دانشکده مهندسی کامپیوتر

تحلیل احساسات نسبت به بیت‌کوین در توییتر

پایان‌نامه برای دریافت درجه کارشناسی

در رشته مهندسی کامپیوتر

سیدمصطفی مسعودی

استاد راهنما:

دکتر صالح اعتمادی – دکتر عادل رحمانی

مهر ماه ۱۴۰۱

دانشکده مهندسی کامپیوتر

تحلیل احساسات نسبت به بیت‌کوین در توییتر

پایان‌نامه برای دریافت درجه کارشناسی

در رشته مهندسی کامپیوتر

سیدمصطفی مسعودی

استاد راهنما:

دکتر صالح اعتمادی – دکتر عادل رحمانی

شهریور ماه ۱۴۰۱


بسم الله الرحمن الرحیم

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پایان‌نامه/رساله

نام دانشکده:

نام دانشجو:

عنوان پایان‌نامه یا رساله:

تاریخ دفاع:

رشته:

گرایش:

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما				
۲	استاد راهنما				
۳	استاد مشاور				
٤	استاد مشاور				
٥	استاد مدعو خارجی				
٦	استاد مدعو خارجی				
۷	استاد مدعو داخلی				
۸	استاد مدعو داخلی				

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب **سید مصطفی مسعودی** به شماره دانشجویی **۹٦٤٦۲۰۱٤** دانشجوی رشته **مهندسی کامپیوتر** مقطع تحصیلی **کارشناسی** تأیید می‌نمایم که کلیه‌ی نتایج این پایان‌نامه/رساله حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: سید مصطفی مسعودی

امضا و تاریخ:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه/ رساله برای همگان بلامانع است.

☐ بهره‌برداری از این پایان‌نامه/ رساله با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه/ رساله تا تاریخ ..... ممنوع است.

نام استاد یا اساتید راهنما:

تاریخ:

امضا:

تشکر و قدردانی:
از اساتید راهنمای گرامی و مهندس امیرحسین امینی‌مهر که من را در طول انجام پایان‌نامه همراهی و راهنمایی کردند، تشکر و قدردانی می‌نمایم.

چکیده

حوزه‌ی رمزارز(به خصوص بیت‌کوین که یکی از محبوب‌ترین آن‌هاست) به تازگی به شدت مورد توجه مردم و متخصصان بازارهای مالی قرار گرفته و تلاش‌های زیادی برای ایجاد سود مالی در این زمینه صورت می‌گیرد. شبکه‌های اجتماعی هم بستر مناسبی برای ارائه‌ی نظرات توسط مردم هستند و روزانه حجم اطلاعات وسیعی توسط اقشار مختلف در آن منتشر می‌شود که توییتر یکی از پرکاربردترین شبکه‌های اجتماعی است. در سوی دیگر، یک فعالیت شناخته شده در هوش مصنوعی، تحلیل احساس نویسنده از متن، است. تحلیل احساسات کاربران شبکه‌های اجتماعی نسبت به بیت‌کوین یکی از زمینه‌های تحقیقاتی است که می‌تواند منجر به پیش‌بینی روند تغییرات قیمت بیت‌کوین شود. در این پایان‌نامه، ابتدا به روش جمع‌آوری داده‌های مرتبط با بیت‌کوین در توییتر، پرداخته شده است و با آن یک مجموعه داده نیز تولید شده. در ادامه همچنین مجموعه داده‌های از قبل منتشر شده را بررسی کرده و روش‌های موجود برای ایجاد برچسب مناسب برای فرآیند تحلیل احساس داده‌ها مطرح شده است. مقالاتی که در این زمینه منتشر شده‌اند اکثرا از روش‌های یادگیری ماشین مثل svm یا random forest استفاده کرده‌اند و کمتر از روش‌های یادگیری عمیق و مدل‌های پیچیده و قدرتمند استفاده شده. استفاده از مدل‌های ساده متشکل از سلول LSTM، مدل‌های زبانی به‌روزتری مثل BERT، مدل‌های بهبود یافته‌ی برت مثل RoBERTa و XLNet، همچنین نمونه‌های آموزش دیده روی متون مالی مثل FinBert، و مقایسه‌ی نتایج تمام این مدل‌ها، محور اصلی فعالیت‌های انجام شده در این پایان‌نامه است. پس از آزمایش‌ها و بررسی‌های انجام شده روی نتایج آموزش مدل‌ها با مجموعه داده‌های آماده شده، استفاده از مدل RoBERTa نتایج بهتری رقم می‌زند. در انتها برای ادامه‌ی کار، توصیه می‌شود تا در زمینه‌ی رمزارز مجموعه داده‌های بزرگ‌تر با برچسب‌های معتبرتری، که در صورت امکان توسط انسان تایید شده باشند، تهیه شود تا بتوان با آموزش روی آن‌ها مدل‌های با دقت بیشتر و مطمئن‌تری ارائه داد.

واژه‌های کلیدی: تحلیل احساسات – رمزارز – BERT – توییتر – یادگیری عمیق

فهرست مطالب

فهرست جداول

فصل ۱:

مقدمه

۱ـ۱ـ مقدمه

امروزه بازارهای مالی گسترش زیادی یافته است و محل مناسبی برای سرمایه‌گذاری با سرمایه‌های کم و زیاد اشخاص حقیقی است. به طور ویژه بازار خرید و فروش رمزارزها هم مشابه با بازارهای مالی رواج پیدا کرده است.

با گسترش تکنولوژی، شبکه‌های اجتماعی در بین مردم پررنگ‌تر شده و در زندگی روزانه مدام مورد استفاده قرار می‌گیرد. از محبوب‌ترین شبکه های اجتماعی می‌توان به توییتر[۱] اشاره کرد که به طور ویژه روزانه در آن توییت‌های[۲] زیادی در رابطه با رمزارزها توسط مردم مختلف منتشر می‌شود.

به کمک هوش مصنوعی می‌توان نظر و احساسات افراد را از پیام‌های آن‌ها در شبکه‌های اجتماعی تا حدودی استخراج کرد. تحلیل احساسات، یک فعالیت مطرح در حوزه‌ی پردازش زبان طبیعی است. برای پردازش و استخراج احساس جمله یا متن، روش‌های مختلفی وجود دارد، روش‌های مبتنی بر قواعد و روش‌های یادگیری ماشین و یادگیری عمیق، از جمله‌ی آن‌ها است که در فصل ۳: به توضیح مفصل آن‌ها پرداخته شده است.

در این پایان‌نامه به تحلیل احساسات کاربران توییتر نسبت به بیت‌کوین در توییت‌های مربوط به آن، با مدل‌های ساده و پیشرفته همچون LSTM، BERT و مشتقات آن‌ها پرداخته شده است. البته لازم به ذکر است که تحلیل احساسات به تنهایی قدرت و دقت لازم برای پیش‌بینی قیمت و روند تغییرات قیمت یک رمزارز را ندارد و بیشتر از نتایج آن استفاده می‌شود. در واقع با ترکیب احساس متن‌های منتشر شده توسط کاربران با پارامترهای دیگر موثر در تغییر قیمت، می‌توان به دقت خوبی برای پیش‌بینی قیمت رسید.

[۱] Twitter

[۲] Tweet: اصطلاحا به پیام‌هایی که کاربران در شبکه توییتر ارسال می‌کنند، توییت گویند.

فصل ۲:

مروری بر ادبیات موضوع

۲ـ۱ـ مقدمه

در زمینه‌ی تحلیل احساسات متن با موضوعات مختلف، کارهای زیادی انجام شده است اما به طور خاص کارهای انجام شده در زمینه‌ی متن‌های مرتبط با رمزارزها خیلی کم است. عموما مقالاتی که تحلیل احساسات را روی داده‌های شبکه‌های اجتماعی و مرتبط با رمزارز انجام داده‌اند از روش های یادگیری ماشین مثل svm، random forest و naïve bayes استفاده کرده‌اند و در بعضی مقالات به سراغ شبکه‌های عصبی ساده و مدل‌های بازگشتی مثل RNN و LSTM رفته‌اند.

یکی از مشکلات بزرگ در زمینه‌ی رمزارز ها، عدم وجود مجموعه داده‌ی متنی استخراج شده از شبکه‌ی اجتماعی، که معتبر و زیاد باشد و توسط انسان برچسب‌زنی شده باشد، است. در تعداد کمی مقاله داده‌ها شخصا جمع‌آوری شده و توسط نیروی انسانی برچسب خورده‌اند و در بیشتر مقالات برای برچسب‌زنی، بجای نیروی انسانی، به صورت اتوماتیک از روش‌های مبتنی بر قاعده و کتابخانه های موجود استفاده کرده‌اند.

۲ـ۲ـ کارهای مرتبط

مقاله‌ی Valencia[۱] از یک روش مبتنی بر قواعد به نام VADER[۲] که هماهنگ‌شده با داده‌های شبکه‌های اجتماعی طراحی شده، استفاده کرده است و با استفاده از ابزار طراحی شده توسط خالقین این الگوریتم، خروجی الگوریتم VADER را بدست آورده و به عنوان برچسب داده‌ها در نظر گرفته است. در این مقاله مدل‌های SVM، RandomForet و MLP تست شده است و بیان می‌کند که روش MLP نتیجه‌ی بهتری دارد. MLP یک شبکه عصبی ساده است. همچنین در این مقاله از روش‌های دیگری برای ادغام قیمت و داده‌های عددی با داده‌های متنی صورت گرفته است تا به پیش‌بینی قیمت بپردازد.

مقاله‌ی Nistor[۳] به تحلیل احساسات با مدل‌های عصبی بازگشتی پرداخته است. در این مقاله برای

آموزش مدل از مجموعه داده‌ی غیرمرتبط به رمزارز استفاده شده است که این یک نقطه ضعف برای آن می‌باشد. سپس مدل‌های LSTM، GRU و Attention را روی مجموعه داده آموزش داده و نتایج نشان می‌دهد که استفاده از Attention مقدار خیلی کمی نتایج را بهبود می‌بخشد. همچنین تعداد لایه‌ها و نورون‌های هر لایه هم با اعداد مختلف تست شده است و بیان می‌کند که دولایه LSTM نتیجه بهتری نسبت به تعداد لایه‌ی بیشتر دارد. در این مقاله سعی شده مدل‌های پیچیده‌تری با مدل‌های ساده مقایسه شود.

اما در مقاله‌ی Pant[۴] یک مجموعه داده جمع‌آوری و توسط نیروی انسانی برچسب‌زنی شده است که این یک نقطه‌ی قوت این کار می‌باشد. در ادامه در این مقاله از ۵ الگوریتم naïve bayes، multinomial naïve bayes، bernoulli naïve bayes، linear svm و random forest استفاده شده است و ادعا می‌کند که به دقت ۸۱ درصد برای تحلیل احساسات رسیده است. در قسمت دوم این مقاله هم به پیش‌بینی قیمت با کمک نتایج قسمت اول یعنی تحلیل احساسات، پرداخته است.

مقاله‌ی Aslam[۵] روی تحلیل احساسات[۱] و همچنین تشخیص حالت[۲] پرداخته است و مدل‌های یادگیری ماشین را با مدل طراحی شده‌ی خود(استفاده از یک لایه LSTM و یک لایه GRU پشت سر هم) مقایسه کرده است. همچنین ۴۰۰۰۰ نمونه جمع‌آوری شده و با استفاده از کتابخانه textblob و text2emotion برچسب‌زنی شده است. گزارش‌های این مقاله نشان می‌دهد استفاده از مدل طراحی شده‌ی دولایه نتیجه‌ی چشمگیر و بهتری نسبت به روش‌های یادگیری ماشین مطرح شده در مقاله مثل svm، linear regression و decision tree دارد.

[۱] Sentiment analysis: تحلیل احساس کاربر به صورتی که مشخص کند احساس کاربر مثبت،منفی یا خنثی است.

[۲] Emotion detection: تشخیص احساس کاربر در دسته‌های مختلف.مثلا خشمگین،ناراحت،خوشحال.

فصل ۳:

روش تحقیق

۳ـ۱ـ مقدمه

هدف این پایان‌نامه جمع‌آوری داده‌های شبکه اجتماعی توییتر و پیاده‌سازی و تحقیق روی تسک تحلیل احساسات و اجرای آن روی مدل‌های مختلف موجود و مطرح است. مراحل انجام این پایان‌نامه به ترتیب به این شکل است: جمع‌آوری داده، برچسب‌زنی داده، پردازش و آماده‌سازی داده، پیاده‌سازی مدل‌های مدنظر، آموزش مدل با داده‌ها و مقایسه و تحلیل نتایج. به صورت کلی نتیجه‌ی بدست آمده از تحلیل احساس متن می‌تواند در تعیین روند نوسانات قیمت رمزارز موثر باشد، این نتیجه ممکن است مستقیم و یا غیرمستقیم و با ترکیب با اطلاعات دیگر مربوط به قیمت، مورد استفاده قرار بگیرد، که در این پایان‌نامه به آن پرداخته نشده است. در این فصل به بیان مراحل و کارهای انجام شده در این پایان‌نامه با جزئیات پرداخته شده.

۳ـ۲ـ جمع‌آوری داده‌های توییتر

در این مرحله به جستجوی روشی برای جمع‌آوری داده از شبکه اجتماعی توییتر پرداخته شد. یکی از این راه‌ها استفاده از API[1] رسمی شرکت توییتر[۲] است که یک سری قابلیت در اختیار کاربران قرار می‌دهد. برای جمع‌آوری داده‌های مرتبط با یک رمزارز، استفاده از قابلیت سرچ توییتر گزینه‌ی مناسبی می‌باشد، اما به دلیل وجود محدودیت در درخواست(دسترسی به توییت‌های فقط یک هفته اخیر[۶]) برای جستجو که برای حساب‌های کاربری رایگان وجود دارد، استفاده از API گزینه مناسبی نمی‌باشد.

بنابراین گزینه‌ی مناسب برای جمع‌آوری داده‌های موردنظر استفاده از خزنده[۳] است، از این رو به بررسی خزنده‌های موجود در اینترنت پرداخته شد تا بهترین آن انتخاب و استفاده شود.

[1] Application Programming Interface: واسط برنامه‌نویسی برای تعامل کاربر با یک سامانه به منظور تبادل اطلاعات می‌باشد.

[۲] https://developer.twitter.com/en/docs/twitter-api

[۳] Crawler

جدول (۳ـ۱) بررسی خزنده‌های موجود برای جمع‌آوری داده از توییتر

نام خزنده	توضیحات	آخرین تاریخ بررسی
Snscrape	فعال است – بدون محدودیت توییتر – تمام امکانات جستجوی پیشرفته[۱] توییتر قابل استفاده است	۲۰ اسفند ۱۴۰۰
Tweepy	فعال است – یک واسط برای استفاده از API توییتر است و محدودیت‌های توییتر را دارد	۲۰ اسفند ۱۴۰۰
Scweet	فعال است – بدون محدودیت توییتر – از سلنیوم[۲]استفاده می‌کند و به کروم برای انجام جستجو نیاز دارد	۲۰ اسفند ۱۴۰۰
Twint	فعال است – بدون محدودیت توییتر – تمام امکانات جستجوی پیشرفته توییتر را پشتیبانی نمی‌کند	۲۰ اسفند ۱۴۰۰
getOldTweets3	غیرفعال است – بدون محدودیت توییتر – به دلیل تغییر آدرس جستجو در توییتر،دیگر کارایی ندارد	۲۰ اسفند ۱۴۰۰
TwitterScraper	غیرفعال است – بدون محدودیت توییتر – به دلیل وجود خطا در هنگام اجرای کد کارایی ندارد	۲۰ اسفند ۱۴۰۰

در جدول (۳ـ۱) نتایج جستجوهای انجام شده در رابطه با خزنده‌های مختلف آمده است که از بین آن‌ها Snscrape و scweet و Twint فعال و مناسب هستند ولی snscrape در تمام جهات کاربردی‌تر و کامل‌تر است و در نتیجه از آن می‌توان استفاده کرد. در این مرحله یک اسکریپت پایتون برای استفاده از snscrape و جستجوی کلمات کلیدی مربوط به بیت‌کوین مثل btc و bitcoin و همچنین هشتگ‌های آن‌ها (#btc, #bitcoin) و کشتگ‌های آن‌ها($btc, $bitcoin)، نوشته شده است و در انتها داده‌های جمع‌آوری شده در یک فایل اکسل ذخیره می‌شود. اطلاعات جمع‌آوری شده توسط خزنده شامل موارد زیادی است که بعضی از موارد مهم‌تر عبارت است از: متن توییت، نام کاربر نویسنده توییت، وضعیت تایید[۳] شدن کاربر نویسنده(تیک آبی)، تعداد لایک و ریپلای و بازتوییت. اما داده‌های جمع‌آوری شده خام هستند و همچنین برچسب ندارند که در ادامه به بررسی این موضوع پرداخته شده است.

[۱] **جستجوی پیشرفته توییتر** به همان شکل که در وبسایت آن نیز موجود است امکانات مختلفی در اختیار کاربر قرار می‌دهد.مثلا: امکان استفاده از عملگر OR بین کلمات کلیدی، تعیین بازه زمانی ارسال توییت، تعیین حداقل تعداد لایک، تعیین حداقل تعداد ریپلای، تعیین حداقل تعداد بازتوییت و گزینه‌های دیگر برای کوچک کردن محدوده جستجو

[۲] selenium

[۳] verify

۳ـ۳ـ برچسب‌زدن داده‌ها

برای تحلیل احساسات متن، برچسب می‌تواند یکی از گزینه‌های مثبت، خنثی یا منفی باشد. همچنین می‌توان برچسب را یک عدد در بازه‌ی ۱- تا ۱ در نظر گرفت با این فرض که هرچه به عدد ۱ نزدیک‌تر باشد احساس مثبت است و هرچه به عدد ۱- نزدیک‌تر باشد احساس منفی است و اعداد نزدیک ۰ هم احساس خنثی دارند. روش‌های مختلفی برای برچسب‌زنی می‌تواند مورد استفاده قرار بگیرد:

۱- با نیروی انسانی، که بهترین روش است اما هم به دانش کافی در حوزه مورد نظر و هم به زمان زیادی نیاز دارد و برای چنین پروژه‌ای بهینه نیست.

۲- با استفاده از الگوریتم‌ها و ابزارهای آماده موجود

a. استفاده از VADER: این مقاله روشی مبتنی بر قواعد ارائه داده است که مناسب تشخیص احساس برای متن‌های نوشته شده در شبکه‌های اجتماعی است. برای استفاده از این روش، می‌توان از کتابخانه آن در زبان پایتون[۱] استفاده کرد و به راحتی خروجی این الگوریتم برای هر متن را به عنوان برچسب در نظر گرفت. این الگوریتم ۴ خروجی دارد. سه خروجی negative، neutral و positive میزان احساس متن را برای هر دسته مشخص می‌کند. جمع این سه مقدار ۱ می‌باشد. به عنوان مثال خروجی یک نمونه می‌تواند به این شکل باشد: negative:0.6 , neutral:0.3 , positive:0.1 . از این نمونه می‌توان اینگونه برداشت کرد که بیشتر احساس متن منفی بوده و مقدار کمی احساس مثبت در آن مشاهده شده است. خروجی ۴ام با نام compound یک عدد در بازه‌ی ۱- تا ۱ است که طبق مستندات اصلی مقاله یک عدد ترکیبی است که اگر مقدار آن از 0.05 بیشتر باشد احساس متن مثبت، اگر از 0.05- کمتر باشد احساس متن منفی و در غیر اینصورت احساس متن خنثی است.

[۱] https://pypi.org/project/vaderSentiment

b. استفاده از کتابخانه TextBlob در پایتون[1] : این کتابخانه امکانات مختلفی در زمینه پردازش زبان طبیعی ارائه داده است که یکی از آن‌ها تحلیل احساس متن است. خروجی آن به شکل یک عدد با عنوان polarity است که در بازه ۱- تا ۱ می‌باشد. استفاده از این ابزار هم می‌تواند برای برچسب‌زنی مفید باشد.

۳ـ۴ـ مجموعه‌داده‌های آماده

در این قسمت به بررسی مجموعه داده‌های موجود و منتشر شده در زمینه‌ی مرتبط با تحقیق، پرداخته شده است. مجموعه داده باید از توییتر جمع شده باشد و مرتبط با رمزارز باشد. برای این کار مجموعه داده‌های منتشر شده در سایت Kaggle[2] و وب‌سایت‌های مشابه بررسی شده است و همچنین تعدادی از مقالات مرتبط در این زمینه مطالعه شده تا در صورت امکان از مجموعه داده‌های آن‌ها استفاده شود.

[1] https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis

[2] https://www.kaggle.com بستری برای انتشار مجموعه‌داده به صورت رایگان توسط هر فردی

جدول (۳ـ۲) لیست مجموعه داده‌های موجود متشکل از داده‌های مرتبط با رمزارز در توییتر

نام	رمزارز	توضیحات	بازه زمانی(میلادی)	برچسب	تعداد
Bitcoin tweets	بیت‌کوین	توییت‌هایی که در آن از هشتگ btc استفاده شده	از ۲۰۲۱/۲/۶ تا (به‌روزمی‌شود)	ندارد	2.5M
Bitcoin tweets-16M	بیت‌کوین	توییت‌هایی که در آن کلمه btc استفاده شده	از ۲۰۱۶/۱/۱ تا ۲۰۱۹/۳/۲۹	ندارد	16M
Bitcoin 17.7 M tweet and price	بیت‌کوین	برچسب داده‌ها باvader مشخص شده و میانگین مقدار آن برای توییت‌ها در بازه‌های یک ساعته ذخیره شده است.در نتیجه برچسب مناسب تحلیل احساسات نیست	از ۲۰۱۷/۸/۱ تا ۲۰۱۹/۱/۲۱	ندارد	13K
Btc tweets sentiment	بیت‌کوین	برچسب به صورت یک عدد(۱- یا ۰ یا ۱) می‌باشد اما نوع برچسب‌زنی توضیح داده نشده است	در ۲۰۱۸/۸/۲۳ از ساعت ۰۰:۰۰ تا ساعت ۹:۰۰	دارد	50K
Cryptocurrency tweets with sentiment analysis	بیت‌کوین لایت‌کوین و۶ مورد دیگر	داده‌های مرتبط با ۸ رمزارز جمع‌آوری شده و برچسب‌زنی با vader انجام شده است	از ۲۰۱۴/۱۰/۱۶ تا ۲۰۲۱/۲/۱۰	دارد	800K
Twitter emotion cryptocurrency	نامشخص	توییت‌های صرفا مرتبط با حوزه رمزارز و در ۶ دسته(عصبانیت، ترس، طمع، تنفر، ناراحتی و شادی)	نامشخص	دارد	3.5K

با مشاهده‌ی جدول (۳ـ۲) اطلاعات مناسبی از مجموعه داده‌ها می‌توان کسب کرد. دو مجموعه داده اول برچسب ندارد و داده خام است. مجموعه داده سوم هم برچسب ندارد و درواقع میانگین احساس پیام‌ها در بازه‌های زمانی یک ساعته را دارد که چون میانگین است و روی هر متن نیست، مناسب تسک تحلیل احساس متن نیست. مجموعه داده چهارم، عددی به عنوان برچسب در بازه ۱- تا ۱ دارد اما توضیحی در باره‌ی اینکه چگونه برچسب‌زنی انجام شده است نداده و نمی‌توان بدون اطلاع، از برچسب‌های آن استفاده کرد. مجموعه داده‌ی پنجم با روش VADER برچسب زده است و تعداد ۸۰۰ هزار نمونه دارد اما داده‌های آن مربوط به چندین رمزارز است که باید در صورت نیاز آن‌ها را فیلتر کرد و فقط از داده‌های مرتبط با بیت‌کوین استفاده کرد. مجموعه داده ششم هم دسته بندی در ۶ دسته انجام داده که مناسب نیست چون در

این پروژه احساسات مثبت و منفی و خنثی مد نظر است. به صورت کلی مجموعه داده معتبری که به صورت انسانی برچسب زده شده باشد و کارهای مختلفی توسط محققان دیگر روی آن انجام شده باشد، وجود ندارد که این یکی از ضعف‌های این زمینه است و ارائه‌ی یک مجموعه داده معتبر و مناسب می‌تواند بسیار مورد استقبال قرار بگیرد که جمع‌آوری آن می‌تواند یکی از کارهای آینده باشد.

۳ـ۵ـ مجموعه داده انتخابی

در این پروژه از مجموعه داده پنجم در جدول (۳ـ۲) استفاده شده است. این مجموعه داده دارای ۸۰۰ هزار نمونه توییتر از سال ۲۰۱۴ تا ۲۰۲۱ میلادی است. توییت‌ها، جمع‌آوری شده بر اساس سرچ روی کلمات کلیدی زیر است.

('ethereum', 'bitcoin', 'litecoin', 'tezos', 'ripple', 'yearn-finance', 'cardano', 'cryptocurrency')

که همه مربوط به حوزه رمزارز است.

برچسب‌های مشخص شده در این مجموعه داده با روش VADER محاسبه شده‌اند و به صورت چهار ستون neg, neu, pos, compound است.

سه ستون neg, neu, pos به صورت softmax هستند که یعنی جمع آن‌ها ۱ می‌شود.

با فیلتر داده‌ها بر اساس سرچ روی کلمه کلیدی bitcoin تعداد داده‌ها به حدود ۲۵۵۰۰۰ نمونه کاهش می‌یابد. توزیع داده‌ها برای سه دسته‌بندی موجود(خنثی،منفی و مثبت) به این روش حساب شد که بیشینه عدد بین سه مقدار خنثی و منفی و مثبت، برچسب نهایی آن داده خواهد بود. مثلا اگر اعداد یک نمونه به این شکل باشد، neg:0.6, pos:0.2, neu:0.2 ، برچسب این نمونه منفی در نظر گرفته می‌شود. با این فرمول توزیع داده‌ها در دسته‌بندی‌های موجود، در جدول آمده است.

جدول (۳ـ۳) توزیع داده‌های بیت‌کوین در مجموعه داده انتخاب شده

منفی	۴۵۷
خنثی	۲۵۳۹۵۷
مثبت	۱۱۶۷

۳ـ۵ـ۱ـ مشکل مجموعه داده انتخابی

یکی از مشکلات این مجموعه داده عدم تعادل داده‌ها در دسته‌بندی‌های مختلف است و تعداد نمونه‌ها در دسته‌ی خنثی بیش از حد بیشتر است. این عدم تعادل باعث می‌شود تا مدل بیشتر نتایج منفی و مثبت را هم خنثی پیش‌بینی کند و در این صورت precision و recall برای دسته‌ی منفی و مثبت کم خواهد بود. پس متعادل کردن مجموعه داده ضروری است.

۳ـ۵ـ۲ـ متعادل کردن مجموعه داده

همانطور که در قسمت «مشکل مجموعه داده انتخابی» بیان شد، نیاز به متعادل کردن داده‌ها بود. مجموعه داده به چهار شکل تغییر داده شد و آزمایش‌های مختلفی روی این چهار حالت مجموعه داده انجام شده است. نام‌گذاری و توضیحات این چهار حالت به شرح زیر است:

۱- **حالت Undersample** – یک روش مرسوم برای متعادل کردن undersample است،یعنی کاهش نمونه‌های هر دسته به، کمترین تعداد نمونه در بین دسته‌ها. در مجموعه داده انتخاب شده کمترین نمونه برای دسته منفی و با ۴۵۷ نمونه است. در این حالت ۴۵۷ نمونه منفی، ۴۵۷ نمونه خنثی و ۴۵۷ نمونه مثبت باید انتخاب شود اما برای دسته‌ی خنثی و مثبت ۵۰۰ نمونه انتخاب شده است که در مجموع ۱۴۵۷ نمونه در مجموعه داده در حالت undersample وجود دارد.

۲- **حالت ChangeLabel** – برای اینکه داده‌های مثبت و منفی بیشتری در مجموعه داده موجود باشد، روشی برای تغییر مقادیر برچسب‌ها در نظر گرفته شده به طوری که تعداد نمونه‌های مثبت و منفی بیشتر شود. برچسب‌ها با روش VADER روی مجموعه داده مشخص شده‌اند و سه مقدار مثبت و منفی و خنثی موجود است، روش تغییر مقادیر برچسب‌ها به این شکل انتخاب شده که مقادیر ستون منفی در صورتی که از ستون مثبت بیشتر باشد، ۲ برابر شود و همچنین مقادیر ستون مثبت برای نمونه‌هایی که از ستون منفی بیشتر است هم ۲ برابر شود و سپس سه مقدار منفی و مثبت و خنثی نرمالایزه شده تا جمع آن‌ها همچنان یک بماند. چند برابر کردن ستون‌ها به صورت چشمی

روی چند نمونه بررسی شدند و مقدار ۲ برابر نتایج بهتری را رقم زد. در نتیجه این تغییرات تعداد کل داده‌ها ۱۱۵۰۰ شد که منفی ۳۵۰۰، خنثی ۳۹۸۸ و مثبت ۴۰۱۰ داده شد.

۳- **حالت NoNeutral** – یک روش برای تسک تحلیل احساسات این است که فقط داده‌های منفی و مثبت را آموزش دهیم و برچسب‌ها فقط مثبت یا منفی باشد.

۱. **حالت NoNeutral_1600** حذف داده‌های خنثی از مجموعه داده اصلی که منجر به ۱۶۲۴ نمونه شد، منفی ۴۵۷ و مثبت ۱۱۶۷.

۲. **حالت Noneutral_7500** حذف داده های خنثی از مجموعه داده ChangeLabel که منجر به ۷۵۰۰ نمونه شد، منفی ۳۵۰۰ و مثبت ۴۰۱۰.

۳ـ۵ـ۳ـ انواع حالت برچسب‌ها

مجموعه داده سه ستون منفی،خنثی و مثبت دارد که جمع آن‌ها ۱ می‌شود. با توجه به این سه ستون ۳ فرم مختلف برای برچسب نهایی در نظر گرفته شده است که به طبع آن برای هر حالت باید در لایه‌ی آخر مدل‌ها هم تغییراتی ایجاد کرد تا متناسب با نوع برچسب باشد.

۱- **فرم softmax** : در این فرم دقیقا از همان مقادیر سه ستون منفی،خنثی و مثبت که جمع آن‌ها ۱ می‌شود استفاده شده و در لایه آخر مدل‌ها هم باید ۳ نورون وجود داشته باشد و از تابع فعال‌ساز softmax استفاده شود. در این حالت چون به صورت ۰ و ۱ اعلام نمی‌شود که برچسب نمونه چیست و مقادیر هر دسته یک عدد اعشاری بین ۰ تا ۱ است، نمی‌توان دقت حساب کرد و باید خطای مدل را با فرمول میانگین خطای مطلق[1] بدست آورد و مقایسه بهبود مدل باید از این معیار صورت گیرد.

۲- **فرم one-hot** : در این فرم ستونی که مقدار بیشینه را دارد به عنوان برچسب مطلق آن نمونه در نظر گرفته می‌شود و به فرم بردار one-hot فقط مقدار همان ستون بیشینه را ۱ گذاشته و بقیه ستون

[1] MAE یا mean absolute error یا میانگین خطای مطلق = $1/n * (\sum | x_i - y_i |)$

ها ۰ خواهند بود. مثلا از حالت neg:0.1, neu:0.4, pos:0.5 به حالت neg:0, neu:0, pos=1 تغییر خواهد کرد. مدل در لایه آخر ۳ نورون خواهد داشت و تابع فعال‌سازی همان softmax خواهد بود و در این فرم می‌توان دقت مدل را گزارش کرد، چون برچسب‌ها به صورت ۰ و ۱ هستند.

۳- **فرم regression** : در این فرم حاصل تفریق ستون مثبت از ستون منفی که یک عدد بین ۱- تا ۱ می‌شود به عنوان برچسب در نظر گرفته می‌شود و مدل باید در لایه اخر ۱ نورون داشته باشد و از تابع فعال‌سازی tanh استفاده کرده و یک عدد در بازه ۱- تا ۱ پیش‌بینی کند. در این حالت هم فقط می‌توان خطا را با فرمول‌هایی مثل میانگین خطای مطلق محاسبه و مقایسه کرد.

۳ـ۶ـ مجموعه داده جمع‌آوری شده با خزنده(CrawledData)

همان‌طور که قبلا بیان شد یک اسکریپت پایتون نوشته شده که از خزنده snscrape استفاده می‌کند و در بازه‌ی زمانی مشخص داده‌های توییتر را جمع‌آوری می‌کند. در این قسمت داده‌های از تاریخ ۲۰ فوریه ۲۰۲۱ تا ۱۵ مارس ۲۰۲۱، با جستجوی مرتبط با بیت‌کوین، با اجرای کد بدست آمده است. سپس برای برچسب زدن آن‌ها از کتابخانه VADER استفاده شده است. توزیع داده‌ها در دسته‌ها به صورت زیر می‌باشد.

جدول (۳ـ۴) توزیع داده‌ها در مجموعه داده جمع‌آوری شده

	اولیه	کوچک شده
منفی	۲۶۶۴	۲۶۶۴
خنثی	۱۸۸۷۲	۳۰۰۰
مثبت	۷۵۵۶	۳۰۰۰

این مجموعه داده با روش undersample کوچک شده و در نهایت با اندازه ۸۶۶۴ نمونه مورد استفاده قرار گرفته است و با اسم CrawledData در ادامه‌ی گزارش استفاده شده است.

۳ـ۷ـ پیش‌پردازش‌های انجام شده روی مجموعه داده

مجموعه داده انتخاب شده نیاز به پیش‌پردازش داشت. بنابراین پیش‌پردازش‌های زیر به ترتیب روی داده‌ها اعمال شد:

- حذف url
- حذف کد html
- حذف mention کاربران در متن توییت
- حذف stop word
- بن‌واژه‌سازی[۱] کلمات
- حذف صورتک‌ها[۲]
- تبدیل بعضی مخفف‌های رایج زبان انگلیسی در شبکه‌های اجتماعی به فرم گسترده آن‌ها مثلا G9=Genius
- تصحیح غلط املایی مثلا haert -> heart
- حذف اعداد
- حذف علائم نگارشی و نشانه‌گذاری[۳]

۳ـ۷ـ۱ـ تصحیح غلط املایی

یکی از مراحل پیش‌پردازش تصحیح غلط‌های املایی است که ممکن است در داده‌های شبکه‌های اجتماعی بیشتر دیده شود. برای تصحیح غلط املایی دو کتابخانه به صورت آزاد موجود است:

- کتابخانه Pyspellchecker[4] که از یک روش به نام norving[1] استفاده کرده.

[۱] Lemmatization: یافتن ریشه‌ی بامعنی کلمه

[۲] emojy

[۳] Punctuation: علائم نگارشی مثل ویرگول، علامت تعجب، علامت سوال، پرانتز، براکت و امثال این‌ها

[4] https://pypi.org/project/pyspellchecker

- کتابخانه symspellpy[2] که نسبت به روش norving سرعت بالاتری دارد اما با بررسی چند مثال دیده شد که بعضی از کلمات را به شکل بدی تغییر می‌دهد. مثلا noooooo را به not good تبدیل می‌کند.

به دلیل اینکه pyspellchecker نتیجه‌ی بهتری برای تصحیح کلمات دارد، از آن استفاده شده است. در این قسمت به بررسی تاثیر تحصیح غلط املایی بر خطای مدل پرداخته شده است. در واقع یک بار داده‌ها با تصحیح غلط املایی و یک‌بار بدون تصحیح غلط املایی پیش‌پردازش شده‌اند و روی مدل ساده که از ۲ لایه BiLSTM تشکیل شده است، آموزش دیده‌اند. به دلیل زمان‌بر بودن تصحیح با کتابخانه pyspellchecker، مجموعه داده اصلی بدون اینکه روی داده‌های مرتبط با بیت‌کوین فیلتر شوند، کوچک شده‌اند. یک مجموعه داده با 1500 نمونه(از هر دسته ۵۰۰ نمونه) و یکی دیگر با تعداد ۳۰۰۰ نمونه(از هر دسته ۱۰۰۰ نمونه) جدا شده است.

سپس نتایج بدست آمده از آموزش داده‌ها روی مدل ذکر شده را مقایسه کرده و **نتیجه آن شد که عدم استفاده از تصحیح غلط املایی، بهتر است و خطای مدل بعد از آموزش کمتر شده است.**

۳ـ۸ـ مدل

در این پایان‌نامه آزمایش‌های مختلفی با مجموعه داده‌های موجود، روی مدل‌های مختلفی انجام شده است. مدل‌های مورد استفاده به شرح زیر است:

۱- سلول LSTM – سلول Long Short-Term Memory(LSTM) یک نوع سلول بازگشتی[3] است اما با سلول‌های بازگشتی ساده که به نام RNN شناخته می‌شوند، تفاوت دارد. این سلول از ۴ گیت به نام cell، input gate، output gate و forget gate تشکیل شده است که در نهایت ایده‌ی اصلی، کنترل جریان اطلاعات در طول رشته‌ی داده است به شکلی که اگر اطلاعات مهمی در ابتدای رشته وجود داشته باشد، فراموش نشود و حفظ شوند. در ادامه مدل‌هایی متشکل از همین سلول LSTM

[1] https://norvig.com/spell-correct.html
[2] https://pypi.org/project/symspellpy
[3] Recurrent

برای مقایسه در نظر گرفته شده‌اند.

a. مدل 1-BiLSTM – از ۱ لایه LSTM دوطرفه[۱] تشکیل شده است.
b. مدل 2-BiLSTM – از ۲ لایه LSTM دوطرفه تشکیل شده است.
c. مدل 4-BiLSTM – از ۴ لایه LSTM دوطرفه تشکیل شده است.
d. مدل 10-BiLSTM– از ۱۰ لایه LSTM دوطرفه تشکیل شده است.

۲- Base-BERT – مدل BERT[۷] یک مدل زبانی قدرت‌مند است که می‌توان از آن برای تسک‌های مختلف هوش مصنوعی مثل تحلیل احساسات استفاده کرد. برت[۲] در واقع مدل و تکنیک Transformer ها که یک مدل مشهور بر پایه attention است را به صورت دوطرفه به کار گرفته است. برت دو نسخه پایه و بزرگ دارد که نسخه پایه از ۱۲ لایه encoder یا همان transformer block تشکیل شده است. برت روی مجموعه داده‌های بزرگ BookCorpus و English Wikipedia Data آموزش دیده است که باعث شده تا دانش خوبی نسبت به کلمات زیادی داشته باشد و بازنمایی مناسبی برای کلمات ارائه کند. برت پایه، به ازای هر کلمه در جمله‌ی ورودی یک بردار با اندازه‌ی ۷۶۸ در خروجی دارد. اندازه طول جمله در این پایان‌نامه، ۱۲۸ در نظر گرفته شده است. در نتیجه در لایه آخر ۱۲۸ بردار با اندازه‌ی ۷۶۸ موجود است. یکی دیگر از خروجی‌های برت میانگین ۱۲۸ بردار خروجی است که در نهایت یک بردار به طول ۷۶۸ خواهد بود که با اسم pooled_output استفاده شده است. در این پایان‌نامه از خروجی pooled_output استفاده شده و بعد از آن یک یا چند لایه‌ی نهایی اضافه شده است که متناسب با خروجی مورد نظر ما برای تحلیل احساسات است. مثلا در تمام آزمایش‌ها در انتها یک لایه شبکه عصبی ساده با ۳ نورون قرار داده شده است که هر نورون به ازای یکی از دسته‌های احساس (منفی، خنثی یا مثبت) است.

۳- FinBert[۸] – همان مدل Base-BERT است که روی متون مالی[۳] آموزش داده شده است. امید است

[۱] Bidirectional LSTM: دو سلول LSTM که از دو سمت مختلف رشته را تحلیل می‌کنند. یکی از کلمه اول جمله شروع می‌کند و دیگری از کلمه آخر جمله شروع می‌کند و در نهایت اطلاعات آن‌ها باهم ترکیب می‌شود.

[۲] BERT

[۳] Finance corpus

که کلمات و بازنمایی کلمات که توسط FinBert بدست آمده‌اند به یادگیری بهتر احساسات داده‌های مرتبط با بیت‌کوین منجر شود. خروجی مورد استفاده و لایه‌ی آخر در این روش هم همانند روش قبل در نظر گرفته شده است.

۴- RoBERTa[۹] – این مدل همان معماری برت را دارد با این تفاوت که با تغییر بعضی از تکنیک‌های برت، یادگیری مدل را بهبود بخشیده و همچنین حجم داده‌هایی که روی آن آموزش دیده است چندین برابر برت است.

۵- XLNet[۱۰] – طبق مطالب بیان شده در مقاله اصلی، این مدل از نظر معماری مشابه برت است و در نحوه‌ی آموزش دیدن با برت متفاوت است. برت از روش autoanecoding استفاده می‌کند و XLNet از روش autoregressive. تفاوت این دومورد، در روش استفاده شده برای آموزش داده‌ها است. طبق ادعای مقاله، این مدل در حدود ۲۰ تسک دیگر از مدل برت بهتر عمل کرده است.

فصل ٤:

نتایج و تفسیر آنها

۴ـ۱ـ مقدمه

در فصل ۳: مجموعه داده‌های انتخاب شده و جمع‌آوری شده و حالت‌های مختلف آن‌ها، با تغییرات و مشخصات هر کدام، به طور دقیق شرح داده و نام‌گذاری شده‌اند. همچنین مدل‌های مختلف مورد استفاده در این پایان‌نامه هم، شرح داده و نام‌گذاری شده‌اند. در مسیر انجام پایان‌نامه آزمایش‌های مختلفی مورد بررسی و مقایسه قرار گرفته شده و تلاش شده در هر آزمایش و مقایسه‌ی انجام شده، تمام شرایط و پارامترها ثابت مانده و فقط یک پارامتر تغییر کند تا نتیجه‌گیری‌ها قابل قبول باشد. در ادامه به بیان آن‌ها پرداخته شده و نتیجه‌گیری هر کدام ذکر شده است.

۴ـ۲ـ مقایسه حالت با و بدون تصحیح غلط املایی در پیش‌پردازش

برای این آزمایش مجموعه داده آماده‌ی انتخاب شده، بدون فیلتر کردن داده‌های بیت‌کوین، مورد استفاده قرار داده شده و یک‌بار به تعداد ۱۵۰۰ نمونه و بار دیگر به تعداد ۳۰۰۰ نمونه کاهش یافته است.
مدل: 2-BiLSTM | لایه خروجی مدل: لایه ساده با ۳ نورون | تابع فعال‌ساز: softmax | حالت برچسب: فرم softmax | تابع ارزش[۱]: MeanAbsoluteError | بهینه‌ساز: Adam

جدول (۴ـ۱) تاثیر تصحیح غلط املایی در پیش‌پردازش در میزان خطای مدل

	خطای داده ارزیابی با تصحیح املا	خطا داده ارزیابی بدون تصحیح املا
۱۵۰۰ نمونه (هر دسته ۵۰۰ نمونه)	**0.067**	**0.064**
۳۰۰۰ نمونه (هر دسته ۱۰۰۰ نمونه)	**0.051**	**0.042**

نتیجه: استفاده از تصحیح غلط املایی خطا را بیشتر می‌کند و ترجیه عدم استفاده از آن است.

[۱] Cost function

۴ـ۳ـ بررسی تاثیر تعداد لایه در مدل متشکل از LSTM

مدل: 1-BiLSTM,…,10-BiLSTM | مجموعه داده: undersample | حالت برچسب: فرم softmax | لایه آخر مدل: ۳ نورون ساده | تابع فعال‌ساز: softmax | تابع ارزش: MeanAbsolteError | بهینه‌ساز: Adam

جدول (۴ـ۲) بررسی تاثیر تعداد لایه در خطای مدل ساده با استفاده از مجموعه داده undersample

	Train MAE[1]	Validation MAE
1-BiLSTM	0.016	0.084
2-BiLSTM	0.017	0.078
4-BiLSTM	0.022	0.078
10-BiLSTM	0.21	0.22

نتیجه: استفاده از لایه‌های بیشتر BiLSTM و عمیق‌تر شدن مدل کمکی به کاهش خطا نمی‌کند و حتی در مدل ۱۰ لایه میزان خطا خیلی افزایش می‌یابد. شاید بتوان دلیل این رویداد را رخ دادن gradient vanishing دانست. Gradient vanishing یا ناپدید شدن گرادیان درواقع به دلیل عمیق شدن لایه‌ها رخ می‌دهد و باعث می‌شود گرادیان عدد خیلی کوچکی شود و مدل نتواند به راحتی و در جای مناسبی همگرا شود. عموما برای رفع این مشکل از تکنیک residual block استفاده می‌شود تا با عمیق شدن مدل، گرادیان خیلی کاهش نیابد. در مجموع استفاده از 4-BiLSTM نتیجه‌ی مناسبی در مقایسه با بقیه داشته و چون ظرفیت این مدل هم به اندازه 1-BiLSTM کم نیست، مفیدتر است چون ممکن است با داده‌های بیشتر نیاز باشد از مدل با ظرفیت بیشتر استفاده شود.

[1] MAE: Mean Absolute Error

۴ـ۴ـ انتخاب لایه های دسته‌بند[۱] بعد از خروجی مدل Base-BERT

بعد از خروجی pooled_output از مدل برت، می‌توان یک یا چند لایه با معماری دلخواه گذاشت و تاثیر آن‌ها را بررسی کرد. در اینجا منظور از دسته‌بند همان لایه‌های نهایی تعبیه شده بعد از خروجی برت است. در این بخش دو روش برای دسته‌بند در نظر گرفته شده و نتیجه در جدول زیر قابل مشاهده است.

مدل: Base-BERT | مجموعه داده: undersample | حالت برچسب: فرم one-hot | تابع فعال‌ساز لایه آخر: softmax | تابع ارزش: CategoricalCrossEntropy | بهینه‌ساز: Adam

جدول (۴ـ۳) مقایسه دو نوع دسته‌بند برای مدل Base-BERT روی مجموعه داده undersample

	Classifier layers	Train accuracy / validation accuracy
۱	Pooled_output+batchNorm+dropout+3dense	1 / 0.91
۲	Pooled_output+batchNorm+dropout+128dense+3dense	0.99 / 0.91

نتیجه: نوع های دیگری از دسته‌بند هم امتحان شده بود که متاسفانه به دلیل عدم ذخیره‌سازی اطلاعات در اینجا گزارش نشده است و تنها این دومدل آورده شده است اما در کل استفاده یا عدم استفاده از لایه‌ی dense اضافی در دسته‌بند ، تاثیر زیادی روی دقت مدل برای مجموعه داده undersample نداشته است.

۴ـ۵ـ مقایسه مدل‌های Base-BERT ، FinBert، 4-BiLSTM و RoBERTa

در این قسمت مدل‌های 4-BiLSTM، Base-BERT، FinBert و RoBERTa روی یک سری از مجموعه داده‌ها آموزش داده شده‌اند و دقت آن‌ها با هم مقایسه شده. مدل RoBERTa فقط برای مجموعه داده‌های undersample و ChangeLabel استفاده شده است.

حالت برچسب: فرم one-hot | لایه آخر مدل: ۳ نورون ساده | تابع فعال‌ساز لایه آخر: softmax | تابع

[۱] Classifier

ارزش: CategoricalCrossEntropy | بهینه‌ساز: Adam

جدول (۴ـ۴) گزارش دقت داده آموزش و داده ارزیابی به ترتیب ،برای مدل‌های مختلف، روی مجموعه داده‌های مختلف در حالت برچسب one-hot

مدل / مجموعه‌داده	4-BiLSTM	Base-BERT	FinBert	RoBERTa
Undersample	0.99/0.88	0.99/0.91	0.99/0.90	0.99/0.93
ChangeLabel	0.99/0.77	0.99/0.86	0.99/0.86	0.99/0.90
NoNeutral_1600	0.99/0.91	1.0/0.97	0.99/0.94	-
NoNeutal_7500	0.99/0.91	0.99/0.94	0.99/0.94	-

نتیجه: در تمام مجموعه داده‌ها، دقت مدل برت از مدل 4-BiLSTM بهتر است. مدل finBert که از وزن‌ها و بازنمایی‌های متفاوتی نسبت به برت استفاده می‌کند و انتظار می‌رفت که از برت بهتر باشد نتیجه بهتری نداشته است و نهایتا می‌توان گفت همان نتیجه مشابه برت را خواهد داشت. یکی از دلایل می‌تواند این باشد که متون مالی‌ای که finBert روی آن آموزش دیده رسمی بوده و مناسب داده‌های شبکه اجتماعی نبوده است. مدل RoBERTa در مجموعه داده‌های undersample و changeLabel بررسی شده است و در هر دو حالت نسبت به برت نتیجه‌ی خیلی بهتری دارد. در نتیجه می‌توان روبرتا را به عنوان بهترین مدل معرفی کرد.

دقت نتایج مجموعه داده ChangeLabel کمتر از دقت نتایج مجموعه داده Undersample هستند، اما نمی‌توان مقایسه‌ی درستی انجام داد، چون در ChangeLabel تعداد داده‌ها بیشتر شده و این خود می‌تواند عاملی باشد که دقت کاهش یافته است چون با افزایش داده‌ها الگوهای بیشتری به وجود می‌آید و یادگیری و تفکیک آن‌ها از هم سخت‌تر می‌شود. اما اگر فرض کنیم افزایش داده‌ها موجب کاهش دقت نشده، می‌توان این نتیجه را گرفت که نوع تغییر برچسب‌ها به شکل درست و مناسبی صورت نگرفته و منجر شده تا نظم داده‌ها بهم ریخته و الگوی آن‌ها از بین برود و مدل نتواند آن‌ها را تفکیک کند.

دو مجموعه داده‌ای که نمونه با برچسب خنثی ندارد به طور کلی قابل مقایسه با نتایج مجموعه داده‌های دیگر نیست و به دلیل عدم وجود داده‌های خنثی در آن‌ها، طبیعتا دقت مدل اعداد بالاتری است.

۴ـ۶ـ مقایسه عملکرد مدل‌ها با برچسب به فرم regression

در جدول زیر گزارش دقت‌ها برای دو مدل 4-BiLSTM و Base-BERT برای برچسب به فرم regression که در بازه ی ۱- تا ۱ هستند، روی دو دیتاست undersample و changeLabel ارائه شده است. اعداد داخل جدول به ترتیب از چپ به راست خطای MAE برای داده های آموزش و خطای MAE برای داده های ارزیابی است.

حالت برچسب: فرم regression | لایه آخر مدل: ۱ نورون ساده | تابع فعال‌ساز لایه آخر: tanh | تابع ارزش: MeanAbsolueError | بهینه‌ساز: Adam

جدول (۴ـ۵) عملکرد مدل‌ها در برچسب با فرم regression

	Undersample	changeLabel
4-BiLSTM	0.028 / 0.14	0.028 / 0.13
Base-BERT+1dense	0.034 / 0.13	0.036 / 0.091

نتیجه: برای regression با داده‌های کم موجود در مجموعه داده Undersample، مدل Base-BERT نتوانسته آنچنان بهتر از مدل ساده 4-BiLSTM عمل کند اما با افزایش تعداد نمونه‌ها در مجموعه داده ChangeLabel با وجود اینکه ممکن است برچسب‌ها خراب شده باشند اما Base-BERT بهتر از مدل ساده عمل کرده است. در مجموع در حالت regression خیلی تفاوت بین مدل ساده و برت قابل مشاهده و لمس نیست چون از خطاها نمی‌توان تعداد نمونه‌های اشتباه تشخیص داده شده را فهمید. تنها نتیجه‌گیری این است که در حالت regression، با تعداد داده‌ی بیشتر که باعث می‌شود الگوها بیشتر شده و فرآیند آموزش سخت‌تر شود، برت می‌تواند نتیجه‌ی بهتری داشته باشد.

۴ـ۷ـ مقایسه مدل‌ها روی مجموعه داده crawledData

تا اینجا بررسی‌ها روی داده‌های آماده‌ی گرفته شده از اینترنت بوده است اما در این قسمت به سراغ مجموعه داده‌ی جمع‌آوری شده(crawledData) رفته و مدل‌های 2-BiLSTm، 4-BiLSTM، 10-BiLSTM، Base-BERT، FinBert، RoBERTa و XLNet روی آن تنظیم دقیق شده‌اند. مدل XLNet برای این قسمت اضافه شده است.

مجموعه داده: crawledData | حالت برچسب: فرم one-hot| تعداد ایپاک: ۲۰ | لایه آخر مدل: ۳ نورون ساده | تابع فعال‌ساز لایه آخر: softmax | تابع ارزش: CategoricalCrossEntropy | بهینه‌ساز: Adam

جدول (۴ـ۶) مقایسه دقت مدل‌ها پس از تنظیم دقیق روی مجموعه داده crawledData

مدل	دقت داده آموزش ٪	دقت داده ارزیابی ٪
2-BiLSTM	**84.4**	**69.4**
4-BiLSTM	**89.2**	**70.5**
10-BiLSTM	**92.1**	**70**
XLNet	**99.5**	**79**
Base-Bert	**99.7**	**79.1**
FinBert	**99.7**	**79.1**
RoBERTa	**98.8**	**82**

نتیجه: طبق جدول (۴ـ۶) می‌توان نتیجه گرفت که (۱) در حالت برچسب one-hot استفاده از مدل‌های LSTM با تعداد لایه بیشتر دقت داده‌های آموزش را افزایش می‌دهد اما در رابطه با دقت داده‌های ارزیابی می‌توان گفت که نسبتا مشابه هستند اما در نهایت طبق اعداد می‌توان گفت با تعداد لایه بیشتر دقت کاهش می‌یابد و نیاز به افزایش انچنانی ظرفیت مدل نیست و مدل ۴ لایه هم کارکر خوبی در داده‌های ارزیابی دارد. (۲) همچنین از بین مدل‌هایی که معماری آن‌ها همان معماری مشابه BERT است، مدل XLNet که ادعا کرده بود در ۲۰ تسک از برت بهتر است، نتیجه بهتری نداشته و نسبتا مشابه برت عمل کرده است که شاید یکی از دلایل آن تعداد کم داده‌هایی باشد که در این اینجا استفاده شده است. (۳) مدل FinBert

همچنان نتیجه بهتری نسبت به برت ندارد و مشابه آن عمل کرده است (۴) اما مدل RoBERTa توانسته به دقت ۸۲ درصد در داده‌های ارزیابی برسد و بهتر از برت عمل کند. یکی از تفاوت‌های مدل روبرتا این بود که با حجم داده‌ی بیشتری آموزش دیده و شاید بتوان به این نتیجه رسید که تعداد داده‌های مورد استفاده برای آموزش بسیار موثر است. در نهایت استفاده از مدل روبرتا توصیه می‌شود.

فصل ۵:

جمع‌بندی و پیشنهادها

۵ـ۱ـ جمع‌بندی

در زمینه‌ی تحلیل احساسات داده‌های مرتبط با رمزارزها در شبکه‌ی اجتماعی توییتر، تحقیقات کمی انجام شده و مجموعه داده‌های معتبری که توسط نیروی انسانی برچسب خورده باشد در این زمینه تهیه نشده است. در بیشتر تحقیقات برای برچسب‌زنی از کتابخانه VADER استفاده کرده‌اند و خروجی آن را به عنوان برچسب در نظر گرفته‌اند.

بیشتر مقالات مدل‌های معروف و نسبتا قدیمی یادگیری ماشین را بررسی کرده‌اند و کمتر سراغ مدل‌های پیشرفته‌ی جدید رفته‌اند. در این مقاله سعی شده روش‌های جدیدتر با یکدیگر مقایسه شوند و بهترین آن‌ها گزارش شود. مدل‌های تشکیل شده از سلول دو طرفه LSTM و مدل BERT و نسخه‌های بهبود یافته‌ی آن، از جمله این مدل‌ها هستند.

در این پایان‌نامه آزمایش‌های مختلفی انجام شده که در نتیجه‌ی آن‌ها، می‌توان به این موارد اشاره کرد:

- عدم استفاده از تصحیح غلط املایی در پیش‌پردازش به افزایش قدرت مدل کمک می‌کند.
- افزایش تعداد نمونه‌ها ممکن است منجر به افزایش الگوها و سخت شدن آموزش مدل و کاهش دقت مدل شود.
- استفاده از دسته‌بندهای مختلف در انتهای مدل BERT تفاوتی در دقت مدل ایجاد نمی‌کند.
- استفاده از مدل‌های با معماری مشابه BERT نسبت به مدل‌های ساده‌تر تشکیل شده از LSTM نتیجه‌ی بهتری دارد، اما مدل‌هایی مثل finBert که روی داده‌های مالی آموزش دیده‌اند الزاما از برت نتایج بهتری نداشته‌اند و نمی‌توان انتظار نتیجه بهتری داشت مگر اینکه روی داده‌های مالی مرتبط با زمینه رمزارز و همچنین منتشر شده در شبکه‌های اجتماعی، آموزش دیده باشد. همچنین از بین مدل‌های با معماری مشابه BERT مدل RoBERTa با اختلاف بهتر از مدل‌های دیگر است و دقت بالاتری دارد.
- در حالتی که لایه آخر ۱ نورون و به صورت regression باشد با داده‌های کم مدل برت نمی‌تواند بهتر از مدل ساده Bi-LSTM عمل کند و نسبتا مشابه هستند اما با افزایش داده‌ها

برت بهتر عمل خواهد کرد.

۵ـ۲ـ پیشنهادها

یکی از کارهای مهم در این زمینه جمع‌آوری یک مجموعه داده جامع و کامل که توسط انسان برچسب زده شده باشد، است. در ادامه‌ی این مسیر باید از نتایج تحلیل احساس متن با بهترین شرایط و مدل، استفاده شود و با ترکیب آن با داده‌های عددی مرتبط دیگر، به پیش‌بینی قیمت رمزارز پرداخته شود.

مراجع

مراجع

[۱] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, "Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning," *Entropy*, vol. 21, no. 6, p. 589, 2019. [Online]. Available: **https://www.mdpi.com/1099-4300/21/6/589**.

[۲] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216-225, 05/16 2014. [Online]. Available: **https//:ojs.aaai.org/index.php/ICWSM/article/view/14550**.

[۳] S. C. Nistor, M. Moca, D. Moldovan, D. B. Oprean, and R. L. Nistor, "Building a Twitter Sentiment Analysis System with Recurrent Neural Networks," *Sensors*, vol. 21, no. 7, p. 2266, 2021. [Online]. Available: **https://www.mdpi.com/1424-8220/21/7/2266**.

[٤] D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel, and B. K. Lama, "Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis," in *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, 25-27 Oct. 2018 2018, pp. 128-132, doi: 10.1109/CCCS.2018.8586824 .

[٥] N. Aslam, F. Rustam, E. Lee, P. B. Washington, and I. Ashraf, "Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model," *IEEE Access*, vol. 10, pp. 39313-39324, 2022, doi: 10.1109/ACCESS.2022.3165621.

[٦] Twitter. "Search Tweets: Standard v1.1." **https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview** (accessed September 12.(۲۰۲۲ ,

[۷] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[۸] D. Araci, *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv.

[۹] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*, vol. abs/1907.11692, 2019.

[۱۰] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*: Curran Associates Inc., 2019, p. Article 517.

Abstract:

The field of cryptocurrency (especially Bitcoin, which is one of the most popular ones) has recently received a great deal of attention from people and financial market experts, and many efforts are being made to create financial profit in this field. Social networks are also a suitable platform for people to express their opinions, and a large amount of information is published by different strata daily, and Twitter is one of the most used social networks. On the other hand, a well-known task in artificial intelligence is the analysis of the author's sentiments from the text. Analyzing the sentiments of social network users towards Bitcoin is one of the research fields that can lead to predicting the trend of Bitcoin price changes. In this thesis, firstly, the method of collecting data related to Bitcoin on Twitter has been discussed and a dataset has been produced with it. In the following, the previously published datasets have been reviewed and the available methods for creating appropriate labels for the process of data sentiment analysis have been proposed. Most of the articles published in this field have used machine learning methods such as SVM or Random-Forest, and deep learning methods and complex and powerful models have been used less. The use of simple models consisting of LSTM cells, more up-to-date language models such as BERT, improved BERT models such as RoBERTa and XLNet, as well as versions trained on financial texts, such as FinBert, and comparing the results of all these models, is the main focus of the activities carried out in this thesis. After the experiments on the results of training the models with the prepared dataset, using the RoBERTa model gives better results. In the end, to continue the work, it is recommended to prepare a larger dataset with more valid labels in the field of cryptocurrency, which have been verified by humans if possible, so that more accurate and reliable models can be provided by training them.

Keywords: Sentiment Analysis – Cryptocurrency – BERT – Twitter – Deep Learning

Iran University of Science and Technology
Computer engineering Department

Bitcoin-related Sentiment Analysis of Twitter

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Undergraduate in Computer Engineering Field

By:
Seyed Mostafa Masoudi

Supervisor:
Dr. Sauleh Etemadi
Dr. Adel Rahmani

October 2022