NLP analysis of Doodles project based on Twitter

Introduction

With billions of pages of data on the Internet, web data as a potential source of data has a huge potential to be utilized for strategic business development of the industry. In addition, data acquisition has a high value of utilization in the field of finance and risk management (Penman et al., 2009). Data acquisition is a prerequisite and a necessary condition for big data analysis and plays an important role in the whole data utilization process.

NFT (Non-Fungible Tokens) relies on blockchain technology to make digital works unique. Its carrier can be in the form of a JDP image, video clip, and so on, but it has a unique Encryption Token. Unlike cryptocurrencies that also use blockchain technology, the serial number in NFT is unique, so each NFT piece is irreplaceable and indivisible. When you purchase an NFT, you acquire an indelible record of ownership of it. Others can use your purchased digital image with your permission, but your information as the copyright owner is recorded. Just like when you buy a work of art, it can be displayed and reproduced, but only you are the actual owner. In reality, many works are not identifiable after multiple handoffs of owner information. Unless one goes looking for a recognized institution to register. In contrast, the most fundamental value proposition of NFT relies on the decentralized nature of the blockchain, which makes it impossible for the information recorded by NFT to be tampered with by any individual or institution. For now, NFT is likely to be a new business growth point for the companies.

Doodles is based on a community-driven NFT digital collection project. It is considered one of the most promising NFT projects. Doodles NFTs sold out in a flash of the public sale, which shows its hotness. Similarly, according to the Nansen website, only 20 of the 103 experienced collectors who own Doodles NFTs have resold them since they were made available to the public, demonstrating the extraordinary value of the Doodles NFT collection. The success of the Doodles project lies not only in the artistic nature of the NFT themselves and the creative additions of famous artists but also in its marketing team and community building. The public sale format and the use of Twitter to build momentum led to stronger community cohesion and ultimately a strong consensus mechanism.

Numerous NFT works have been considered as collectible and unique artworks. Doodles, which already has specific value support, will become very meaningful for mining its information and data as NFT gradually becomes a consensus. This will help to understand the community's views and perceptions of NFT and also help people to identify the value of NFT. This report aims at demonstrating how to use technology to retrieve information from the Doodles platform and perform an analysis of the data.

Main body

API

To obtain information about Doodles, we considered news, research reports, or some publicly available data. As NFTs are a recent project, academic research and traditional media news on doodles are scarce, making it difficult to obtain valid data. NFT project owners or creators mainly discuss them through social media, or private communities. We chose to collect data about the Doodles project through the Twitter API, considering that information cannot be legally and effectively accessed through private communities like Discord through the API. Twitter is one of the main social media channels for NFT work holders, creators and interested parties to express their views. In addition to this, Twitter allows users to use their NFT holdings as avatars, which adds to the relevance of NFT to twitter. This is why it is important to examine the impact of sentiment on Twitter on the NFT project.

Twitter's API is based on a client-server approach, whereby a developer account creates a client application to connect to Twitter's servers to receive certain information. The request is triggered by the client and the developer account can retrieve the information it needs about the twitter object. Twitter has released the Twitter API v2 to give users more levels of access, which includes three levels: Essential, Elevated and Academic Research. In order to break the limits on how fast and how much data an Essential account can access, we have requested the Elevated level of access.

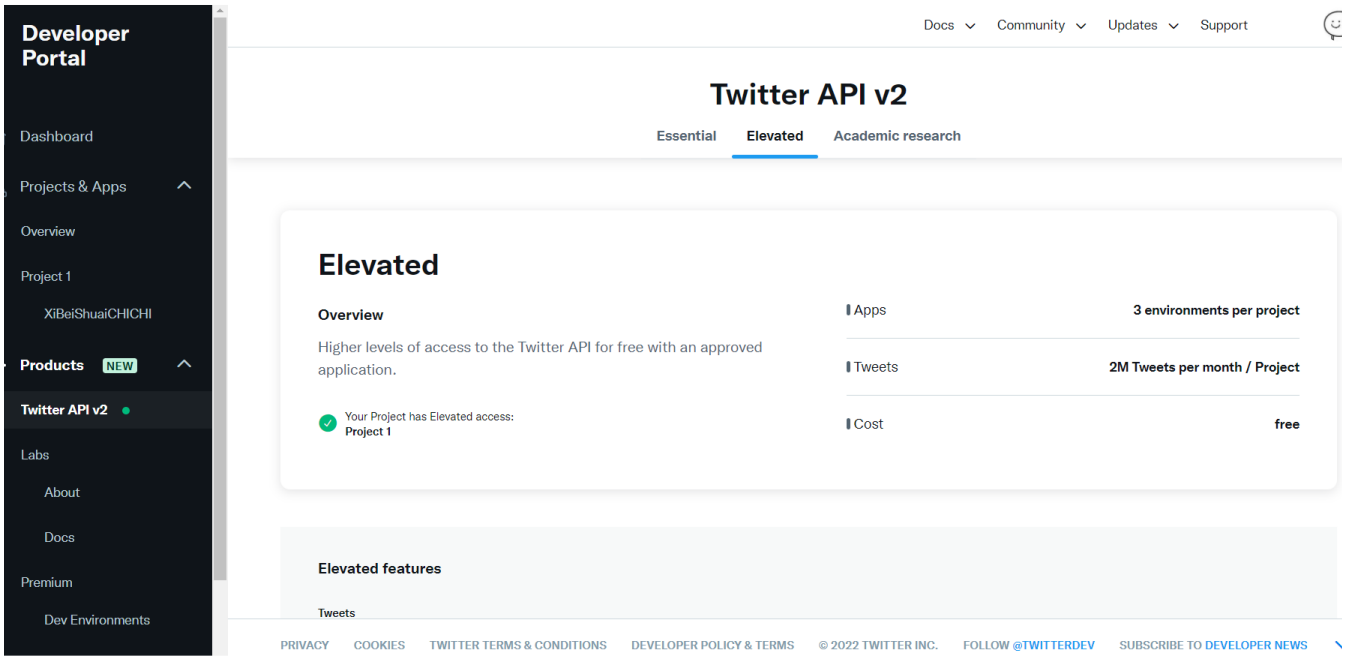

Figure 1: Twitter API developed portal

After obtaining API_Key, API_Key_Secret, Access_Token, Access_Token_Secret via the Twitter API v2, we created an ini configuration file to store this information.

```
# read configs
config = configparser.ConfigParser()
config.read('config.ini')

api_key = config['twitter']['api_key']
api_key_secret = config['twitter']['api_key_secret']

access_token = config['twitter']['access_token']

access_token_secret = config['twitter']['access_token_secret']

```

In the code implementation, we use the configparser package in python to read the Consumer Keys and Authentication Tokens and then use the tweepy package to access the twitter API.

Our main research object is the moods of the discussion participants of the doodles project on social media, including the interaction between the official account and the community. Therefore, we limited the code to retrieving the data under the official account of doodles, but the results are very limited. It mainly focuses on official activities and cannot capture the discussion information of the whole community of doodles.

Then, we limit the keyword to doodles, because the word ‘doodles’ represents a wide range of semantics. For example, doodle represents a kind of painting. The content we retrieved is very messy and noisy. Finally, we propose a hypothesis: the twitter user of @ doodles is the user who wants to publish information about the doodles project and the official account of doodles. So we chose the keyword '@ doodles', which is to retrieve all tweets from user who addresses doodles official accounts. After trying, the final result matches the information we want to get. In the process of running the code, due to the limitations of twitter API requests and rate, we use the wait_on_rate_limit parameter to let the code run automatically wait for the rate limit to be supplemented.

```
# authentication
auth = tweepy.OAuthHandler(api_key, api_key_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)

# search tweets
keywords = '@doodles'
# keywords =
limit=500000

tweets = tweepy.Cursor(api.search_tweets, q=keywords, count=200, tweet_mode='extended').ite
# tweets = api.user_timeline(screen_name=user, count=limit, tweet_mode='extended')

#help(tweepy.Cursor)

# create DataFrame ,
columns = ['Time','User', 'Tweet']
data = []

for tweet in tweets:
    data.append([tweet.created_at,tweet.user.screen_name, tweet.full_text])
#
print(data)

df = pd.DataFrame(data, columns=columns)

print(df)

df.to_csv('Doodles_Final.csv')


```

To study recent tweets related to the Doodles project, we plan to obtain as much data as possible. Due to

permissions restrictions on the twitter developer account, we only got data for a maximum of 9 days. After selecting the maximum limit of 500,000 tweets in the code, running the code we successfully obtained 138,503 tweets from 2022-03-17 to 03-09. The text information includes three main contents, Twitter username, tweet content, and acquisition time

Natural Language Process

3.1 Procedures

Natural language processing is a field in computer science and computational linguistics for the study of the interaction between human (natural) language and computers. When people see text, they can usually understand the meaning. When computers see text, they can only see strings of characters and cannot translate them to real-world things or understand the ideas they contain. As humans become increasingly dependent on computing systems, it is becoming increasingly important for computers to understand text and language. This is where Natural Language Processing (NLP) comes in. This project focuses on the application of NLP to pre-process doodles from Twitter for later sentiment analysis. The main processes and techniques used were Tokenize, Stopwords, Part of Speech tagging and Lemmatization.

3.1.1 Tokenize

The first step in processing text is to break it down into words. Words are called tokens and the process of splitting text into tokens is called tokenization, and the model or tool used for tokenization is called a tokenizer. NLTK provides the Tokenizer class for pre-processing text documents for deep learning.

```
def decompose_word(doc):
    txt = []
    for word in doc:
        txt.extend(word.split())
    return txt

# decompose a list of sentences into words by self-defined function
tokenslist = decompose_word(doc_out)
# decompose a list of sentences into words from NLTK module
 tokens = nltk.word_tokenize(doc_out)
```

The function nltk.word_tokenize is used to split the sentence into individual words as the figure1 shows.

3.1.2 Stopwords

Stopwords are words that are filtered out in natural language processing, usually meaningless definite articles, indefinite articles, conjunctions, etc. There is no standard for this, but rather for specific tasks and documents.

```
STOPWORDS = ["an", "a", "the", "or", "and", "thou", "must", "that", "this", "self", "unless
```

```
def remove_stopwords(txt):
    """Delete from txt all words contained in STOPWORDS."""
    words = txt.split()
    # words = txt.split(" ")
    for i, word in enumerate(words):
        if word in STOPWORDS:
            words[i] = " "
    return (" ".join(words))
```

Since the result of the splitting contains word interference items such as punctuation, the result should be filtered by means of functions and stopwords. So we define the function to remove stopwords. After the first attempt, we found that there were still many words in the results that were not useful for the case study, such as AMP, httpstcofkugazrt and doodles, so we decided that the usual stopword list was no longer sufficient for processing Twitter data. We added AMP, and HTTP prefixes to the stopwords as the figure2 shows to avoid useless data.

3.1.3 Part of Speech tagging

Part of Speech tagging is a supervised learning solution that uses features such as previous word, next word, initial capitalization, etc. NLTK has the ability to acquire part of speech and start processing word lexicality after the sentence break tokenization process.

In NLTK, it is easy to use the nltk.pos_tag() function to get the lexicality of the word in the sentence. Once the part of speech of the words has been obtained, it is easy to perform lemmatization.

It is important to specify the part of speech of the word when using it specifically, otherwise, lemmatization may not work well, as in the following figure.

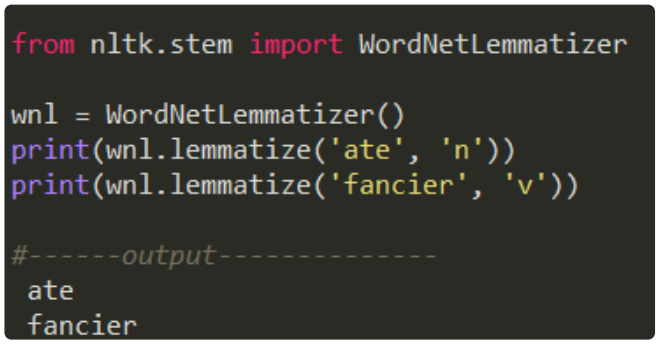

Figure 2: wrong lemmatization

3.1.4 Lemmatization

In simple terms, lemmatization is the process of removing the affixes from a word and extracting the main part of the word, usually from the dictionary, unlike stemming, where the extracted word does not necessarily appear in the word. For example, the word "cars" is reduced to the word "car" and the word "ate" is reduced to the word "eat ".

In Python's nltk module, WordNet is used to provide us with a robust function for lemmatization

```
# nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer
```

```
lemmatizer = WordNetLemmatizer()
tokens=[]
for i in tokenslist:
    tokens.append(lemmatizer.lemmatize(i))
print(tokens)

def lemmatize_all(doc_out):
    wnl = WordNetLemmatizer()
    for word, tag in pos_tag(word_tokenize(doc_out)):
        if tag.startswith('NN'):
            yield wnl.lemmatize(word, pos='n')
        elif tag.startswith('VB'):
            yield wnl.lemmatize(word, pos='v')
        elif tag.startswith('JJ'):
            yield wnl.lemmatize(word, pos='a')
        elif tag.startswith('R'):
            yield wnl.lemmatize(word, pos='r')
        else:
            yield word

```

Thus, we define the function lemmatize all in the figure4. In the above code, the wnl.lemmatize() function can perform lemmatization, the first argument is the word, and the second argument is the part of speech of the word, such as nouns, verbs, adjectives, etc. The result returned is the result of the word form reduction of the input word.

Sentiment analysis

Text mining and analysis of social media and product reviews has become a great tool to study data patterns in the majority of service business. This type of dataset can aid business in better understanding their customers and decision-making. Besides, the value of NFTs is primarily determined by the degree to which their inventors are recognized by customers. Therefore, the sentiment of NFT participants may have some impact on NFT prices. Kapoor et al., (2022) discovered that branding and metadata (Twitter and OpenSea features) had a greater influence on value than the NFT product itself. In this proposal, by collecting 130,000 tweets about Doodles from 9th March to 17th March, the authors found that Twitter users had considerably more positive views (32,961) about Doodles than negative views (5,891) and it is seen in Figure 1. As seen in Figure 2, there is a clear upward trend in Doodles' floor price over the same period. It is reasonable to suppose that the price of NFT products may be influenced by the opinion expressed on Twitter users towards the product.

Combining a statistical list of favorable phrases with the tweets in which they are found, we found over 4,000 comments regarding the highly praised SXSW event (e.g. cool, amazing, excited and awesome). From 12th March and 14th March, which is the time of the SXSW Doodles Event, the floor price and volume of Doodles increased constantly, with the volume of Doodles trading reaching a period high on 14th March. It is reasonable to presume that the event went well and was well welcomed by the participants, indicating the fact that the company's recent sales had considerable outpaced those of the majority of other companies in the field of NFT. Additionally, over 11,000 tweets directly expressed their admiration for Doodles, and 440

tweets suggested a bullish attitude towards the price of Doodles. Thus, the positive sentiment towards Doodles is currently strong among Twitter users

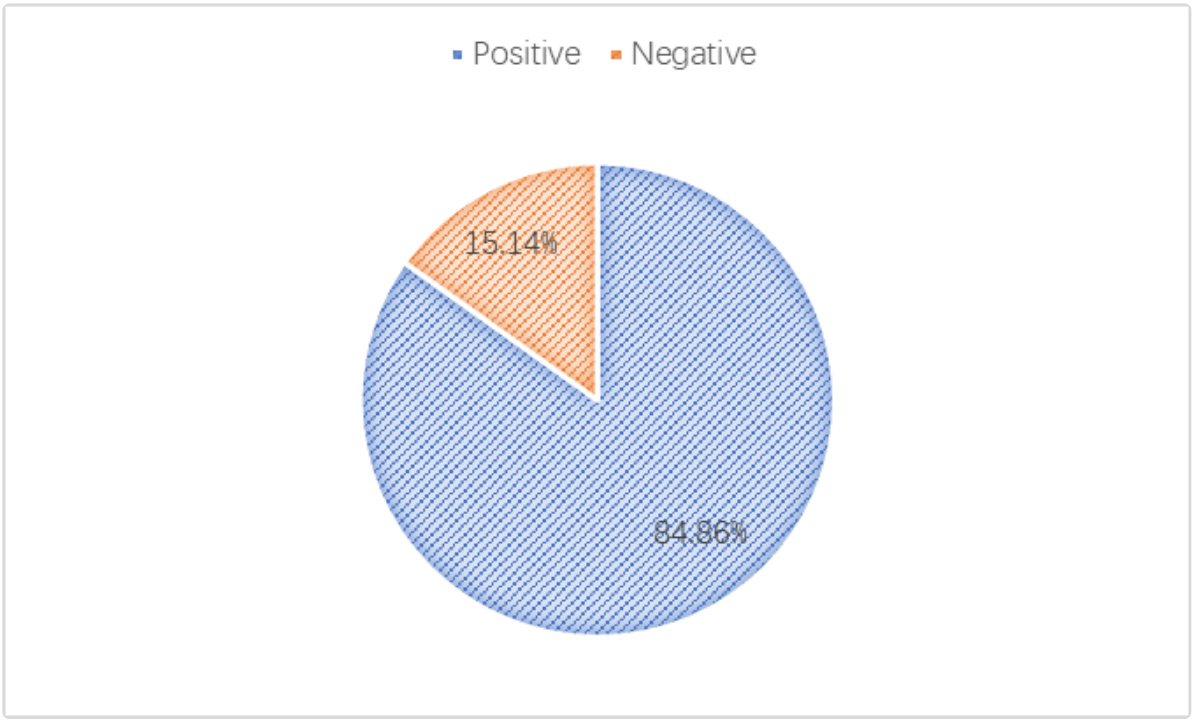

Figure 3: The proportion of sentiment words related to Doodles in twitter

Figure 4: Floor price of Doodles between 9th March and 17th March

The daily histogram of positive and negative sentiment words in tweets with is depicted in Figure 3. It is found that the Pearson correlation coefficient (ρ) between the number of tweets with sentiment/ positive sentiment and the floor price of Doodles NFT were between 0.7 and 0.8, indicating moderate positive correlations. However, whether or not there is a significant economic relationship between the tweets with sentiment views and the price of Doodle NFT still needs to be further determined. Overall, the analysis implies that the present high level of favorable sentiment regarding Doodles NFT among Twitter users may contribute to an increase in this project's floor price.

```
import numpy as np
price = [8.65,9.38,10.2,10.3,10.8,11.7,13.18,13,12.5]
total_tweets = [772,2103,2227,3215,5132,6532,11604,8382,1825]
pos_tweets = [643,1736,1771,2687,4198,5191,10110,7321,1449]

co_tweets = np.corrcoef(price, total_tweets)
co_protweets = np.corrcoef(price, pos_tweets)
print(co_tweets)
print(co_protweets)
```

```
[[1.         0.77233406]
 [0.77233406 1.        ]]
[[1.         0.77041387]
 [0.77041387 1.        ]]
```

Figure 5: The Pearson correlation coefficient (ρ) between the number of tweets with sentiment/ positive sentiment and the floor price of Doodles NFT

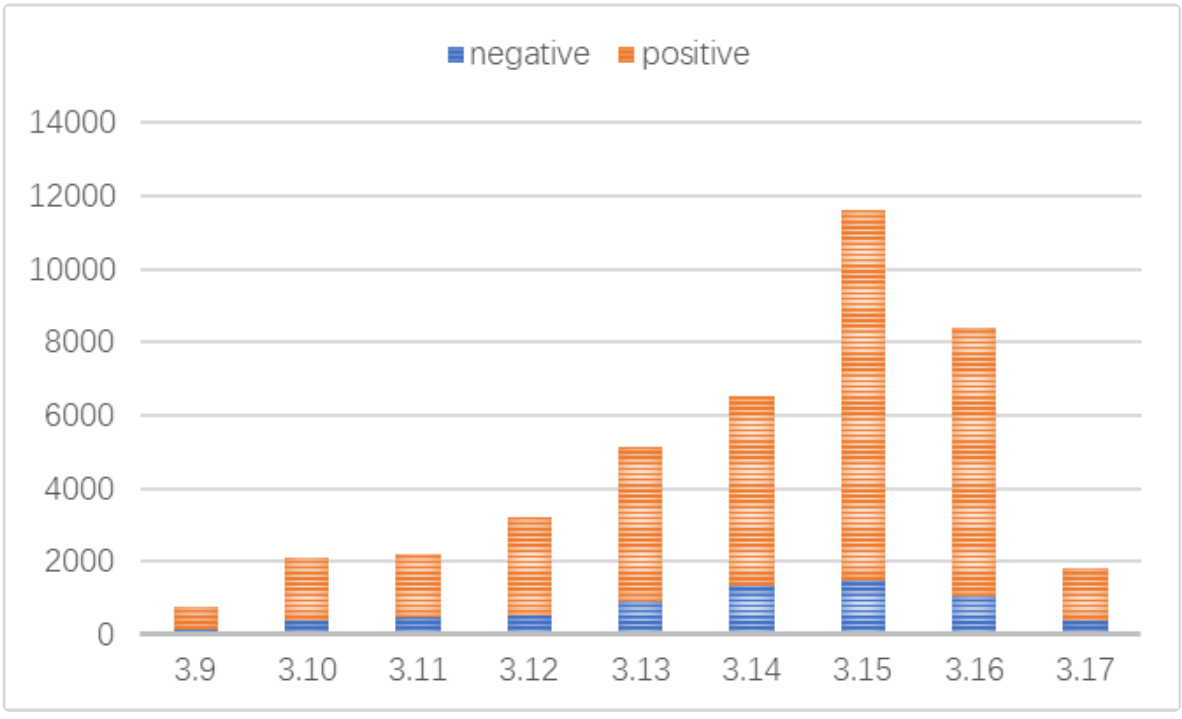

Figure 6: The number of words related to positive and negative sentiment between 9th March and 17th March.

Wordcloud

We created a word cloud from all of the data and then classified the words displayed. Four major groups of terms were employed in total: general words, event words, keyword words, and name words. Due to the massive quantity of data collected, we hope to analyze the project community's daily word cloud to reach more objective and detailed conclusions.

Figure 7: Wordcloud for all data

Combining the total word cloud with the daily word cloud, RT dominates the word cloud, reflecting the fact that the Doodles community has maintained a fairly high level of discussion. Additionally, the other evident terms in all of the word clouds are SHOPIFY and GIVING AWAY, two subsidiaries of SHOPIFY, a Canadian multinational e-commerce firm that collaborated on the project with doodles. This commercial engagement prompted the Doodles community to become more active and greatly increased the frequency with which they retweeted Doodles-related messages on Twitter, which almost certainly resulted in some good community conversation and created buzz for both shopify and Doodles. The vocabulary of SXSW, a major offline physical exhibition of Doodles, became particularly evident during the two days of 12-13 March. When Doodles turned virtual NFT images into real art on display, visitors were provided with an unparalleled immersive experience. This event, which pulls the Doodles community from the online virtual world into reality, may have stimulated a degree of community cohesion.

Figure 8: Wordcloud for 12th March

Figure 9: Wordcloud for 13th March

Apart from that, RTs, as indicated by retweets, continued to dominate the word cloud for the duration of the day, but were joined by a few new terms. For instance, HOOMANETH, ASTIN, and MIKEVAYNERCHUK are all terms that symbolize the NFT scene's netroots on Twitter. The three founders of doodles are BURNTTOAST, EVANKEAST, and POOPIE. These terms reflect the NFT community's proximity to the influencer on Twitter, and the influencer's material may affect the direction of the doodles community. Additionally, we observe multiple high-quality terms associated with NFT projects in the word cloud. COOLCATS, BOREDAPEYC, and AZUKIZEN are just a few examples. This may symbolize the relationship between the DOODLE and other NFT communities.

Wordclouds have shown to be quite useful in the NFT project's continuing study. Previously, it was believed that digital currencies, NFT, and other related fields have a speculative performance. However, upon closer examination, particularly with the use of applicable tools such as NLP, we discover that behind the surface phenomena of pyramid plans, there is considerable actual development on the part of the project partners. For instance, organising offline exhibits, collaborating on projects, and so on. While we may not be able to reach a final conclusion on the true worth of these behaviours, we will learn more if we withhold judgement and continue investigating and analysing.

This report used tools such as NLP to illustrate and analyse the optimism level among the Doodles community towards the project. Sentiment analysis and word clouds indicate that Shopify's twitter sweepstakes has stimulated some activities in the Doodles community. Additionally, the Doodles event at SXSW received much appreciation from the community.

Conclusion

These are likely to have contributed to the project's floor price. However, there are some limitations. Due to data availability, the data set for the report only contains tweets from 9th March to 17th March. Besides, the majority of discussions on the Doodles community are among people who have some knowledge of the subject. Therefore, the sentiment reviews in the twitter do not fully reflect the public sentiment towards NFT. The degree to which Doodles are influenced by community emotion has to be confirmed.

Reference

Aharon, D. Y., & Demir, E. (2021). NFTs and asset class spillovers: Lessons from the period around the COVID-19 pandemic. Finance Research Letters, 102515.

Ante, L. (2021). The non-fungible token (NFT) market and its relationship with Bitcoin and Ethereum. Available at SSRN 3861106.

Dowling, M. (2022). Is non-fungible token pricing driven by cryptocurrencies?. Finance Research Letters, 44, 102097.

Kapoor, A., Guhathakurta, D., Mathur, M., Yadav, R., Gupta, M., & Kumaraguru, P. (2022). TweetBoost: Influence of Social Media on NFT Valuation. *arXiv preprint arXiv:2201.08373*.

Nadini, M., Alessandretti, L., Di Giacinto, F., Martino, M., Aiello, L. M., & Baronchelli, A. (2021). Mapping the NFT revolution: market trends, trade networks, and visual features. Scientific reports, 11(1), 1-11.

Penman, S. H., Richardson, S. A., & Tuna, I. (2007). The Book-to-Price Effect in Stock Returns: Accounting for Leverage. *Journal of Accounting Research, 45*(2).

Wang, Q., Li, R., Wang, Q., & Chen, S. (2021). Non-fungible token (NFT): Overview, evaluation,

opportunities and challenges. arXiv preprint arXiv:2105.07447.