Bitcoin vs. Twitter — Zusammenhänge des Bitcoin-Kurses und zeitgleicher Tweets

Alicia Hirt,* Tristan Kreuziger,* and Jan-Philipp Praetorius*

Faculty of Mathematics and Computer Science, Friedrich-Schiller-Universitt Jena, Germany

E-mail: alicia.hirt@uni-jena.de; tristan.kreuziger@uni-jena.de; jan-philipp.mueller@uni-jena.de

Einführung

Neben etablierten Wertpapieren und Indizes, lassen sich seit einiger Zeit auch digitale Zahlungsmittel an offiziellen Marktplätzen und Börsen handeln. Während Währungen wie der Doller einen relativ gut messbaren Wert besitzen und dieser Wert durch aktive Maßnahmen reguliert und beeinflusst werden kann, sind Kryptowährungen weitestgehend unabhängig. Dabei haben digitale Währungen wie der Bitcoin eine extreme Volatilität. Im Rahmen dieser Arbeit soll ermitteln werden, ob es einen Zusammenhang zwischen dem Kursverlauf des Bitcoin und der Meinung von Twitter-Nutzern über den Bitcoin gibt. Dazu wird eine Sentiment-Analyse von Tweets durchgeführt.

Daten

Um einen Zusammenhang zwischen dem Bitcoin und der Meinung von Twitter–Nutzern ermitteln zu können, werden Daten aus dem gleichen Zeitintervall benötigt. Der Kursverlauf des Bitcoin lässt sich unkompliziert über mehrere Quellen beschaffen. Insbesondere Internetseiten wie *www.yahoo.com* bieten gute Möglichkeiten sekundengenaue Informationen über den Kursverlauf zu erhalten. Für diese Arbeit wurde der Kurs im Minutentakt vom 10. Januar 2018 bis zum 2. Februar 2018 von der Seite *www.coindesk.com* extrahiert. In welchem Format die Bitcoin–Kurse vorliegen ist in Abbildung 1 zu sehen.

	A	B	C
1	Date	Open Price	Close Price
2	2018-01-10 00:00	14073,16	14073,16
3	2018-01-10 01:00	14073,16	14193,43
4	2018-01-10 02:00	14193,43	14290,4
5	2018-01-10 03:00	14290,4	14139,18
6	2018-01-10 04:00	14139,18	14374,32
7	2018-01-10 05:00	14374,32	14182,03
8	2018-01-10 06:00	14182,03	14240,25
9	2018-01-10 07:00	14240,25	14115,24
10	2018-01-10 08:00	14115,24	13500,39
11	2018-01-10 09:00	13500,39	13790,67
12	2018-01-10 10:00	13790,67	13862,86
13	2018-01-10 11:00	13862,86	14073,29
14	2018-01-10 12:00	14073,29	13898,03
15	2018-01-10 13:00	13898,03	13907,03
16	2018-01-10 14:00	13907,03	13778,42
17	2018-01-10 15:00	13778,42	14420,99
18	2018-01-10 16:00	14420,99	14498,19

Abbildung 1: Bitcoin-Kursverlauf

Den Bitcoin–Kursen werden im Rahmen dieser Arbeit Tweets gegenübergestellt. Ein Tweet hat maximal 280 Zeichen. In Abbildung 2 ist ein Beispiel fr einen solchen Tweet zu sehen.

Abbildung 2: Beispiel-Tweet

Um eine repräsentative Stichprobe zu haben, wurden knapp 120.000 englischsprachige

Tweets untersucht. Neben dem Text eines Tweets, werden außerdem Informationen über denjenigen ausgelesen, der den Tweet verfasst hat, wann diese Kurznachricht verfasst wurde, eine eindeutige Identifikationsnummer, retweets und jede Menge andere Informationen. Für diese Arbeit wurde jedoch nur das Erstellungsdatum und der Text eines Tweets gesammelt. Allgemein wurden nur Nachrichten aufgezeichnet, in denen die Zeichen: *Bitcoin*, *#Bitcoin* und *cryptocurrency* vorkommen. Auerdem sind nur Tweets erfasst worden, die innerhalb der USA und einem Großteil von Kanada veröffentlicht wurden. Da die Sentiment-Analyse am effizientesten funktioniert, wenn alle zu untersuchenden Texte in einer Sprache sind, wurde mit Hilfe der Bibliothek „TextBlob[1] “ für jeden einzelnen Tweets die entsprechende Sprache ermittelt. Alle Nachrichten, die nicht in der Sprache Englisch waren, wurden daraufhin aus dem Datensatz entfernt.

Umsetzung in Spark

Der Arbeitsfluss in Spark (siehe Code 1) besteht grob aus vier Phasen.

Die bereinigten Daten werden aus einer Textdatei direkt von Spark eingelesen und liegen dann als RDD vor. Danach folgt im ersten Schritt eine Vorverarbeitung, bei der das Datum von einem Text in ein Objekt konvertiert wird, damit der Umgang damit später einfacher ist. Hier werden auch Fälle herausgefiltert, bei denen die Konvertierung fehlgeschlagen ist. In einem solchen Fall hat das Mapping einen ungültigen Wert zurückgeliefert, der bekannt ist und dann explizit entfernt werden kann.

Dann folgt der Schritt, in dem der Text des Tweets analysiert wird und die heuristisch bestimmte Stimmung zurückgegeben wird. Die Typen der Tupel ändern sich hier von (Datum, Text) zu (Datum, Float). Die Stimmungswerte liegen zwischen -1 und 1, sodass bei Fehlern in diesem Mapping-Schritt einfach ein Wert zurückgegeben werden kann, der deutlich aus dem Rahmen fällt. Dieser wird dann wie zuvor beim Datum herausgefiltert.

Im Folgenden folgt ein Mappen der Datumsangaben und Stimmungen auf diskretere

[1]Siehe http://textblob.readthedocs.io/en/dev/ .

Punkte. Die Zeitstempel der Tweets liegen sekundengenau vor, was für die Analyse nicht hilfreich ist. Das Ziel ist es alle Stimmungen in einem gewissen Datumsbereich, z.B. an Tag X von 18 bis 20 Uhr zusammenzufassen. Dann lassen sich die Stimmungen in diesen Zeitintervallen mit Veränderungen im Bitcoin-Kurs über den gleichen Zeitraum vergleichen. Je größer die Anzahl an - einigermaßen zeitlich gleichverteilten Tweets - ließen sich diese Intervalle immer kleiner machen, sodass eine punktuelle Betrachtung möglich wird.

In der Praxis kommen Stimmungen auf dem ganzen Spektrum [-1,1] vor, was wie auch bei den Zeitangaben in diesem Szenario nicht hilfreich ist. Folglich werden zu Beginn des Programmablaufs Intervalle angegeben, auf die die Stimmungen abgebildet werden. Im später folgenden Beispiel sind das [-1.0, -0.1], [-0.1, 0.1], [0.1, 1.0]. Werte werden jetzt auf diese Intervalle diskret abgebildet und nur noch der Index des Intervalls gespeichert. Die geringe Größe des Intervalls um 0 herum, also neutrale Stimmungen liegt daran, dass die Analyse der Stimmungen überproportional stark zur Einteilung neutral tendiert, was hiermit etwas abgeschwächt wird.

Im letzten Schritt wird für jedes Stimmungsintervall gezählt, wie oft die Stimmung an jedem Zeitpunkt vorgekommen ist. Das finale Ergebnis besteht dann aus dem Zeitstempel und der Liste der Häufigkeiten für jedes Intervall.

Code 1: Spark Code

```
data_lines = sc.textFile(txt)


prepared = data_lines.map(lambda x: split_data_line(x))
converted = prepared.map(lambda x: convert_date(x))
    .filter(lambda x: x[0].year != 1990)
sentiments = converted.map(lambda x: check_sentiment(x))
    .filter(lambda x: x[1] != 666)
sentiments = sentiments.map(lambda x: cluster_sentiment(x, levels))
```

```
sentiments = sentiments.map(lambda x: cluster_date(x, levels))


for i in range(len(sent_levels)):
    temp = s.filter(lambda x: x[1] == i)
            .groupByKey()
            .map(lambda x: (x[0], (i, len(list(x[1])) - 1)))
    final = final.union(temp)
```

Auswertung

Die Meinung von Twitternutzern ist dem Bitcoin–Kurs in Abbildung 3 gegenübergestellt. Die verschiedenen Tweets wurden mit Hilfe der Sentimentanalyse in negative, neutrale und positive Nachrichten unterteilt und in den Säulen farblich gekennzeichnet. Die Höhe der Säulen sentspricht der Anzahl an Nachrichten, die innerhalb eines Zeitintervalls von einer Stunde registriert wurden. Hohe Säulen weisen somit auf viele Nachrichten hin, whrend flache Säulen für wenig Twitter-Meldungen zum Thema Bitcoin stehen. Zusätzlich ist der Trend des Bitcoin-Kurses aufgetragen. Der jeweilige Wert wurde als Differenz aus Opening- und Closing-Betrag, wie in Abbildung 1 zu sehen, bestimmt. Ein positiver Wert entspricht somit einem steigenden Kursverlauf, ein negativer Wert entsprechend einem fallenden.

Es ist zu beachten, dass die dargestellte Zeitachse nicht kontinuierlich verluft. Viel mehr wurden pro Datum nur Nachrichten aus wenigen Stunden extrahiert. Dieses Verhalten ist auf die Bibliothek „tweepy[2]“ zurückzuführen, mit deren Hilfe die Tweets heruntergeladen wurde. Auf Grund der limitierten Downloadrate bei Twitter und den Einstellungen der Bibliothek wurden Anstelle des gewünschten Zeitraums vom 10.01.2018 – 01.02.2018 lediglich Daten weniger, kurzer Zeitabschnitte heruntergeladen.
Aus den vorhandenen Daten lassen sich trotzdem einige Ergebnisse ableiten.

[2]Siehe http://www.tweepy.org/ .

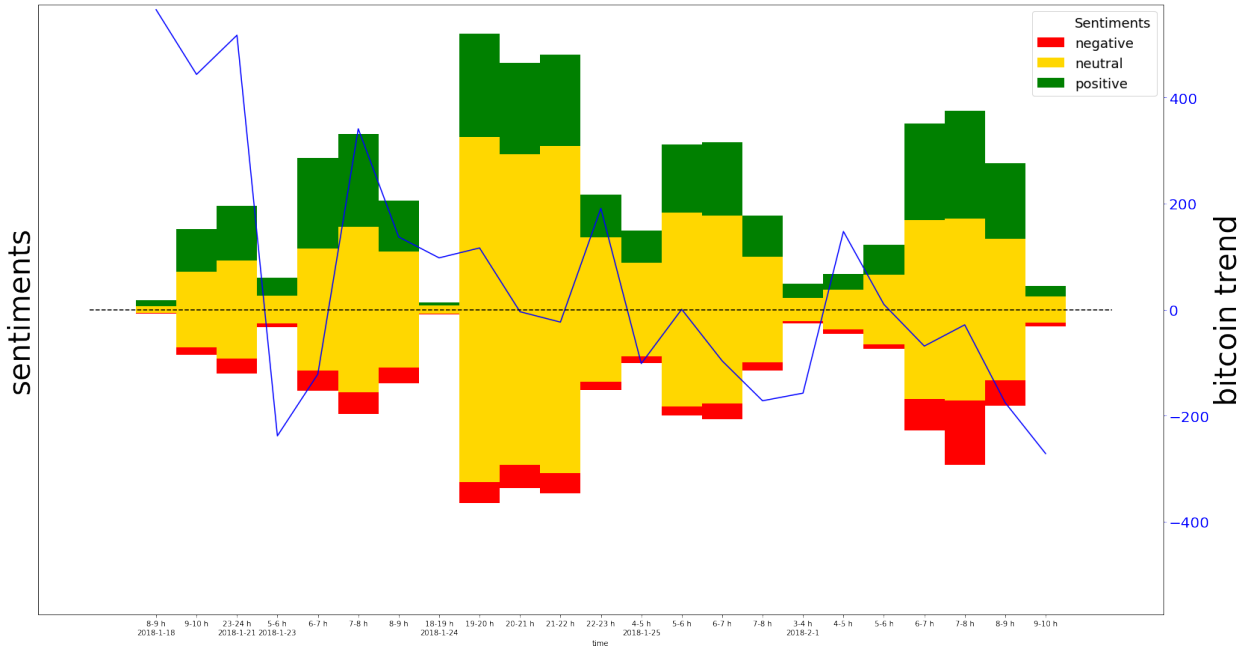

Abbildung 3: Übersicht der Auswertungsergebnisse. Die blaue Kurve entspricht dem Bitcon-Trend. Die farblich unterteilten Säulen stellen die Anzahl an Meinungen dar, die den Clustern positiv (grün), neutral (gelb) und negativ (rot) dar. Die Säulenhöhe steht für die Anzahl an Nachrichten pro Stunde.

- Wie zu erwarten, ist die Anzahl der Meldungen in den frühen Morgenstunden geringer als am Abend.
- Der Anteil der neutralen Meldungen ist auffallend hoch.
- Es gibt im Allgemeinen mehr positive Tweets zum Thema Bitcoin als negative.
- Der Bitcoin-Trend ist am 18.01. und am 21.01.2018 positiv. Zu den jeweiligen Zeitpunkten wurden jedoch nur verhältnismässig wenig Tweets registriert.
- Am 24.01.2018 ist das Interesse am Bitcoin deutlich gestiegen.
- Der negative Trend am 01.02.2018 wurde von mehr negativen Tweets begleitet, als zu allen anderen betrachteten Zeitpunkten. Jedoch ist auch die Anzahl der positiven Nachrichten gestiegen.

Insgesamt lässt sich kein Zusammenhang zwischen den Bitcoin-Trend und den Twitter-

Nachrichten erkennen. Dies ist zum einen auf die geringe Stichprobe und die unregelmäßigen Zeitpunkte der Twitterdaten zurückführen. Zum anderen deuten die vielen, als neutral eingestuften Meldungen darauf hin, dass der von *TextBlob* verwendete Klassifikator nicht ausreichend gut auf die Problemstellung angepasst ist.

Zusammenfassung und Ausblick

In dieser Arbeit wurde eine Sentiment-Analyse von Twitterdaten mit Hilfe von Spark durchgeführt. Da dabei Python genutzt wurde, sind bei der Textanalyse Schwierigkeiten mit der Kodierung aufgetreten, die eine Bereinigung der Twitterdaten nötig machen.
Die Sentiment-Analyse mit Hilfe von *TextBlob* kann ausschliesslich auf englischsprachigen Texten durchgeführt werden. Auch wenn eine Erkennung der Sprache angeboten und eine anschließende Übersetzung mit Hilfe von *Google Translate* möglich ist, ist dieses Verfahren sehr zeitaufwendig. Aus diesem Grund ist die Analyse in dieser Arbeit auf englischsprachige Tweets beschränkt. Trotzdem könnte eine detailliertere Auswertung des kompletten Datensatzes interessant sein.
Um bessere Rückschlüsse auf einen möglichen Zusammenhang zwischen Bitcoin-Kurs und Twitterdaten zu erlauben, ist ein konsistenter Datensatz notwendig, der keine zahlreichen zeitlichen Lücken aufweist. Zusätzlich könnten neben Twitter auch andere Quellen genutzt werden, um die öffentliche Meinung zum Thema Bitcoin auszuwerten. Ein Beispiel dafür wäre das Soziale Netzwerk *Facebook*.
In Abbildung 3 wird deutlich, dass der Anteil an neutralen Nachrichten besonders groß ist. Um ein genaueres Bild dieser Meldungen zu erhalten sollte der gewählte Klassifikator genauer betrachtet werden. Eine Möglichkeit ist eine detallierte Untersuchung des Verhaltens des Klassifikators auf verschiedene Texte, um gegebenenfalls die gewählten Schwellwerte anzupassen. Eine andere Option ist das Trainieren eines eigenen Klassifikators, der auf die Problemstellung angepasst ist.

Schlussendlich ist eine analoge Analyse nicht nur für den Begriff *Bitcoin*, sondern auch für andere Kryptowährungen möglich.