دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

گزارش کار پروژه نهایی درس یادگیری ماشین

پیش بینی تأثیر پیام‌های توییتری مرتبط با رمز ارزها روی قیمت bitcoin

محمد لشکری، امیر صادقی، ایمان ملکیان

زمستان ۱۴۰۰

۱ مقدمه

همانطور که می‌دانیم در دنیای امروزی، رمز ارزها و مبادلاتی که در این حوزه انجام می‌شود از اهمیت زیادی برخوردار است. به گونه ای که اکثریت افراد قصد سرمایه گذاری یا استخراج رمز ارزها را دارند. یکی از موضوعاتی که تأثیر به سزایی در این رابطه می‌گذارد، توییت‌های افراد صاحب سرمایه در رابطه با رمز ارزهاست. توییت های این افراد عموما تأثیرات مثبت یا منفی زیادی در نوسانات قیمت رمز ارزها می‌گذارد [۱]. ما در این پروژه تأثیر توییت‌های افراد مختلف را روی قیمت bitcoin بعد از گذشت ۲۴ ساعت از زمان ارسال پیام بررسی می‌کنیم. یکی از روش‌هایی که می‌توان با استفاده از مسئله را حل کرد استفاده از BERT و روش دیگر مبتنی بر یادگیری انتقالی [۱] است که مدل آن نیز BERT است [۲] [۳]. ما در این پروژه، از دو روش بردارهای پشتیبان [۲] و بیز ساده [۳] استفاده کرده و آنها را با مدل از پیش آموزش دیده BERT مقایسه می‌کنیم. علاوه بر این ما با استفاده از یادگیری انتقالی مدل BERT را روی داده‌های خود تنظیم [۴] کرده و نتایج را مقایسه می‌کنیم. از چالش‌هایی که در این راه وجود دارد می‌توان به طول جملات، یافتن حداکثر تعداد ویژگی‌های مؤثر و محدودیت سخت افزاری به دلیل بزرگ بودن ابعاد دادگان اشاره کرد.

۲ کارهای پیشین

یکی از روش‌های ارائه شده برای حل این مسئله استفاده از BERT از پیش آموزش دیده، بردارهای پشتیبان و بیز ساده با توزیع برنولی است [۴]. ما از مدل بیز گاوسی، بردارهای پشتیبان(تنظیم هایپرپارامترها به کمک جستجوی شبکه‌ای [۵]) و مدل تنظیم شده BERT روی داده‌های مسئله استفاده می‌کنیم. چون مدل از پیش آموزش دیده BERT با اطلاعات جملات مربوط به دنیای رمزارزها آشنا نیست ما BERT را روی دادگان خود تنظیم و نتایج را با دیگر مدل‌ها مقایسه می‌کنیم.

۳ دادگان

۱.۳ ساخت دادگان از روی مجموعه داده‌های موجود

برای این پروژه از دادگان موجود در سایت Kaggle که شامل پیام‌های توییتری سال ۲۰۲۱ در ماه‌های ۷ و ۸ و ۹ میلادی با برچسب مرتبط با بیت کوین بودند استفاده کرده‌ایم [۵]. برای برچسب گذاری هر توییت برحسب میزان تأثیر آن روی قیمت بیت کوین از داده‌های میزان تغییر قیمت بیت کوین که در هر دقیقه تهیه شده اند استفاده کرده‌ایم [۶]. این مجموعه داده شامل ۱۲ ویژگی است. ما فقط از دو ویژگی open time, close استفاده کرده‌ایم. ستون open time نشان‌دهنده ابتدای دقیقه شروع بازه است و close بیانگر قیمت بیت کوین در انتهای بازه است (بازه ها دقیقه هستند.). فرض کنیم پیام t در دقیقه m پست شده باشد و P_m

[1] transfer learning
[2] Support vector machine
[3] Naïve Bayes
[4] fine-tune
[5] Grid Search

نشان‌دهنده قیمت بیت کوین در انتهای دقیقه m باشد. میزان تأثیر یک پیام روی قیمت بیت کوین در ۲۴ ساعت آینده به صورت زیر محاسبه می‌شود:

$$\Delta P(t) = \frac{P_{m+۱۴۴۰} - P_m}{P_m}$$

اگر $\Delta P(t) > ۰/۰۱$ یعنی حداقل یک درصد تأثیر مثبت داشته است و ما برای آن برچسب ۱ یعنی کلاس مثبت را در نظر گرفته‌ایم. اگر $\Delta P(t) < -۰/۰۱$ برچسب آن را -۱ یعنی کلاس منفی قرار داده‌ایم. اگر هیچ یک از دو شرط قبل برقرار نبود، داده برچسب ۰ یعنی خنثی گرفته است.

۲.۳ پیش پردازش داده‌ها

ما ایموجی‌ها، حروف اضافه و کاراکترهای اضافی را از جملات حذف کرده‌ایم. کلمات ت‌ءقف مانند not,in,or را حذف کردیم چون ویژگی خاصی ندارند و به مدل اطلاعات مفیدی را از آن استخراج نمی‌کند. فعل‌ها را ساده‌سازی [۶] کردیم مثلاً کلمه removed به remove تبدیل شده است. نمونه‌ای از دادگان پیش پردازش شده را می‌توانید در جدول ۱ مشاهده کنید.

id	sentence
7334	complet normal bitcoin http co lu bjqmeat
7472	night hour best time receiv cosmo coin atom coin walletx ...
8365	good luck athlet gear compet respect sport openingceremoni ...
8453	mysquarefin thank share wonder opportun hope project better ...
9560	btc updat rang set matter break first read chart full note ...

جدول ۱: نمونه‌هایی از دادگان پیش پردازش شده

۴ مدل‌سازی

۱.۴ بردارهای پشتیبان

این الگوریتم که آن را به اختصار SVM می‌نامیم بر پایه طبقه بندی دودویی است و خروجی آن یک مدل متمایزکننده است کوتاه ترین فاصله بین مدل و هر دو کلاس را بیشینه می‌کنیم. اگر نزدیک ترین نقاط از هر دو کلاس به مدل را بردارهای پشتیبان و w را پارامتر مدل در نظر بگیریم، هدف این الگوریتم، یافتن پاسخ برای مسئله زیر است:

$$\begin{aligned} \min \quad & \frac{۱}{۲}\|w\|_۲^۲ \\ \text{s.t.} \quad & y_i(w^T \mathrm{x}_i + b) \geq ۱ \end{aligned}$$

[6] stemming

که در آن به $\forall i \;\; y_i \in \{-۱, +۱\}$.

۲.۴ بیز ساده

طبقه بندی کننده‌های Naive Bayes مجموعه ای از الگوریتم های طبقه بندی بر اساس تئوری Bayes هستند که هر جفت ویژگی طبقه بندی شده در الگوریتم‌های بیز مستقل از یکدیگر است. ما از بیز ساده گاوسی [۷] که به اختصار آن را GNB می‌نامیم استفاده کرده‌ایم. قضیه‌ی بیز، احتمال رخ دادن یک پیشامد را هنگامی که پیشامد دیگر اتفاق افتاده باشد بدست می‌آورد. همانطور که در معادله ی زیر مشاهده می کنید، با استفاده از تئوری بیز، خواهیم توانست $P(C = c_j|X = \mathrm{x})$ را بدست آوریم.

$$P(C = c|X = \mathrm{x}) \propto P(X = \mathrm{x}|C = c)P(C = c)$$

که درآن $X = \mathrm{x}|C = c$ توزیع نرمال با میانگین μ_c و انحراف معیار $\sigma_c^۲$ است. مقدار فوق به ازای تمام کلاس‌ها محاسبه و کلاسی که بیشترین مقدار را داشته باشد برچسب نمونه x خواهد شد.

۳.۴ الگوریتم BERT

Bert یک مدل از پیش آموزش داده شده توسط گوگل است که سال ۲۰۱۸ منتشر شده است. Bert در جقیقت مخفف Transformers from Representations Encoder Bidirectional است. این مدل ها از نوع Transformers Attention-based هستند [۲]. معماری این دسته از مدل ها در شکل ۱ قابل مشاهده است [۷].

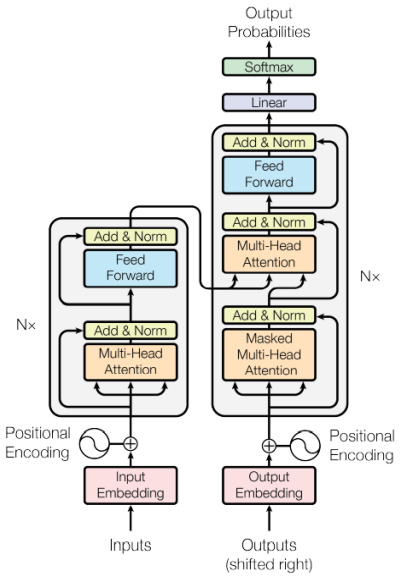
ما از مدل های BERT میتوانیم برای یادگیری انتقالی استفاده کنیم، برای هر دو دیتای با نظارت و بی نظارت. در بخش پیش آموزش این دسته از مدل ها از منابعی مثل BooksCorpus (شامل ۸۰۰ میلیون کلمه) و Wikipedia (شامل ۲۵۰۰ میلیون کلمه) استفاده می‌شود [۲].

۵ آزمایش و نتیجه‌گیری

ما به دلیل محدودیت‌های سخت‌افزاری فقط ده هزار نمونه به صورت تصادفی به عنوان دادگان آموزشی انتخاب کردیم.ما ابتدا برچسب‌های $-۱, ۰, ۱$ را به ترتیب با $۰, ۱, ۲$ جایگزین کرده‌ایم. چون مدل برچسب منفی نمی‌پذیرد. سپس برای مدل‌های SVM و GNB با استفاده از الگوریتم کیف کلمات [۸] با تعداد ۱۰۰۰ ویژگی، جملات را برداری کرده‌ایم. ما BERT را روی داده‌های مورد تحقیق خودمان تنظیم کردیم و خروجی گرفتیم. منظور از مدل BERT در اینجا نیز همین است. نتایج مدل‌ها را می‌توانید در جدول ۲ مشاهده کنید:

برای SVM با استفاده از سرچ شبکه‌ای برای کرنل `rbf` و `C=1` به دست آمد و دقت اعتبار سنجی آن ۰/۵۵ شد که دقیقاً همان دقت تست است. دقت اعتبار سنجی GNB برابر ۰/۳۳ به دست آمد. اعتبار سنجی به کمک Cross Validation با ۱۰ فلد انجام شده است. صحت دو مدل BERT, GNB برای کلاس

[7] Gaussian Naive Bayes
[8] Bag of Words



شکل ۱: معماری مدل‌های transformers

خنثی بسیار کم است که علت آن می‌تواند نامتوازن بودن دادگان باشد. مدل BERT هیچ متنی را در کلاس خنثی پیش‌بینی نکرد که دلیل آن نزدیک بودن توییت‌های کلاس صفر به کلاس یک است. بنابراین ما برای BERT از کلاس خنثی صرف نظر کرده و ۲۰۰۰۰ داده متوازن از دادگان با برچسب‌های مثبت و منفی برای آموزش این مدل، نمونه گرفتیم که نتایج آن در ۳ قابل مشاهده است که صحت و f1-score در آن میانگین ماکرو دو کلاس مثبت و منفی هستند.

مدل	دقت تست	f1-score	صحت تست		
			کلاس منفی	کلاس خنثی	کلاس مثبت
SVM	۰/۵۵	۰/۳۱	۰/۶۵	۰/۷۱	۰/۵۴
BERT	۰/۵۳	۰/۲۷	۰/۵۶	۰/۰۰	۰/۵۳
GNB	۰/۳۵	۰/۳۵	۰/۳۰	۰/۲۷	۰/۵۹

جدول ۲: نتایج مدل‌سازی

دقت	صحت	f1-score
۰/۶۰	۰/۶۰	۰/۶۰

جدول ۳: نتایج BERT برای حال دو کلاسه

دلیل کم بدون دقت BERT می‌تواند غلط‌های املایی، طولانی‌بودن جملات و نامفهوم بودن توییت‌ها از لحاظ گرامری باشد که مدل را به اشتباه انداخته است. BERT در ساختار خود از embedding استفاده کرده و این اشتباهات نگارشی تأثیر مستقیمی روی عملکرد آن خواهد داشت. نتیجه نهایی پروژه این است که از روی تحلیل احساسات توییت‌های ویرایش نشده از نظر املا و دستور زبان نمی‌توان قیمت بیت کوین را پیش بینی کرد.

۶ نحوه همکاری

ساخت دادگان توسط محمد لشکری انجام شد. پیش پردازش دادگان توسط امیر صادقی انجام شد. SVM توسط محمد لشکری، BERT و تنظیم آن روی دادگان توسط امیر صادقی و GNB توسط ایمان ملکیان انجام شد. ساخت فایل گزارش نهایی و ویرایش توسط محمد لشکری انجام شد. هر یک از اعضا در تهیه محتوای کار خود برای گزارش کار نقش داشتند.

مراجع

[1] Matta, M., Lunesu, I., & Marchesi, M. (2015, June). Bitcoin Spread Prediction Using Social and Web Search Media. In UMAP workshops (pp. 1-10).

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] Bao, X., & Qiao, Q. (2019, August). Transfer learning from pre-trained bert for pronoun resolution. In Proceedings of the first workshop on gender bias in natural language processing (pp. 82-88).

[4] Abdali, S., & Hoskins, B. (2021). Twitter Sentiment Analysis for Bitcoin Price Prediction. Stanford University.

[5] K. Suresh, "Bitcoin Tweets, Version 23," Kaggle, 30-11-2021. [Online]. Available: https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets. [Accessed: 11-Dec2021].

[6] Monthly Klines, BTCUSDT, lm in Binance Market Data. [Online]. https://data.binance.vision/?prefix=data/spot/monthly/klines/BTCUSDT/1m/. [Accessed: 11-Dec2021]

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).