به نام خدا

مباحثی در علوم کامپیوتر
گزارش پروژه
۷ بهمن ۹۹

استاد درس: دکتر اکبری
دانشگاه صنعتی امیرکبیر
دانشکده‌ی ریاضی و علوم کامپیوتر

پیش‌بینی قیمت سهام از شبکه‌های اجتماعی

چکیده

در این مقاله، یک روش برای پیش‌بینی روند کلی ارزش بیت‌کوین[1] ارائه شده و سپس خروجی آن، روی یک اپلیکیشن موبایل پیاده شده است. در حال حاضر، بیت‌کوین بزرگ‌ترین رمزارز[2] از لحاظ سهم از بازار است. با این حال، قیمت آن در یک بازه‌ی روزانه یا حتی بلند مدت، بسیار پرنوسان است.

توییتر یک شبکه‌ی مجازی بسیار محبوب است که گاهی به عنوان یک منبع خبری تاثیرگذار برای تصمیم‌گیری در خرید ارزهای دیجیتال مورد استفاده قرار می‌گیرد. تحلیل توییت‌ها و فهمیدن نظر عموم مردم درباره‌ی روند قیمت بیت‌کوین، می‌تواند کمک شایانی به تصمیم‌گیری در مورد مبادله‌ی آن باشد.

پس از جمع‌آوری توییت‌های مرتبط با بیت‌کوین و قیمت بیت‌کوین در بازه‌ی زمانی توییت‌ها، این نتیجه حاصل شد که به دلیل تعدد متغیرها، استفاده از رگرسیون[3] راه درستی نیست و درنهایت با استفاده از مدل‌های Classification متفاوت و مقایسه‌ی خروجی هر مدل با ارزیابی، به نتیجه‌گیری‌هایی که به آن خواهیم پرداخت، دست یافتیم.

۱ مقدمه

امروزه عده‌ی زیادی بیت‌کوین را به عنوان طلای دیجیتال می‌شناسند. این شناسایی دلیل خوبی هم دارد. این رمزارز حجم معاملات بسیار بالایی داشته و با توجه به تعداد محدود آن و تقاضای روزافزون مردم، به راحتی می‌توان به تشابه آن با طلا پی برد.

اما با تمام این‌ها، نوسان شدید این رمزارز باعث شده تا پیش‌بینی زمان مناسب برای خرید یا فروش آن، بسیار دشوار شود.

میان فاکتورهای تاثیرگذار روی بیت‌کوین، رسانه تاثیر به سزایی روی قیمت آن می‌گذارد. به کمک شبکه‌های مجازی، هر خبر می‌تواند به سرعت پخش شده و قیمت بیت‌کوین را تحت تاثیر قرار دهد. همچنین بسیاری از پیش‌بینی کننده‌های بیت‌کوین، نظر عمومی حال حاضر مردم را در پیش‌بینی خود، در نظر نمی‌گیرند. بنابراین ما بر آن شدیم تا روشی برای پیش‌بینی روند کلی ارزش بیت‌کوین ارائه کنیم.

---

[1] Bitcoin

[2] Crypto Currency

[3] Regression

به این منظور، ابتدا توییت‌های مربوط به بیت‌کوین به همراه قیمت بیت‌کوین در زمان توییت‌های جمع‌آوری شده، به روشی که در بخش مجموعه داده‌ها به تفصیل بیان شده است، جمع‌آوری و اطلاعات لازم از آن استخراج و پیش‌پردازش شده است.

همچنین برای ارتباط میان احساسات[۴] توییت و جهت حرکت قیمت بیت‌کوین، از مدل‌های Classifier استفاده شده که اطلاعات بیشتر در این باره، در بخش مدل قابل مشاهده است.

۲ مجموعه داده

۲.۱ توییت‌ها

توییت‌ها در دو دسته جمع‌آوری شدند؛ توییت‌های قدیمی و جدید.

توییت‌های قدیمی مربوط به سال ۲۰۱۷ هستند که از مجموعه داده‌ی موجود در لینک https://github.com/teoYQ/Bitcoin-Twitter-sentiment-analysis/tree/master/2019%20Tweets%20Final جمع‌آوری شده و توییت‌های جدید، مربوط سال ۲۰۲۱ با استفاده از API توییتر استخراج شده‌اند.

معیار انتخاب توییت‌ها در ابتدا، تنها وجود کلمه‌ی "bitcoin" و انواع مخلف نوشتاری آن مانند "BTC" در متن و برچسب‌های آن بود. (bitcoin OR Bitcoin OR BITCOIN OR btc OR Btc OR BTC) پس از استخراج تعدادی از توییت‌ها، متوجه شدیم ریتوییت‌ها[۵] به عنوان یک توییت جدا به دیتاست وارد شده و موجب بروز اختلال در مدل‌ها می‌شدند که با اعمال محدودیت ریتوییت نشده بودن، این مشکل برطرف شد. (-is:retweet) همچنین تنها توییت‌های به زبان انگلیسی جمع‌آوری شدند. (lang:en)

از جمله اطلاعاتی که از این توییت‌ها استخراج شد نیز، می‌توان به زمان توییت، تعداد علاقه‌مندی[۶]، ریتوییت، پاسخ[۷] و نقل قول[۸] اشاره کرد. (tweet.fields=public_metrics,created_at)

https://api.twitter.com/2/tweets/search/recent?query=(bitcoin OR Bitcoin OR BITCOIN OR btc OR Btc OR BTC) -is:retweet lang:en&tweet.fields=public_metrics,created_at

---

[4] Sentiment

[5] Retweet

[6] Favorite

[7] Reply

[8] Quote

Price
Market
New
Bitcoin
Crypto
Buy
Cryptocurrency
Blockchain
Exchange
sell
money
Cash
Top
now
time
via
Wallet
Mining
Ethereum
XRP
ETH
LTC
BCH
USD
Twitter
Analysis
Binance
trade
coin
network
Payment
Bitcoins
trading
day
last
say
future
see
change
year
hour
make
hold
value
think
today
price
dont
world
one
go
use
long
high
low
project
start
Bull
Bear
update
Rage
Pump
Dump
Hard
Fork
Wallet
Mining
Analysis
Payment
Twitter
Binance
Exchange
Blockchain
Bitcoins
Ratio
Bank
Bitcoinback
current
LTC
BCH
USD
going
worth
Join
Bull
Bear

۳ مدل

در ابتدا تلاش شد تا مسئله، به روش رگرسیون حل شود. به همین منظور، یک رگرسیون چندجمله‌ای[۱۱] نوشته شد اما جواب‌های قابل قبولی از آن به دست نیامد و قابل استفاده نبود.

سپس تلاش شد تا مسئله به صورت طبقه‌بندی[۱۲] حل شود. به همین منظور از classifierهای زیر با نتایج قابل مشاهده، استفاده شد.

۳.۱ مدل kNN

برای تنظیم ابرپارامتر k، با استفاده از روش بازو[۱۳]، عدد بهینه، ۴ به دست آمد.

```
In [12]: knn_()
1 [[ 819 1011]
 [ 613 1257]]
2 [[ 819 1011]
 [ 613 1257]]
3 [[ 819 1011]
 [ 613 1257]]
4 [[ 819 1011]
 [ 613 1257]]
5 [[ 819 1011]
 [ 613 1257]]
6 [[ 819 1011]
 [ 613 1257]]
7 [[ 819 1011]
 [ 613 1257]]
8 [[ 819 1011]
 [ 613 1257]]
9 [[ 819 1011]
 [ 613 1257]]
10 [[ 819 1011]
 [ 613 1257]]
11 [[ 819 1011]
 [ 613 1257]]
12 [[ 819 1011]
 [ 613 1257]]
```

ماتریس سردرگمی به ازای kهای متفاوت

نتایج به دست آمده:

Accuracy: 56.1%

F1: 0.49

Precision: 0.57

Recall: 0.44

۳.۲ مدل Naïve Bayes

با استفاده از روش گاوسی نتایج زیر به دست آمد:

Accuracy: 52.4%

F1: 0.60

Precision: 0.75

Recall: 0.51

---

[11] Polynomial Regression

[12] Classification

[13] Elbow

۳.۳ مدل Random Forest

در این روش، ابرپارامتر، یعنی تعداد درخت‌های مورد استفاده، برای اعداد ۱ تا ۱۰۰ آزموده و در نهایت مقدار بهینه ۱۱ به دست آمد و نتایج زیر از آن کسب شد:

Accuracy: 59%	F1: 0.66
Precision: 0.78	Recall: 0.56

۳.۴ مدل Decision Tree

نتایج زیر به دست آمد:

Accuracy: 59.5%	F1: 0.65
Precision: 0.8	Recall: 0.56

۴ خروجی نهایی

در نهایت مدل به دست آمده روی یک سرور قرار گرفت و در یک اپلیکیشن موبایل، با وارد کردن متن یک توییت، با استفاده از API ساخته شده، نتیجه‌ی پیش‌بینی مدل، مبنی بر بالا یا پایین‌رونده بودن قیمت بیت‌کوین، به کاربر نشان داده می‌شود.

همچنین این اپلیکیشن، از قابلیت‌های دیگری همچون اعلام قیمت لحظه‌ای بیت‌کوین، نمایش نمودار قیمت بیت‌کوین و آخرین توییت‌های مرتبط با بیت‌کوین نیز برخوردار است.

در آینده می‌توان، قابلیت پیش‌بینی قیمت با توجه به آخرین توییت‌های موجود در توییتر را نیز به این اپلیکیشن اضافه کرد.

۵ نتیجه

یکی از چالش‌های اصلی، جمع‌آوری مجموعه داده‌ی مناسب بود که با توجه به آن‌چه در بخش مجموعه داده توضیح داده شد، مجموعه داده‌ی مناسب برای آموزش مدل‌ها، جمع‌آوری و سپس پردازش شد.

برای جمع‌آوری توییت‌ها از API توییتر، key نیاز بود که برای به دست آوردن آن، یک فرم پر شد که در آن مواردی همچون هدف استفاده از این API و موارد دیگر ذکر شد که پس از چند روز، این دسترسی داده شد.

به منظور بهبود یادگیری شبکه، بازه‌های زمانی‌ای که بیت‌کوین هم دارای روند نزولی و هم دارای روند صعودی باشد نیاز بود که با تحلیل بازه‌های زمانی متفاوت، یک بازه‌ی مناسب یافت شد و مجموعه داده‌ی همان بازه، مورد استفاده قرار گرفت.

همچنین در این پروژه، زمان زیادی صرف آزمایش مدل‌های مختلف شدو با این حال، در بهترین مدل، دقت ۶۰٪ به دست آمد که با توجه به زمان مجدود بهینه‌سازی مدل‌ها، بهتر از چیزی بود که گمان می‌رفت.

۴ اعضای گروه

- امیرحسین امامی ۹۷۱۲۰۰۲
- سارا بابایی ۹۷۱۳۰۰۵
- ارشیا مجیدی ۹۷۱۳۰۶۱