NLP:
Stimmungsanalyse rund um das Thema Ethereum

Betreuer: Tatiana Bladier, Kilian Evang

Verfasst von Aldi Halili: 3013735, Brahim Chouikha: 2837831, Saber Klai: 2846796

Table of contents

I. 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 Spacy und Regex 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.

II. 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“,, „Spacy“, „VADER“ und „WordCloud.“

1. **Tweepy:**

Tweepy[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

2. **Spacy & Regex:**

SpaCy ist eine Open Source Bibliothek für die Programmiersprache Python**,** die für das Natural Language Processing eingesetzt wird. SpaCy ermöglicht den Text automatisch computerbasiert analysieren und verstehen zu können. Für die Verarbeitung von Text und das Verstehen des Inhalts bietet SpaCy unterschiedliche Funktionen an wie z.B.:

- Tokenisierung: Segmentierung von Text in Wörter und Satzzeichen
- Part-of-speech (POS) Tagging: Zuordnen von Wortarten
- Lemmatisierung: Zuordnung der Grundformen der Wörter
- Dependency Parsing: Zuordnung von Syntaxabhängigkeiten zwischen den identifizierten Token

[1] Die offizielle Webseite von „Tweepy“: https://docs.tweepy.org/en/stable/

- Named Entity Recognition (NER): Benennen von Objekten, Personen oder Orten
- Ähnlichkeitsanalyse von Wörtern, Sätzen oder kompletten Texten

Bezüglich unseres Projekts würde SpaCy für drei grundlegenden Schritte für die Vorverarbeitung von Tweets angewendet, um festzustellen welche Ergebnisse der VADER – Algorithmus bezüglich unten genannte Vorverarbeitungsschritte uns liefert, nämlich:

- **Tokenisierung**

  Die Tokenisierung ist der erste Schritt im jedem NLP-Modell. Die Hauptfunktion ist eine Eingabetext automatisch in Einheiten namens Tokens zu segmentieren. Normalerweise könnte ein Token ein Wort, Zahl oder Interpunktion sein.
  Für die Tokenisierung in diesem Projekt wurde Spacy Bibliothek durch Laden eines englischen Models von SpaCy Objekt verwendet. SpaCy teilt den Text automatisch in Token auf, wenn ein Text mit dem Modell erstellt wird. Ein Token bezieht sich einfach auf einen einzelnen Teil eines Satzes, der einen semantischen Wert hat.
  Durch Anwendung dieses ersten Schrittes (Tokenisierung) wurde festgestellt, dass die Wörter wie are ‘nt und don’t getrennt, betrachtet werden. Das bedeutet „are“ und „nt“, als zwei unterschiedliche Wörter auftauchen. Unten wird die Tokenisierung als Beispiel von unserem Korpus dargestellt. Das könnte bei Sentiment Analyse ein Nachteil sein, weil Tokens wie "n't" 'too', 'much' durch nächsten Schritt (Stoppwörter Entfernung) entfernt werden und somit die Bedeutung des Satzes beeinflussen (Siehe Fig. 1).
  Tweet 2020: “$ETH There are too many tokens not enough use cases and the economy is lacking People en businesses don't have enough money and too much debt”.
  Tokenisierung: “['$', 'ETH', 'There', 'are', 'too', 'many', 'tokens', 'not', 'enough', 'use', 'cases', ‘and’, 'the', 'economy', 'is', 'lacking', 'People', 'en', 'businesses', 'do', "n't", 'have', 'enough', 'money', 'and', 'too', 'much', 'debt']”.
  Nach der Tokenisierung, der Text wurde mithilfe der „join()“ Methode in einem String umgewandelt, die als Eingabe für den Stoppwörter Entfernung Vorverarbeitungsschritt gilt.

- **Stoppwörter Entfernung**

  Im englischen Wortschatz gibt es viele Wörter wie "I", "the","are", und "you", die sehr häufig im Text vorkommen, aber sie tragen keine wertvollen Bedeutungen für NLP. Es

wird empfohlen die Stoppwörter im Rahmen der Verarbeitungsschritt zu entfernen. Durch Entfernung der Stoppwörter verringert sich die Größe der Textkorpus und erhöht sich somit die Leistung des NLP-Modells. Auf der anderen Seite kann für unser Projekt benachteiligt sein, weil durch Stoppwörter Entfernung die Bedeutung des Textes (Tweets) geändert werden kann. Wenn man diesen Satz “This is not a good way to talk” betrachtet, merkt man, dass, nach der Entfernung der Stoppwörter der Satz von negativ, positiv bewertet wird: “good way talk”.

In unserem Projekt haben wir SpaCy Bibliothek importiert. Wir laden das englische Sprachmodell des SpaCy-Objekts und speichern die Liste der Stoppwörter in einer Variable. Anschließend erstellen wir eine leere Liste, um Wörter zu speichern, die keine Stoppwörter sind.

Mit einer for-Schleife, die über den Text, nach dem Tokenisierung iteriert, überprüfen wir, ob das Wort in der Stoppwörterliste vorhanden ist, wenn nicht, fügen wir es an die Liste hinzu.

Schließlich bekommen wir die Liste der Wörter, die keine Stoppwörter enthalten, indem wir die Funktion "join()" verwenden, und somit haben wir eine endgültige Ausgabe, bei der alle Stoppwörter aus der Zeichenfolge entfernt werden. Diese Ausgabe wird als Eingabe für den letzten Vorverarbeitungsschritt (Lemmatisierung) angewendet.

- **Lemmatisierung**

Lemmatisierung ist das letzte Vorverarbeitungsschritt in diesem Projekt, in dem ein Wort in sein Lemma repräsentiert wird**.** Ein Lemma ist normalerweise die Wörterbuchversion eines Wortes z.B.: die Wörter "spielen", "gespielt" und "spielt" haben alle das gleiche Lemma des Wortes "spielen".

Auf das geladene englische Sprachmodel geben wir den Text (Tweets) ein, die in einer Variable gespeichert wird. Mit einer for-Schleife iterieren wir über den Text und mit der ‚.lemma_‘ Funktion bekommen wir das Lemma für jedes Wort. Die Lemmatisierung-Ausgabe wird als Eingabe für den VADER-Algorithmus verwendet.

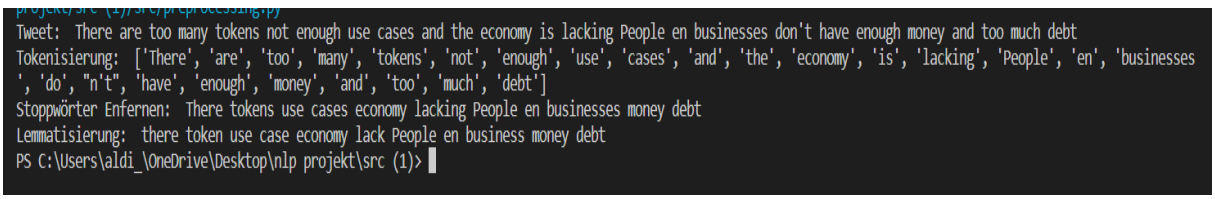


Fig. 1 Hier wird dargestellt, wie ein Tweet sich durch alle Verarbeitungsschritte verringert.

Regex wird in dem Projekt angewendet, um die Tweets zu reinigen. Die Tweets kommen aus sozialen Medien, wobei auf die Schreibweise nicht geachtet wird, und somit viele Sonderzeichen im Text vorkommen. Daher spielt die Entfernung der Sonderzeichen bei der Datenvorbereitung eine wichtige Rolle. Gleichzeitig bringt die Extrahierung der Daten aus dem Internet eine große Herausforderung bezüglich der Datenreinigung mit sich. Was unsere Tweets angeht, besteht der Bedarf an Entfernung von vielen Sonderzeichen wie zum Beispiel: @mentions, hyperlinks, website link wie z.B. http/https/www. Andere Sonderzeichen die durch Regex entfernt worden sind: #&\()*+,-/:;@[\]^_`{|}~ξ. Unsere Tweets enthalten auch Emoji, die durch Regex entfernt werden können.
Die Ausrufzeichen, Fragezeichen und Emoji können die Bedeutung eines Satzes hinsichtlich des Vader Algorithmus verändern. Aus Diesem Grund speichern wir die vorbereiteten Daten in zwei unterschiedliche Dateien. In einer sogenannte „preprocessed_minimal" Datei werden, sowohl Ausrufzeichen, Fragezeichen als auch Emoji beibehalten und die mit SpaCy Vorverarbeitungsschritte werden nicht angewendet. In der zweiten „preprocessed_maximal" Datei wird der Vorverarbeitungsschritt in jeder Hinsicht vollständig durchgeführt. Das Ziel ist es wie oben erwähnt wurde, die von VADER gelieferte Ergebnisse zu vergleichen.

3. VADER:

Valence Aware Dictionary and sEntiment Reasoner oder „VADER"[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

[2] Eine ausführliche Beschreibung von VADER:
https://github.com/cjhutto/vaderSentiment/blob/master/README.rst

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[

4. Wordcloud:

Wordcloud [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.

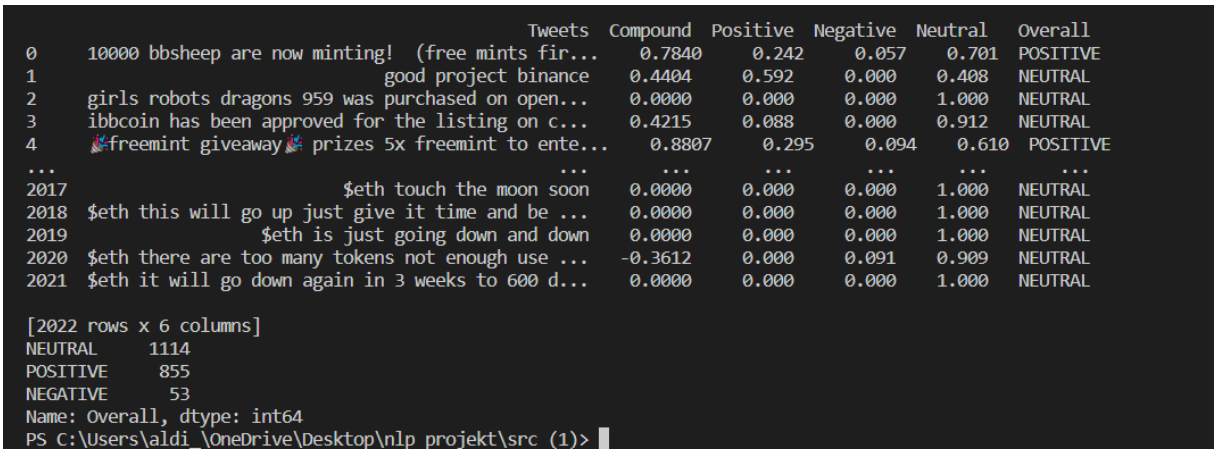
Jedes Wort in der Wolke hat eine variable Schriftgröße und einen variablen Farbton. Somit hilft diese Darstellung dabei, hervorstechende Wörter zu bestimmen. Eine größere Schriftgröße eines Wortes zeigt seine Hervorhebung relativ zu anderen Wörtern im Cluster. Die Anzahl der Wörter spielt beim Erstellen 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.

III. Ergebnisse:

Das Ergebnis des durchgeführten Projektes lässt sich anhand der Darstellungen Fig 2., Fig. 3., Fig. 4., und Fig.5 erläutern.
In Fig. 2 wird das sogenannte „preprocessed_minimal“ Datei für VADER Algorithmus angewendet, wobei außer mentions, hyperlinks und website link nichts entfernt worden ist und Vorverarbeitungsschritt nicht übernommen wird.

[3] Die offizielle Webseite von „wordcloud“: https://pypi.org/project/wordcloud/

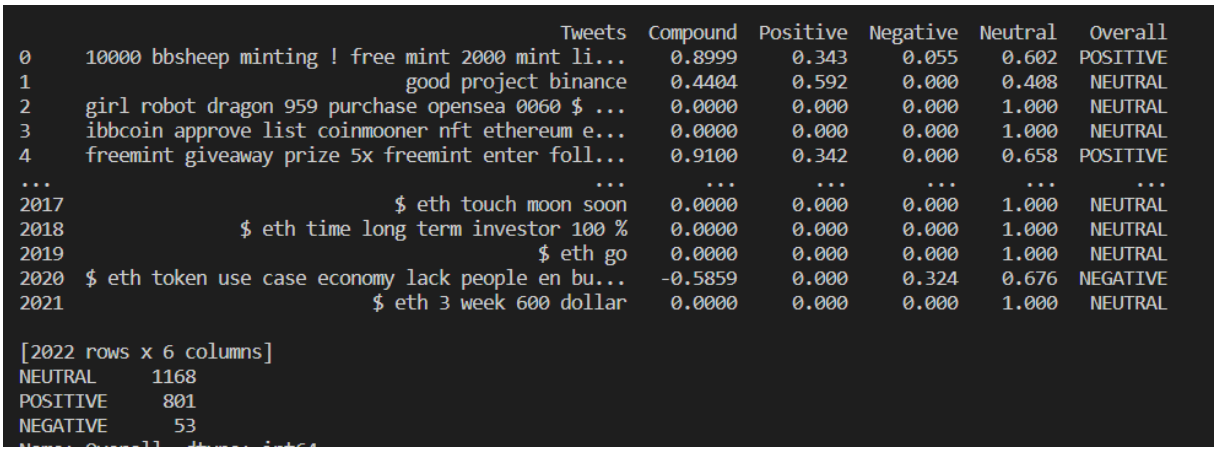
In Fig 2. ist erkennbar, dass 2022 Tweets als Textkorpus durch VADER analysiert wurden. Das Ergebnis der Analyse durch VADER sind 1114 Zählungen für die durch VADER identifizierte Empfindung „neu“, 855 Zählungen für die durch VADER identifizierte Empfindung „pos“ und 53 Zählungen für die durch VADER identifizierte Empfindung „neg“ hinsichtlich „Ethereum“.



```
                                              Tweets  Compound  Positive  Negative  Neutral   Overall
0     10000 bbsheep are now minting!  (free mints fir...    0.7840     0.242     0.057    0.701  POSITIVE
1                                good project binance    0.4404     0.592     0.000    0.408   NEUTRAL
2     girls robots dragons 959 was purchased on open...    0.0000     0.000     0.000    1.000   NEUTRAL
3     ibbcoin has been approved for the listing on c...    0.4215     0.088     0.000    0.912   NEUTRAL
4     🎉freemint giveaway🎉 prizes 5x freemint to ente...    0.8807     0.295     0.094    0.610  POSITIVE
...                                                 ...       ...       ...       ...      ...       ...
2017                              $eth touch the moon soon    0.0000     0.000     0.000    1.000   NEUTRAL
2018  $eth this will go up just give it time and be ...    0.0000     0.000     0.000    1.000   NEUTRAL
2019                    $eth is just going down and down    0.0000     0.000     0.000    1.000   NEUTRAL
2020  $eth there are too many tokens not enough use ...   -0.3612     0.000     0.091    0.909   NEUTRAL
2021  $eth it will go down again in 3 weeks to 600 d...    0.0000     0.000     0.000    1.000   NEUTRAL

[2022 rows x 6 columns]
NEUTRAL     1114
POSITIVE     855
NEGATIVE      53
Name: Overall, dtype: int64
PS C:\Users\aldi_\OneDrive\Desktop\nlp projekt\src (1)>
```

Fig. 2 minimal preprocessed Datei



```
                                              Tweets  Compound  Positive  Negative  Neutral   Overall
0     10000 bbsheep minting ! free mint 2000 mint li...    0.8999     0.343     0.055    0.602  POSITIVE
1                                good project binance    0.4404     0.592     0.000    0.408   NEUTRAL
2     girl robot dragon 959 purchase opensea 0060 $ ...    0.0000     0.000     0.000    1.000   NEUTRAL
3     ibbcoin approve list coinmooner nft ethereum e...    0.0000     0.000     0.000    1.000   NEUTRAL
4     freemint giveaway prize 5x freemint enter foll...    0.9100     0.342     0.000    0.658  POSITIVE
...                                                 ...       ...       ...       ...      ...       ...
2017                              $ eth touch moon soon    0.0000     0.000     0.000    1.000   NEUTRAL
2018               $ eth time long term investor 100 %    0.0000     0.000     0.000    1.000   NEUTRAL
2019                                      $ eth go    0.0000     0.000     0.000    1.000   NEUTRAL
2020  $ eth token use case economy lack people en bu...   -0.5859     0.000     0.324    0.676  NEGATIVE
2021                           $ eth 3 week 600 dollar    0.0000     0.000     0.000    1.000   NEUTRAL

[2022 rows x 6 columns]
NEUTRAL     1168
POSITIVE     801
NEGATIVE      53
```

Fig. 3 maximal preprocessed Datei

In Fig.3 wobei alle Vorverarbeitungsschritte übernommen werden und der Text vollständig gereinigt wird, erfolgen folgende Ergebnisse:

1168 Zählungen „neu“

801 Zählungen „pos“

53 Zählungen „neg“

In Fig 4. ist erkennbar, dass auf den durch VADER verarbeiteten Textkorpus aus allen Tweets von „preprocessed_minimal“ Datei über „Ethereum“, 55% „neu“, 42% „pos“ und 3% „neg“ Empfindungen entfallen. Andersrum liefert „preprocessed_maximal“ Datei ergebnisse wie: 58% „neu“, 40% „pos“ und 3% „neg“ Siehe Fig. 5

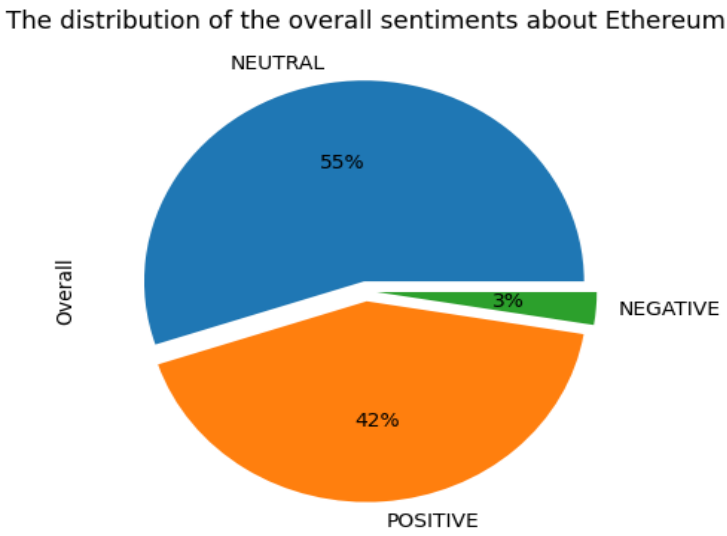


Fig. 4 „preprocessed_minimal“ Datei

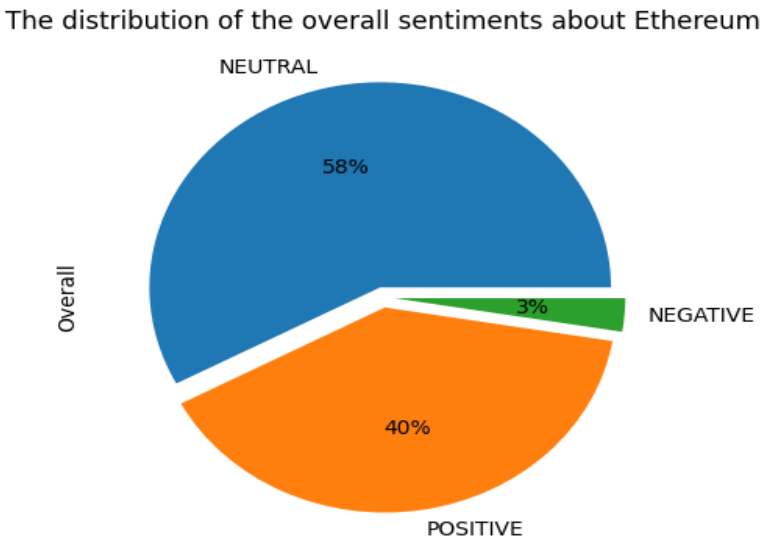


Fig. 5 „preprocessed_maximal“ Datei

Die Wordcloud (Fig. 6) gibt den folgenden Worten eine Hervorhebung: „ethereum“, „nft“, „free“, „mint“, „winner“, „friend“, „tag“, „follow“, „retweet“,

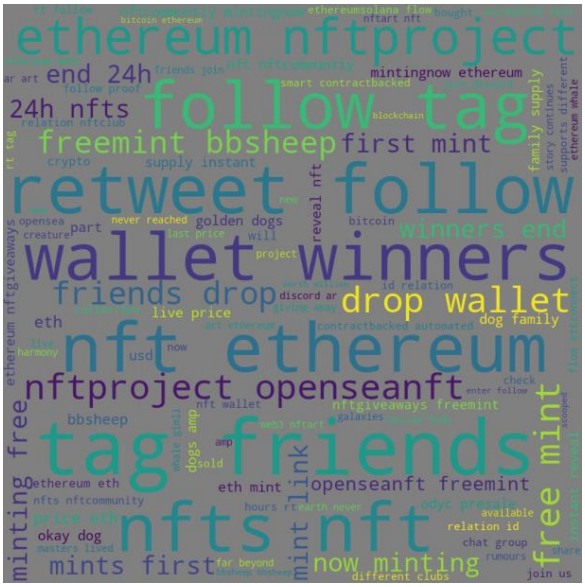


Fig. 6

IV. Diskussion:

Die Resultate des Projektes zeigen auf den ersten Blick eine erhoffte Verteilung von Emotionen innerhalb des Textkorpus zum Thema Kryptowährung und Ethereum: 55% „neu“, 42% „pos“ und 3% „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.