**NLP:**

**Stimmungsanalyse rund um das Thema Ethereum**

**Table of contents**

1. **Einleitung**
2. **Methoden**
3. **Tweepy**
4. **NLTK & Spacy**
5. **VADER**
6. **Wordcloud**
7. **Ergebnisse**
8. **Diskussion**
9. **Zusammenfassung**
10. **Einleitung**

Dieses Projekt beschäftigt sich mit dem zentralen Thema des „Natural Language Processing“, das bedeutet mit der natürlichen Verarbeitung von Sprache durch Computer. Das Projekt behandelt diesen Bereich der NLP indem es sich näher mit dem Aspekt „Opinon Mining“ befasst. Das bedeutet, es wurde untersucht, wie es mit der Hilfe eines Algorithmus ermöglicht werden kann, einen Textkorpus zu bearbeiten und den emotionalen Ton hinter dem Textkorpus zu identifizieren. Man möchte eine augenblickliche allgemeine Meinung der Textverfasser ermitteln können.

Der Hintergrund für dieses Projektes war die Überlegung wie es möglich sein könnte herauszufinden, wie sich im Allgemeinen die Meinungen auf dem aktuellen Krypto Markt bezüglich der Kryptowährung Ethereum verhalten, um daran angelehnt eine Investitionsentscheidung zu treffen. Die Hypothese ist hierbei, dass es möglich sein sollte, eine allgemeine positive, negative oder neutrale Haltung zu der Thematik „Ethereum“ aus Texten hinsichtlich der Thematik herauszuarbeiten sei. Fraglich ist hierbei, wie nachvollziehbar und akkurat diese Klassifizierung durch das Verfahren sein kann.

Das Projekt wird zur Überprüfung der Hypothese einen Textkorpus verwenden, der aus 2022 Tweets besteht. Die Tweets werden durch „Scraping“ von Twitter Nutzeraccounts, die sich auf der Plattform mit dem Thema Kryptowährung beschäftigen, heruntergeladen. Dieser Korpus wird mittels der NLP-Tools „NLTK“ und „Spacy“ bereinigt und vorbereitet. Im Anschluss wird dieser bereinigte Datensatz mittels „VADER“, ein Algorithmus zur Stimmungsanalyse, weiterbearbeitet. Zusätzlich wird das Modul „Wordcloud“ verwendet. Diese beiden Schritte sollen die ermittelten Tendenzen des Textkorpus zusätzlich veranschaulichen.

Das Projekt erwartet eine eindeutige Ermittlung der aktuell herrschenden Tendenzen in der Kommunikation der Interessenten an Kryptowährung.

1. **Methoden**

IN diesem Projekt wurden verschiedenen Module von Python verwendet, die die Bearbeitung des Textkorpus und die Ermittlung der ausgedrückten Empfindungen ermöglichen. Die Module sind „Tweepy“, „NLTK“, „Spacy“, „VADER“ und „WordCloud.“

1. **Tweepy:**

Tweepy[[1]](#footnote-1) ist ein Open-Source-Python-Paket, das es ermöglicht, mit Python auf die Twitter-API zuzugreifen. Die Twitter-API gibt Entwicklern den Zugriff auf die meisten Funktionen von Twitter. Sie können die API verwenden, um Informationen zu Twitter-Entitäten wie Tweets, Benutzer und Trends zu lesen und zu schreiben. API macht folglich HTTP-Endpunkten verfügbar, die sich auf Folgendes beziehen: Tweets, Retweets, Likes, Direktnachrichten, Favoriten, Tendenzen und Medien.

Die Twitter-API verwendet OAuth, ein offenes Autorisierungsprotokoll, um alle Anfragen zu authentifizieren. Bevor man die Twitter-API aufrufen kann, müssen Authentifizierungsdaten erstellt und konfiguriert werden. Es müssen also die erforderlichen Authentifizierungsdaten erstellt werden, um die API verwenden zu können. Diese Anmeldeinformationen sind vier Textzeichenfolgen:

1. Consumer key
2. Consumer secret
3. Access token
4. Access secret
5. **NLTK & Spacy:**

|  |
| --- |
|  |

1. **VADER:**

Valence Aware Dictionary and sEntiment Reasoner oder „VADER“[[2]](#footnote-2) ist ein regelbasiertes Modell zur Stimmungsanalyse. VADER fungiert als Lexikon. Es kann Vokabeln, Abkürzungen, Großschreibungen, wiederholte Satzzeichen, Emoticons usw., die normalerweise auf Social-Media-Plattformen verwendet werden, um die eigene Stimmung auszudrücken, effizient verarbeiten, was es zu einer großartigen Lösung für die Textanalyse der Stimmung in den sozialen Medien macht.

VADER hat den Vorteil, die Stimmung eines beliebigen Textes zu bewerten, ohne dass eine vorherige Schulung, wie es bei Machine-Learning-Modellen der Fall sein kann, erforderlich ist. Das von VADER generierte Ergebnis ist ein Wörterbuch mit 4 Schlüsseln: „neg“, „neu“, „pos“ und „compound“. Die Bezeichnungen „neg“, „neu“ und „pos“ bedeuten jeweils negativ, neutral und positiv. Ihre Summe sollte bei Float-Operation gleich 1 oder nahe daran sein. Die Bezeichnung „compound“ entspricht der Summe der Wertigkeiten jedes Wortes im Lexikon und bestimmt den Grad der Stimmung und nicht den tatsächlichen Wert im Gegensatz zu den vorherigen. Sein Wert liegt zwischen -1 (extremste negative Stimmung) und +1 (extremste positive Stimmung). Die Verwendung der zusammengesetzten Punktzahl kann ausreichen, um die zugrunde liegende Stimmung eines Textes zu bestimmen, denn für:

* eine positive Stimmung, Compound ≥ 0,05
* eine negative Stimmung, Verbindung ≤ -0,05
* eine neutrale Stimmung, die Verbindung liegt zwischen]-0,05, 0,05[

1. **Wordcloud:**

Wordcloud [[3]](#footnote-3)ist im Grunde eine Datenvisualisierungstechnik zur Darstellung von Textdaten. Es geht darum die Häufigkeit von Wörtern in einem Text darzustellen, wobei die Größe jedes Wortes seine Häufigkeit oder Bedeutung angibt. Signifikante Textdatenpunkte können mit einer Wortwolke hervorgehoben werden. Wortwolken werden häufig zur Analyse von Daten von Websites sozialer Netzwerke verwendet.

Jedes Wort in der Wolke hat eine variable Schriftgröße und einen variablen Farbton. Somit hilft diese Darstellung dabei, hervorstechende Wörter zu bestimmen. Eine größere Schriftgröße eines Wortes zeigt seine Hervorhebung relativ zu anderen Wörtern im Cluster. Die Anzahl der Wörter spielt beim Erstellen einer Wortwolke eine wichtige Rolle. Eine größere Anzahl von Wörtern bedeutet nicht immer eine bessere Wortwolke, da sie unübersichtlich und schwer zu lesen ist. Eine Wortwolke muss immer semantisch aussagekräftig sein und darstellen, wofür sie gedacht ist.

1. **Ergebnisse:**

Das Ergebnis des durchgeführten Projektes lässt sich anhand der Darstellungen Fig 1., Fig. 2 und Fig.3 erläutern.

In Fig 1. ist erkennbar, dass 2022 Tweets als Textkorpus durch VADER analysiert wurden. Das Ergebnis der Analyse durch VADER sind 1143 Zählungen für die durch VADER identifizierte Empfindung „neu“, 830 Zählungen für die durch VADER identifizierte Empfindung „pos“ und 49 Zählungen für die durch VADER identifizierte Empfindung „neg“ hinsichtlich „Ethereum“.

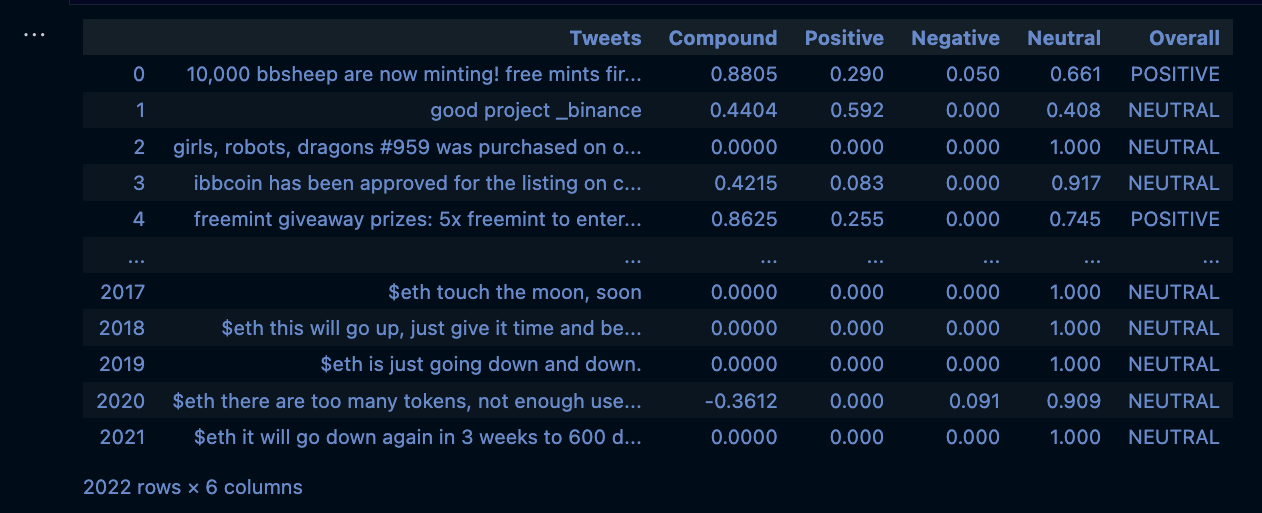


Fig. 1

In Fig 2. ist erkennbar, dass auf den durch VADER verarbeiteten Textkorpus aus allen Tweets über „Ethereum“, 57% „neu“, 41% „pos“ und 2% „neg“ Empfindungen entfallen.

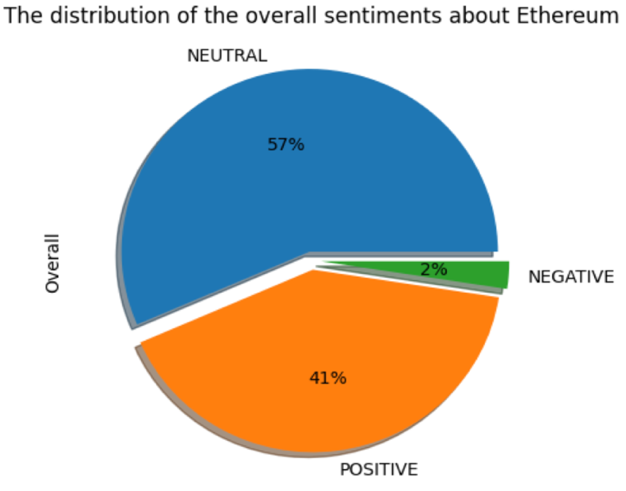


Fig. 2

Die Wordcloud (Fig.3) gibt den folgenden Worten eine Hervorhebung: „ethereum“, „nft“, „free“, „mint“, „winner“, „friend“, „tag“, „follow“, „retweet“,



Fig. 3

1. **Diskussion:**

Die Resultate des Projektes zeigen auf den ersten Blick eine erhoffte Verteilung von Emotionen innerhalb des Textkorpus zum Thema Kryptowährung und Ethereum: 57% „neu“, 41% „pos“ und 2% „neg“. Das Ergebnis ließe, unhinterfragt, somit schlussfolgern, dass die Krypto Community einer Investition in Ethereum allgemein befürworten würde. Dieses Argument wird durch die von der Wortwolke gelieferten Ergebnisse untermauert, die die Wörter „ethereum“, „tag“, „friend“, „retweet“ und „winner“ hervorheben. Diese Worte könnten durchaus als Werbung für die Kryptowährung interpretiert werden.

Allerdings würden Kenner des aktuellen Marktes und des Krypto Marktes vermutlich zu einer genteiligen Prognose gelangen. Ein Sachkundiger würde wesentlich umfangreichere Aspekte in seine Bewertung einfließen lassen, wie z.B. die aktuelle weltpolitische und wirtschaftliche Lage.

Es ist überdies fraglich, wie korrekt die Ergebnisse der Analyse mit VADER tatsächlich sind. Schaut man sich im Detail die Analyse des Textkorpus mittels VADER an, so fallen einige Ergebnisse nicht so aus, wie ein Mensch sie bewertet hätte. In Reihe 2017 z.B. der Tweet *„$eth touch the moon, soon*” wird durch VADER mit dem „overall“ Ergebnis „neu“, d.h. neutral, ausgewertet. Ein sachkundiger Mensch weiß jedoch, dass dieser Tweet in der Krypto Community als sehr positiv zu verstehen gilt. Das weiß der Sachkundige durch die Erfahrung in der Kommunikation mit Gleichgesinnten und die Bewertung mit „pos“ ist also sozial geprägt. VADER macht hier folglich eine Fehlanalyse.

Es gilt wohl auch bei der Diskussion um die Resultate zu bedenken, dass der Einsatz eines Algorithmus wie VADER oder eines Modules wie WordCloud dazu dienen kann einen ersten Eindruck zu vermitteln, welcher jedoch folgend von einem Leser eingeordnet, validiert oder falsifiziert werden muss. Dies kann nicht über den Einsatz eines einzigen „tools“ erfolgreich gelingen. Eine Kombination aus mehreren zusätzlichen Ansätzen und Aspekten (Börsenkurse, Preisbewegungen, politische und gesellschaftliche Faktoren wie der Krieg in der Ukraine oder ein Tweet von Elon Musk) erscheint notwendig um die Ergebnisse der Analyse/n einordnen zu können. Außerdem wird diese Einordnung, Validierung oder Falsifizierung immer im Kontext der herangezogenen Zusatzinformationen des Lesers stehen und ist daher dann ein individuelles Ergebnis.

1. **Zusammenfassung:**

This is the summing up of your argument and should relate back to the introduction.

The conclusion should only consist of a few sentences and should reiterate the findings of your research.

If appropriate, suggest how to improve the procedure, and what additional experiments or research would be helpful.

|  |
| --- |
|  |

1. Die offizielle Webseite von „Tweepy“: https://docs.tweepy.org/en/stable/ [↑](#footnote-ref-1)
2. Eine ausführliche Beschreibung von VADER: <https://github.com/cjhutto/vaderSentiment/blob/master/README.rst> [↑](#footnote-ref-2)
3. Die offizielle Webseite von „wordcloud“: <https://pypi.org/project/wordcloud/> [↑](#footnote-ref-3)