاقچه و دیگری با استفاده از دیتاست جمع آوری شده توسط ما در راستای این پژوهش آموزش دیده اند. این مدل ها بر پایه مدل پارس برت ایجاد شده و بهینه سازی² شده اند. مدل پارس برت توسط آزمایشگاه

¹ ParsBert

² Fine Tune

هوش‌واره بر پایه مدل برت گوگل ایجاد شده و بر روی تعداد زیاد داده به زبان فارسی آموزش دیده است. برای انجام وظیفه تحلیل احساسات، این مدل بر روی داده‌های آموزشی ما بهینه‌سازی شده است.

ورودی شبکه برت برداری از شناسه‌های عددی کلمات واژه‌نامه آن است که در زمان پیش‌آموزش آن، آموزش دیده است. این واژه‌نامه همراه شبکه برت عرضه شده و به صورت عمومی قابل استفاده می‌باشد. این عملیات توسط tokenizer موجود در کتابخانه transformers انجام می‌شود. با دادن یک جمله به این tokenizer، ابتدا جمله به واحدهای کوچکتر زیرواژه^۱ شکسته شده و سپس شناسه معادل هریک از زیرواژه‌ها در واژه‌نامه آموزش دیده برت به صورت یک بردار (آرایه) بازمی‌گردد. این بردار به عنوان ورودی برای انجام وظیفه تحلیل احساسات داده شده و در لایه ورودی برت تبدیل به بردارهای Embedding به ازای هر توکن می‌شود. این لایه Embedding همان بردار ویژگی تولید شده به صورت خودکار است. طول پیشفرض این بردارها ۷۶۸ است.

خروجی شبکه برت نیز همانند ورودی آن برداری به طول ۷۶۸ است. جهت انجام طبقه‌بندی این خروجی را از یک لایه خطی از نورون‌ها عبور می‌دهیم. سائز ورودی این لایه ۷۶۸ و سائز خروجی آن به تعداد کلاس‌های طبقه‌بندی یعنی ۳ است.

همان‌گونه که قبل‌تر اشاره شد این مدل با دو دیتاست طاقچه و دیتاست تولید شده آموزش دیده‌اند. با توجه اینکه برجسب‌های دیتاست طاقچه به صورت سه کلاسه نبوده و به صورت امتیاز داده شده توسط کاربر به کالاهای، در قالب عدد ۱ تا ۵ است، امتیازات کمتر از ۳ به عنوان احساس منفی و بالاتر از آن به عنوان احساس مثبت تلقی شده است. مدل آموزش دیده با این دیتاست علاوه بر تست بر روی داده‌های تست دیتاست طاقچه، بر روی داده‌های تست دیتاست تولید شده توسط ما نیز ارزیابی شده و نتایج آن در فصل پنجم گزارش شده است.

4-2-2- آموزش مدل با استخراج ویژگی به صورت دستی

در این بخش به آموزش دو مدل جهت تحلیل احساسات پرداخته شده است که ویژگی‌های مورد استفاده در آن‌ها به صورت دستی با توجه به داده‌ها ایجاد شده‌اند. این دو مدل نیز همانند مدل‌های قبل، یکی بر روی دیتاست طاقچه و دیگری بر روی دیتاست تولید شده ما آموزش دیده‌اند.

جهت تولید بردارهای ویژگی داده‌ها از روش tf-idf استفاده شده است. ابتدا کل واژه‌های موجود در دیتاست در یک واژه‌نامه جمع‌آوری شده و در زمان پیش‌پردازش هر نمونه جهت طبقه‌بندی، مقدار tf-

^۱ SubWord

idf آن محاسبه و در خانه مخصوص آن در بردار ویژگی ها قرار می گیرد. در این بخش از طبقه بند Naïve Bayes استفاده شده است. نکته قابل توجه در مورد این روش این است که، به دلیل ایجاد دستی بردارهای ویژگی، امکان اعمال تغییرات و ضرایبی در آن وجود داشته و می توان با اعمال ضرایبی برای برخی از لغات، در عملکرد مدل تاثیر نهاد. بدین منظور پرکاربردترین کلمات در دیتاست ها به صورت دستی انتخاب شده و ترتیبی داده شده است که پس از محاسبه مقادیر tf-idf، مقدار این واژه ها تقویت شود. در فصل بعد نتیجه انجام این ضرایب گزارش شده است.

4-3- تشخیص سیگنال

در این بخش به ایجاد و آموزش دو مدل برای انجام وظیفه تشخیص و طبقه بندی سیگنال با استفاده از دیتاست تولید شده پرداخته شده است. این طبقه بندی داده ها را به سه طبقه بدون سیگنال (neutral)، سیگنال خرید (buy) و سیگنال فروش (sell) تقسیم می نماید.

4-3-1- آموزش مدل با استخراج ویژگی خودکار

در این بخش نیز، همانند مدل های قبل که از سیستم استخراج ویژگی خودکار در آن استفاده شده بود، شبکه برت فارسی به کار گرفته شده است. با وجود اینکه داده های آموزشی در این بخش نسبت به آموزش مدل با استفاده از دیتاست طاقچه بسیار کمتر است، با این حال به دلیل پیش آموزش خوبی که شبکه برت بر روی داده های فارسی دیده است، انتظار کسب نتایج قابل قبول، دور از انتظار نیست.

4-3-2- آموزش مدل با استخراج ویژگی دستی

همانند مدل با استخراج ویژگی دستی در وظیفه تحلیل احساسات، در این مدل نیز از معیار tf-idf در تولید بردارهای ویژگی استفاده شده است. با توجه به اینکه دیتاست طاقچه برای وظیفه تحلیل احساسات ایجاد شده است، در این بخش نمی توان از آن استفاده کرد. بنابراین تمامی عملیات آموزش و تست این مدل با استفاده از دیتاست تولید شده انجام خواهد شد. در این بخش نیز، در راستای پژوهش همان کلمات قسمت قبل انتخاب شده و وزن های آن ها، جهت بررسی اثر این عمل تقویت شده است.

فصل پنجم

بررسی نتایج و تحلیل آن‌ها

بررسی نتایج و تحلیل آن‌ها

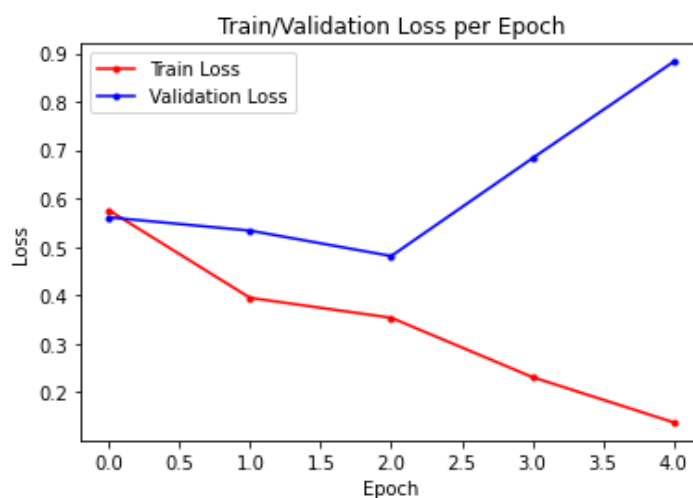
در این بخش از گزارش، به بیان نتایج حاصل از آزمایشات صورت گرفته بر روی مدل‌های ایجاد شده، در قالب جداول و نمودارها پرداخته شده است. همچنین سعی در تحلیل نتایج و بررسی علل برخی از رویدادها خواهد شد.

1-5- تحلیل احساسات

در این بخش چهار مدل آموزش داده شده است که نتایج تست آن‌ها در ذیل گزارش شده است. تمامی مدل‌ها با استخراج ویژگی خودکار به تعداد ۴ دوره^۱ آموزش دیده‌اند.

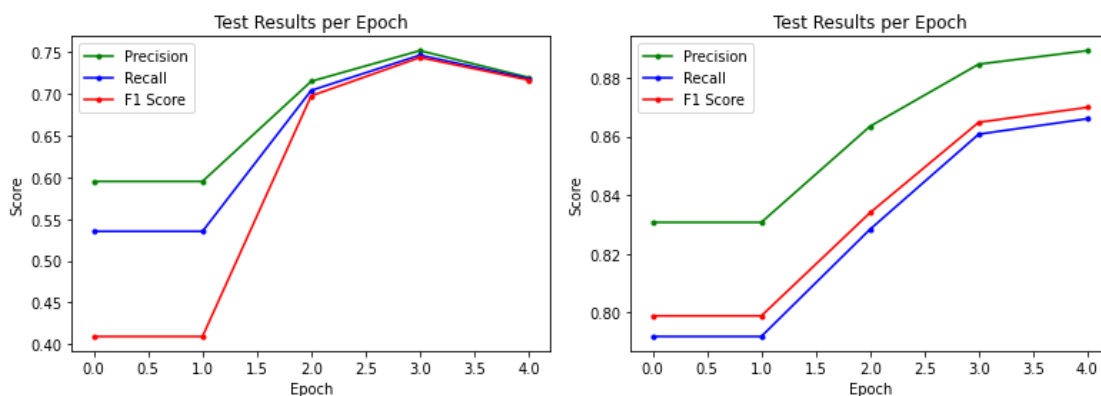
1-1-5- آموزش مدل با استخراج ویژگی خودکار:

استفاده از دیتاست طاقچه جهت آموزش:



شکل ۱۹. نمودار خطای آموزش و ارزیابی مدل.

^۱ Epoch

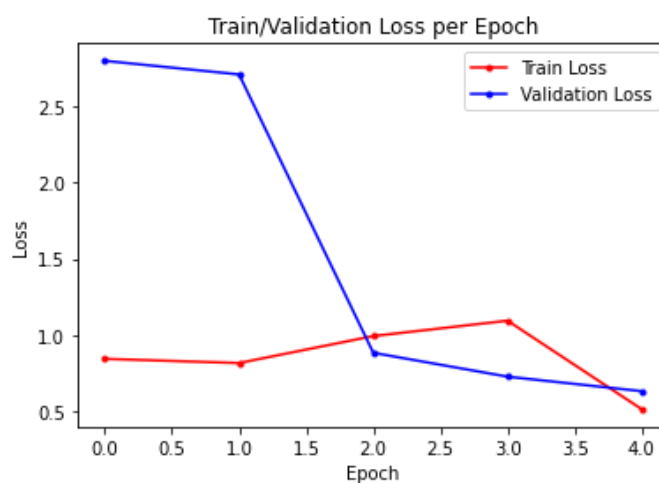


شکل ۲۰. نمودار معیارهای ارزیابی مدل بر روی داده‌های تست. سمت راست: داده‌های طاقچه. سمت چپ: داده‌های تست تولید شده.

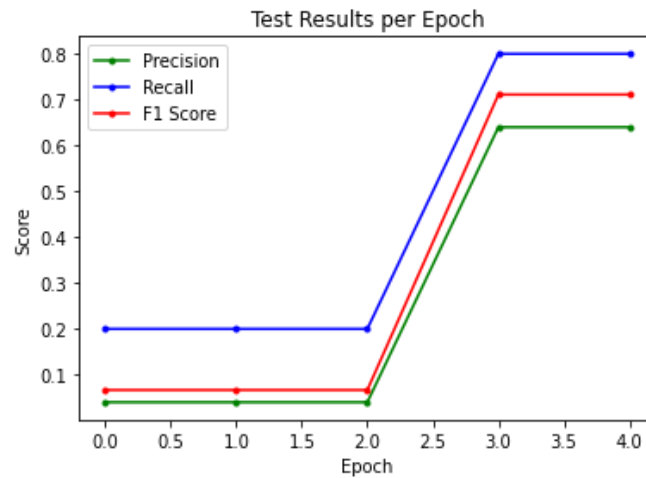
جدول ۳. مقدار معیار ارزیابی f1 مدل در دوره‌های آموزشی و دیتاست‌های گوناگون.

دیتاست تولیدی	دیتاست طاقچه	
0.4091	0.7988	دوره ۱
0.6977	0.8340	دوره ۲
0.7437	0.8648	دوره ۳
0.71672	0.8699	دوره ۴

استفاده از دیتاست تولید شده جهت آموزش:



شکل ۲۱. نمودار خطای آموزش و ارزیابی مدل.



شکل ۲۲. نمودار معیارهای ارزیابی مدل بر روی داده‌های تست تولید شده توسط خودمان.

جدول ۴ مقدار معیار ارزیابی f1 مدل در دوره‌های آموزشی و دیتاست‌های گوناگون.

دیتاست تولیدی	
0.06666	دوره ۱
0.06666	دوره ۲
0.71111	دوره ۳
0.71111	دوره ۴

5-1-2- آموزش مدل با استخراج ویژگی دستی:

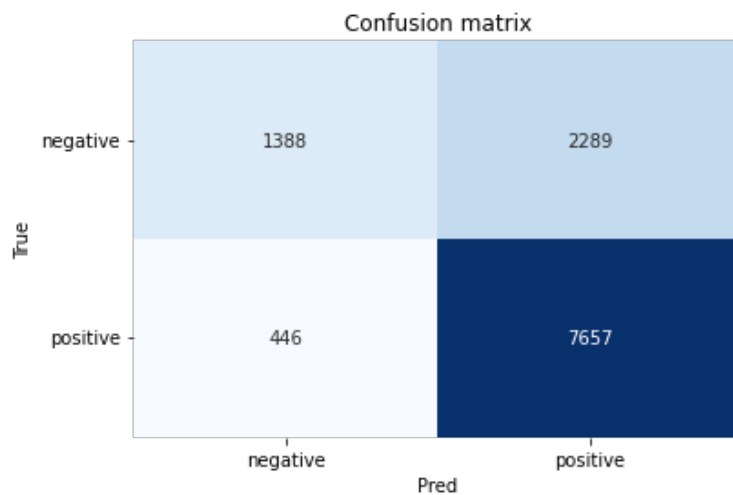
همان‌گونه که در بخش قبل گفته شد، استخراج ویژگی در این مدل‌ها با استفاده از مقادیر tf-idf صورت گرفته و از Naïve Bayes جهت طبقه‌بندی سه کلاس استفاده شده است.

آموزش مدل با استفاده از دیتاست طاقچه:

در این بخش مدل تحلیل احساسات با استفاده از دیتاست طاقچه آموزش دیده و نتایج تست آن در جدول ذیل ذکر شده است.

جدول ۵. نتایج ارزیابی مدل.

معیار f1	معیار Recall	معیار Precision	
0.77	0.77	0.77	آموزش با طاقچه



شکل ۲۳. ماتریس آشفتگی مدل.

آموزش مدل با استفاده از دیتاست تولید شده:

در این بخش مدل تحلیل احساسات با استفاده از دیتاست طاقچه آموزش دیده و نتایج تست آن در ذیل ذکر شده است.

جهت بررسی تاثیر وزن‌های tf-idf، تعدادی از کلمات پرکاربرد در سیگنال‌های بورسی به همراه ضرایبی برای هر کدام در نظر گرفته شده، و پس از محاسبه tf-idf در آن ضرب می‌شود. جدول ذیل شامل کلمات انتخابی می‌باشد.

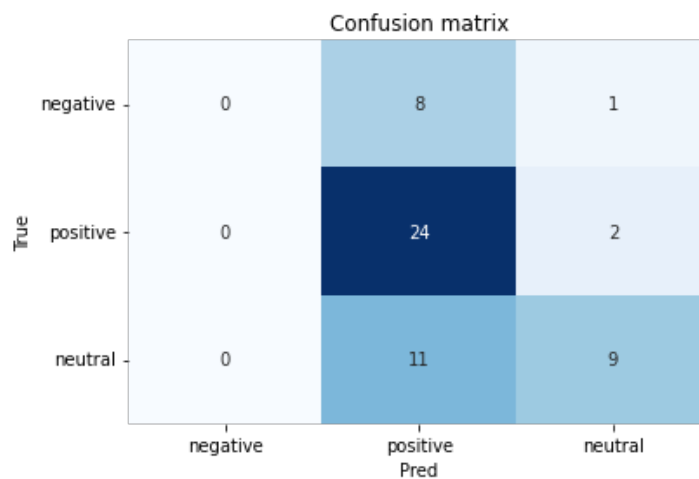
جدول ۶. نمونه‌ای از کلمات مثبت و منفی تاثیرگذار.

منفی	مثبت
کاهش	افزایش
افت	رشد

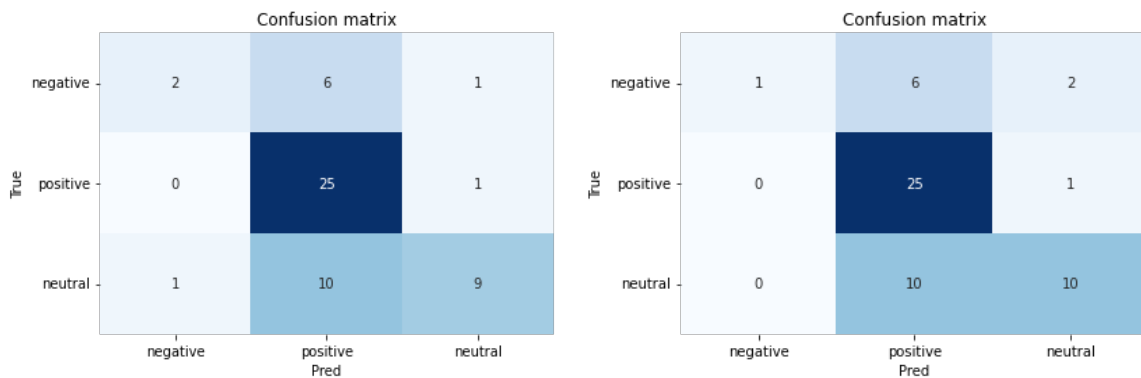
بهبود	حاصل نکرده
صف خرید	صف فروش
خرید	ریزش

جدول ۷. نتایج ارزیابی مدل.

معیار f1	معیار Recall	معیار Precision	
0.44	0.46	0.42	آموزش با ضریب یک
0.52	0.52	0.79	آموزش با ضریب ۱۰ برای کلمات
0.55	0.54	0.70	آموزش با ضریب ۲۰ برای کلمات



شکل ۲۴. ماتریس آشفتگی مدل با ضرایب ۱ برای کلمات.



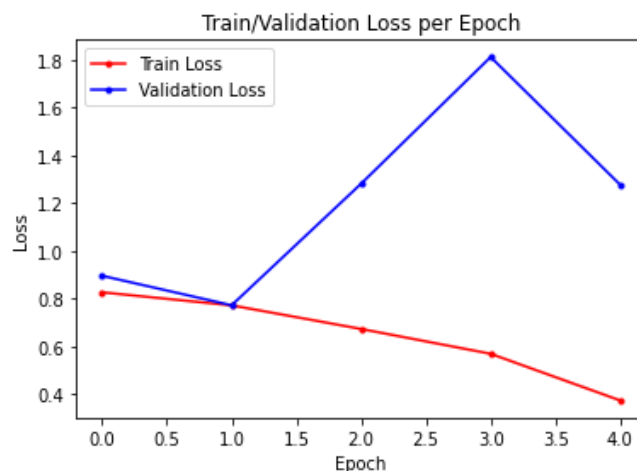
شکل ۲۵. ماتریس آشفتگی مدل با ضرایب ۱۰ برای کلمات. شکل ۲۶. ماتریس آشفتگی مدل با ضرایب ۲۰ برای کلمات.

همان‌گونه که پیش‌بینی می‌شد، انتخاب کلمات مهم در تشخیص احساسات پیام‌های مربوط به بورس و تقویت مقادیر آن در بردار tf-idf باعث افزایش کیفیت مدل شده است. این مقدار تا افزایش ضریب به میزان ۱۰ به سرعت زیاد شده و پس از آن از تاثیر آن کمتر می‌شود.

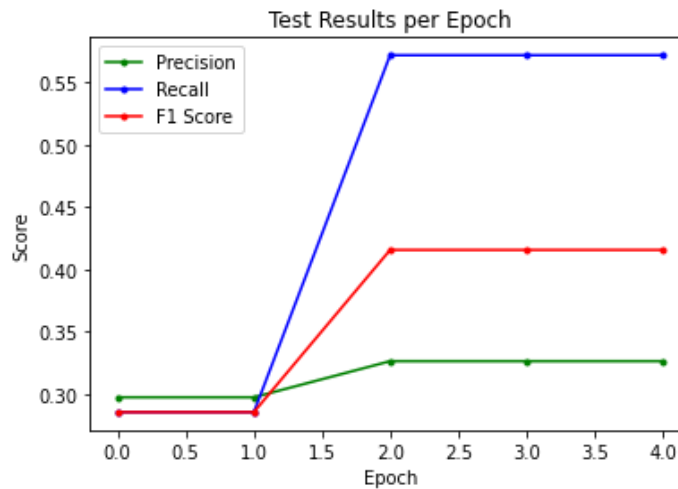
2-5- تشخیص سیگنال

در این بخش دو مدل با استخراج ویژگی خودکار و دستی آموزش داده شده است. آموزش این مدل‌ها با استفاده از داده‌های تولید شده توسط خودمان صورت می‌گیرد. در ادامه به ذکر نتایج حاصل از تست این مدل‌ها پرداخته شده است.

2-5-1- آموزش مدل با استخراج ویژگی خودکار:



شکل ۲۷. نمودار خطای آموزش و ارزیابی مدل.



شکل ۲۸. نمودار معیارهای ارزیابی مدل بر روی داده‌های تست تولید شده.

جدول ۸. مقدار معیار ارزیابی f1 مدل در دوره‌های آموزشی گوناگون.

معیار f1	
0.2857	دوره ۱
0.4155	دوره ۲
0.4155	دوره ۳
0.4155	دوره ۴

5-2-2- آموزش مدل با استخراج ویژگی دستی:

در این بخش نیز استخراج ویژگی در این مدل‌ها با استفاده از مقادیر tf-idf صورت گرفته و از Naïve Bayes جهت طبقه‌بندی سه کلاس استفاده شده است. این مدل با استفاده از دیتاست تولید شده آموزش دیده و نتایج تست آن در ذیل ذکر شده است.

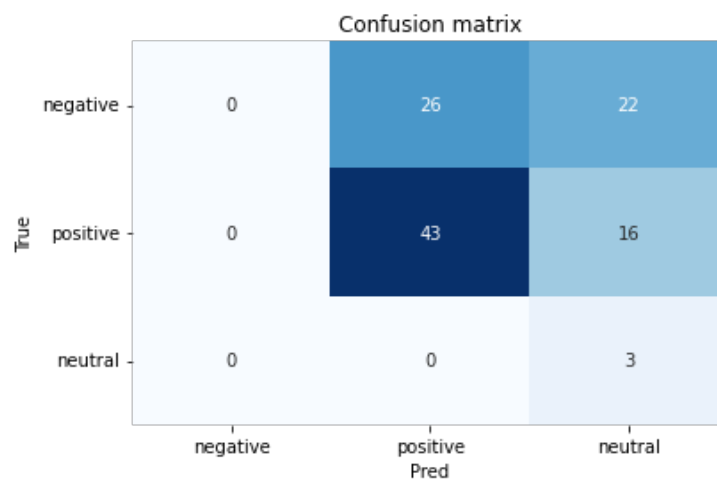
جهت بررسی تاثیر وزن‌های tf-idf، تعدادی از کلمات پرکاربرد در سیگنال‌های بورسی به همراه ضرایبی برای هر کدام در نظر گرفته شده، و پس از محاسبه tf-idf در آن ضرب می‌شود. جدول ذیل شامل کلمات انتخابی می‌باشد.

جدول ۹. نمونه‌ای از کلمات مثبت و منفی تاثیرگذار.

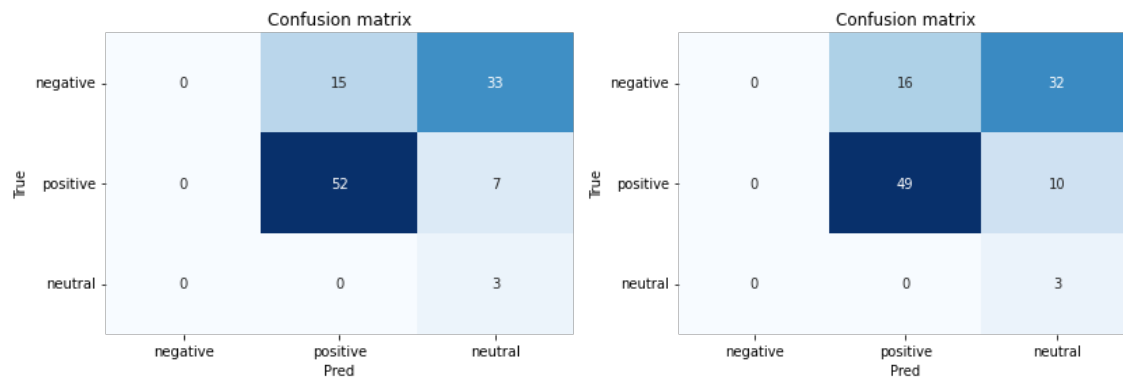
منفی	مثبت
کاهش	افزایش
افت	رشد
حاصل نکرده	بهبود
صف فروش	صف خرید
ریزش	خرید

جدول ۱۰. نتایج ارزیابی مدل.

معیار f1	معیار Recall	معیار Precision	
0.27	0.58	0.23	آموزش با ضریب یک
0.31	0.61	0.27	آموزش با ضریب ۱۰ برای کلمات
0.32	0.63	0.28	آموزش با ضریب ۲۰ برای کلمات



شکل ۲۹. ماتریس آشفتگی مدل با ضرایب ۱ برای کلمات.



شکل ۳۰. ماتریس آشفتگی مدل با ضرایب ۱۰ برای کلمات. شکل ۳۱. ماتریس آشفتگی مدل با ضرایب ۲۰ برای کلمات.

همان‌گونه که پیش‌بینی می‌شد، انتخاب کلمات مهم در تشخیص احساسات پیام‌های مربوط به بورس و تقویت مقادیر آن در بردار tf-idf باعث افزایش کیفیت مدل شده است. این مقدار همانند مدل تحلیل احساسات تا افزایش ضریب به میزان ۱۰ به سرعت زیاد شده و پس از آن از تاثیر آن کمتر می‌شود. همچنین در این مدل افزایش ضرایب کلمات، تاثیر کمتری در نتیجه نسبت به مدل تحلیل احساسات دارد.

فصل ششم

جمع‌بندی و نتیجه‌گیری و پیشنهادات

جمع‌بندی و نتیجه‌گیری

در این پژوهش به بررسی ابزارهای آنالیز متون در پیام‌های بورسی با استفاده از تکنیک‌های یادگیری ماشین پرداخته شد. این ابزارها شامل سه وظیفه یافتن خودکار نمادها از پیام‌های موجود در کانال‌های بورسی، تحلیل احساسات پیام‌ها و تشخیص سیگنال در پیام‌ها می‌باشد.

جهت تشخیص نمادها در پیام‌ها از یک جستجوی رشته‌ای در متون بر اساس لیست نمادهای استخراج شده از سایت شرکت مدیریت فناوری بورس تهران عمل شده است. جهت تحلیل احساسات ۴ مدل آموزش داده شده است که دو عدد از آن‌ها ابتدا با داده‌های دیتاست طاقچه آموزش داده شده و بر روی داده‌های تولید شده تست شده‌اند، و دو عدد دیگر با استفاده از داده‌های تولید شده آموزش دیده و تست شده‌اند. جهت تشخیص سیگنال در پیام‌ها نیز، دو مدل با استفاده از داده‌های تولید شده آموزش دیده است.

در بین مدل‌های آموزش داده شده برخی با استفاده از شبکه پارس‌برت طراحی شده‌اند که از سیستم استخراج ویژگی خودکار برت بهره می‌برند و برخی بر اساس طبقه‌بند Naïve Bayes ایجاد شده‌اند. در این مدل‌ها استخراج ویژگی با استفاده از محاسبه مقادیر tf-idf صورت گرفته است.

دو عدد از نتایج مهم کسب شده در این پژوهش عبارتند از اینکه مدلی که با استفاده از داده‌های طاقچه آموزش دیده و به صورت یادگیری انتقالی (Transfer Learning) بر روی داده‌های تولید شده تست شد، دارای عملکرد بسیار قابل قبولی بوده است. همچنین دریافتیم که با تولید دستی بردارهای ویژگی و تنظیم آن با کلمات مهم‌تر می‌توان به نتایج بهتری دست یافت.

پیشنهادهای

یکی از نقاط دارای چالش در این پژوهش، نبود دیتاست مخصوص و همچنین محدودیت و کم بودن تعداد داده‌های تولید شده است. در رفع محدودیت تعداد داده‌های آموزشی، می‌توان مدل‌های بهتری را آموزش داده و به نتایج بهتری دست یافت. کمک گرفتن از روش‌های جمع‌سپاری جهت تولید داده یا تولید داده به صورت نیمه خودکار، می‌تواند کمک شایانی به پیشرفت در این زمینه نماید.