Natural Language Processing Mini Project Report

**Analyzing Cryptocurrency Trends and Sentiments through Reddit Comments**

Domain: Sentiment Analysis and Topic Modelling

Team Members

Aditi Bhatia D17A 06

Namrata Avhad D17A 03

Bhavesh Bhatia D17A 07

**Problem Statement:**

The cryptocurrency market's dynamic and decentralized nature presents challenges in understanding market sentiment and trends. Traditional financial indicators often fall short in capturing the complexities of this emerging market. The problem at hand is the lack of a reliable method to analyze and interpret the sentiments expressed within the vast volume of Reddit comments discussing cryptocurrencies. This poses a barrier to informed decision-making for traders, investors, and researchers who seek insights into the market's trajectory and community sentiment.

The challenge includes the need to process and comprehend a large and noisy dataset of Reddit comments, accurately gauge the sentiment and emotions behind these comments, and establish connections between the sentiments expressed and actual market movements. Addressing these challenges would provide a valuable tool for individuals and entities navigating the cryptocurrency landscape, enabling them to anticipate market trends and make more informed decisions based on the collective sentiment of the online community.

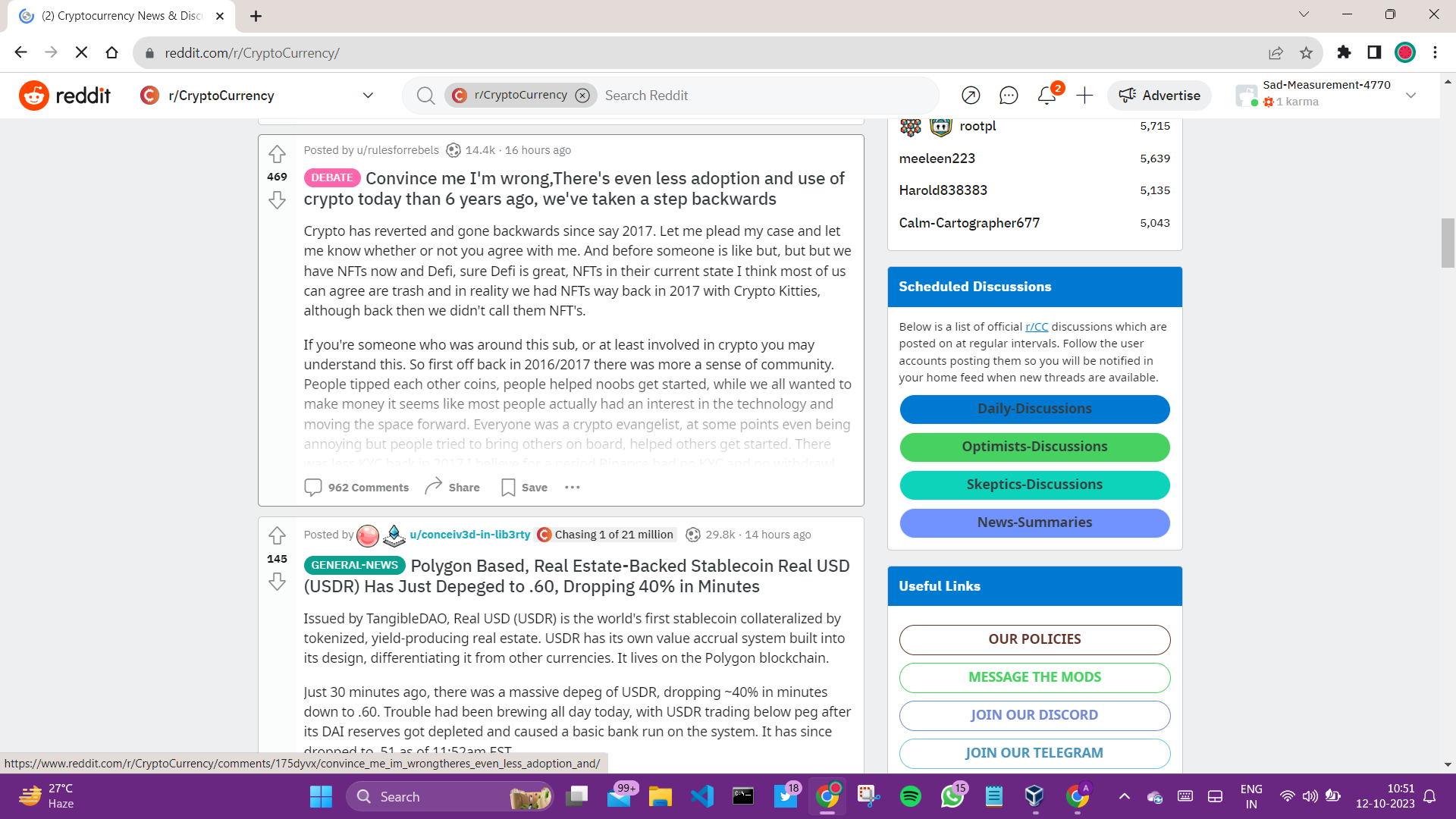
**Steps:**

1. Data Collection: Gathering over 17,000 Reddit comments from the r/Cryptocurrency subreddit using the PRAW API is a significant data collection effort. Focusing on top posts from specific years helps in narrowing down the dataset while still capturing relevant discussions.
2. Text Preprocessing: Preprocessing steps involve multiple layers of data cleaning and normalization, including removing links, punctuation, and non-alphabetical words. Applying lowercase conversion, stopword removal, stemming, and lemmatization enhances the quality of the text data.
3. Exploratory Data Analysis (EDA): EDA using techniques like bar graphs and word maps provides insights into the most common words and trends in the dataset. This process helps you better understand the data before delving into more complex analyses.
4. Topic Modeling: Implementing different topic modeling algorithms, including Latent Semantic Allocation (LSA), Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF), shows a thorough exploration of techniques to uncover underlying themes in the text data.
5. Topic Findings and Interpretation: Identifying and interpreting nine common topics within the cryptocurrency community is a substantial outcome. Understanding these topics can provide valuable information about what's driving discussions and interest.
6. Sentiment Analysis: Incorporating Vader Sentiment Analysis on uncleaned data is an additional layer of insight. This provides an overview of the overall sentiment within the text and helps gauge the community's mood.
7. Price Analysis of various cryptocurrencies based on the sentiments analysed of different comments analysed

**Methodology**

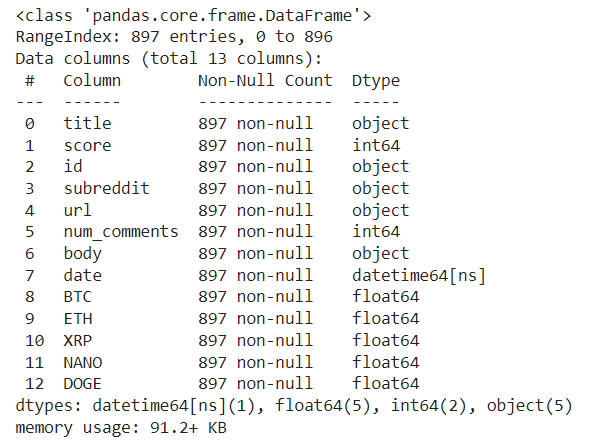
1. **Reddit’s PRAW API and r/Cryptocurrency community**

PRAW (Python Reddit API Wrapper) is a Python library that allows developers to interact with Reddit's API. It simplifies the process of accessing and working with Reddit's data, making it easier to create applications, automate tasks, and analyze content on the platform. PRAW provides a convenient way to access and manipulate Reddit's vast amount of user-generated content.



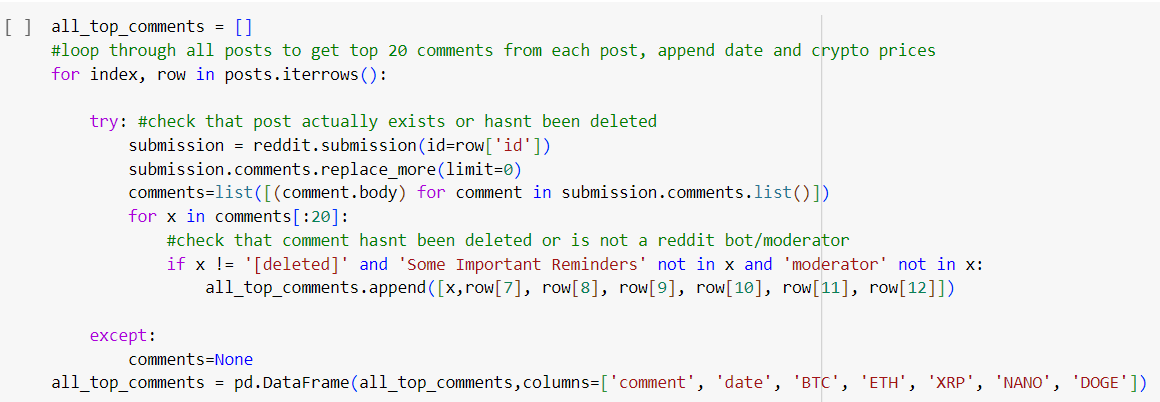
1. **Get open day prices for BTC, ETH, DOGE, NANO, and XRB based on date of post**

To do so, we have used the CryptoCompare API which is a popular cryptocurrency data provider offering a wide range of cryptocurrency market data. It provides access to real-time and historical pricing, trading volume, market capitalization, and various other metrics for numerous cryptocurrencies and exchanges. Developers and traders can use this API to integrate cryptocurrency data into their applications, perform market analysis, and monitor crypto-related activities. CryptoCompare's API is a valuable resource for staying updated on the dynamic world of cryptocurrencies and making informed investment decisions.



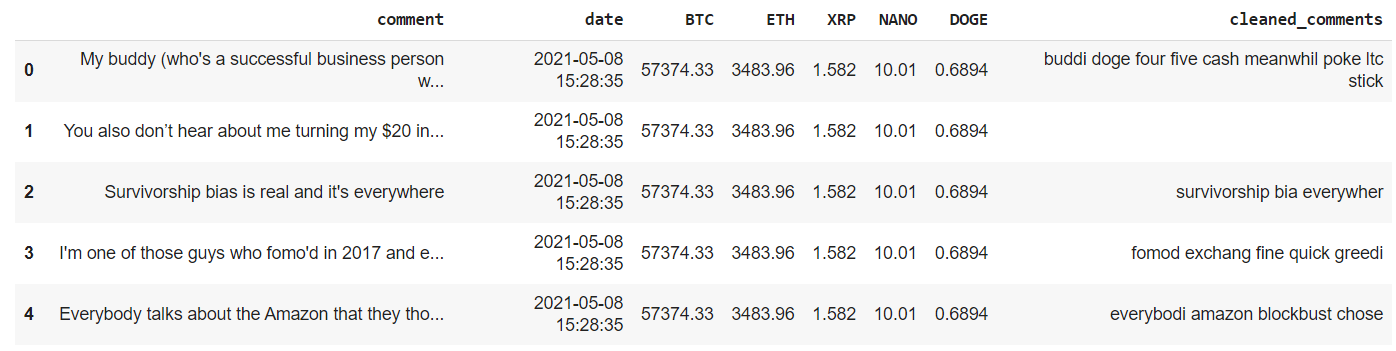
1. **Get top comments from each post**

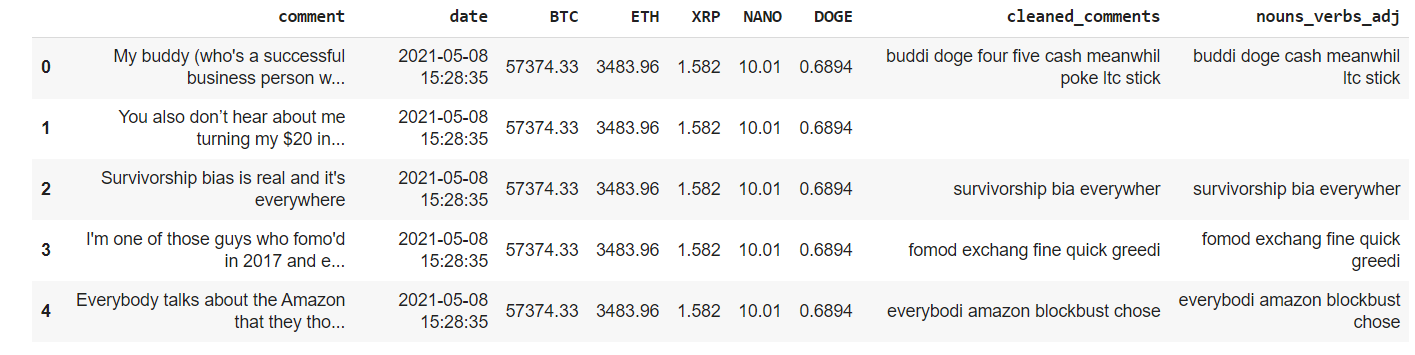
Loop through all posts to get top 20 comments from each post then append date and crypto prices



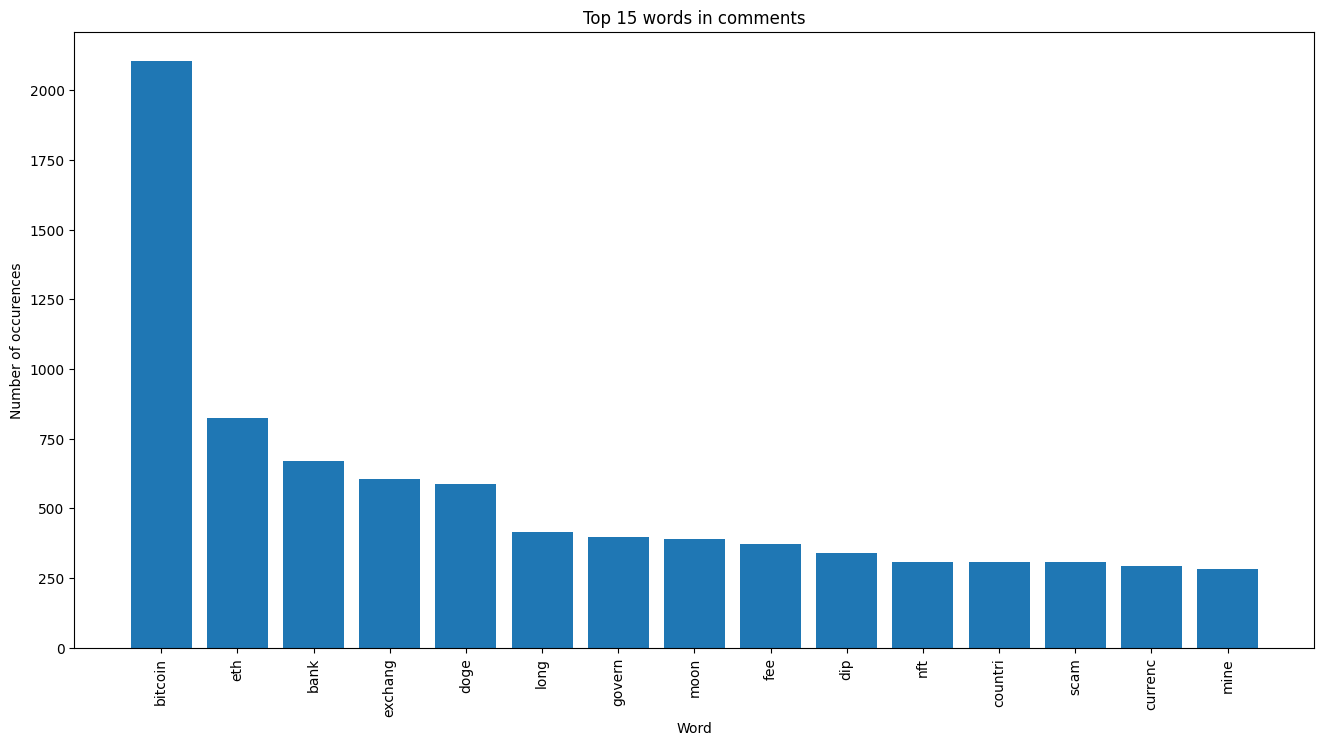
1. **Text Cleaning**

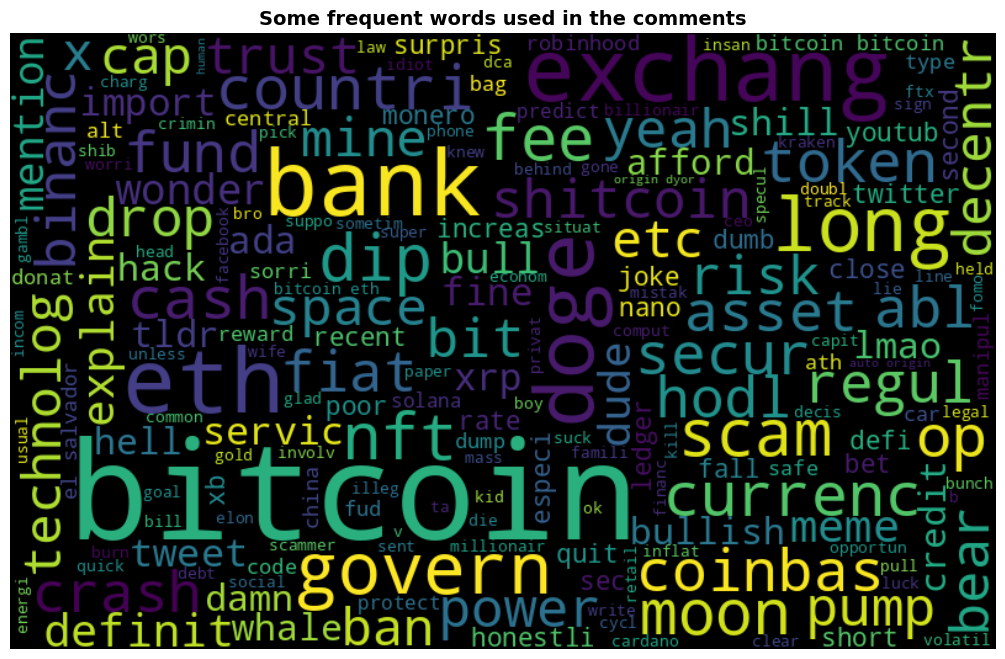
Stop word removal is a text preprocessing technique used in natural language processing (NLP). Stop words are common words like "the," "and," "in," which are often removed from text data to reduce noise and improve the efficiency of text analysis. By eliminating stop words, NLP models can focus on more significant words, enhancing the accuracy and performance of tasks like sentiment analysis, text classification, and information retrieval.

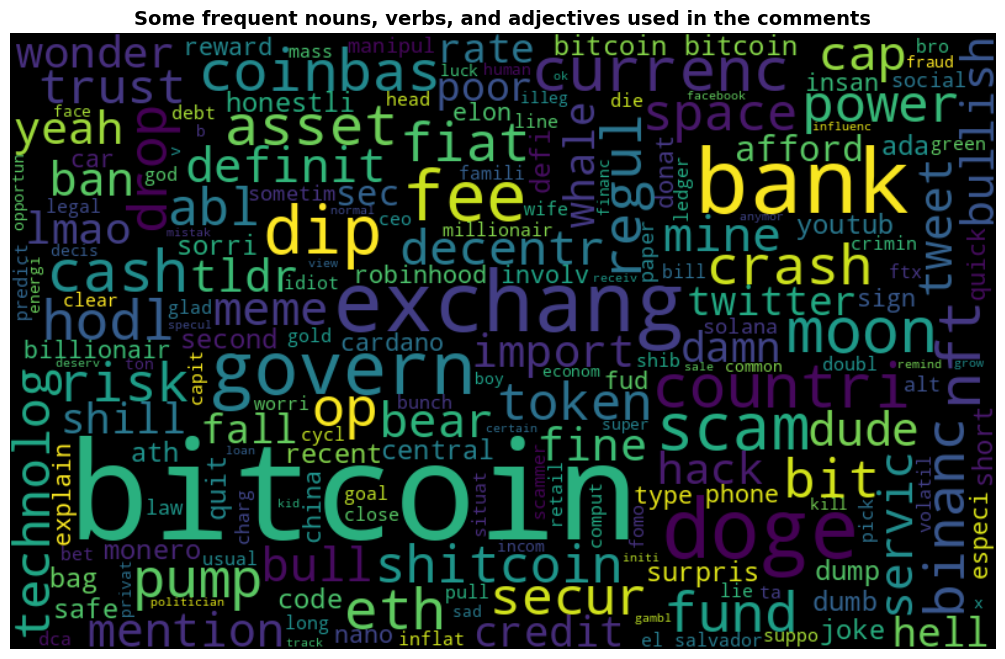


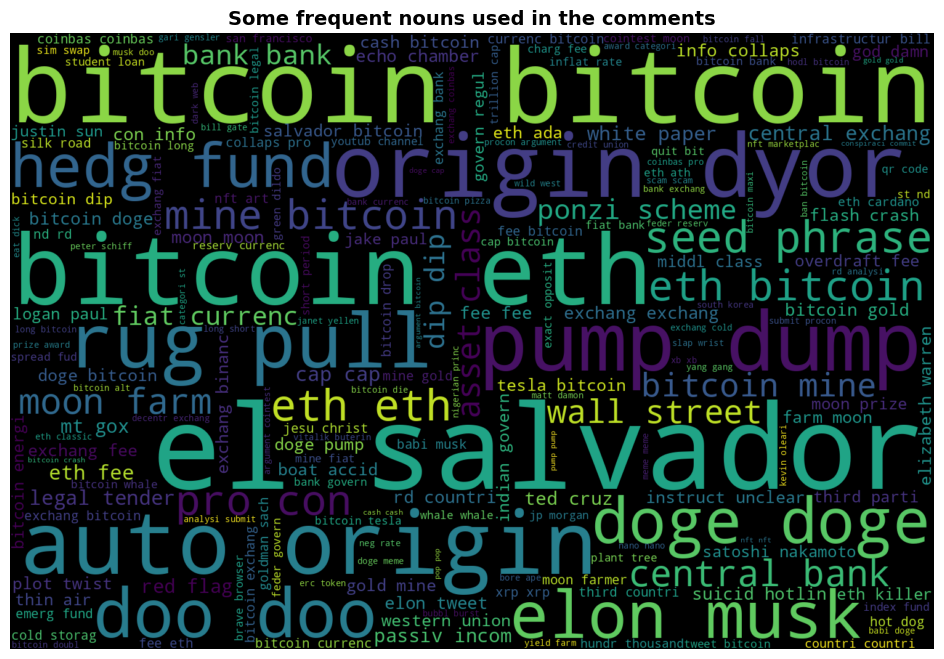
**Find all nouns, verbs and adjectives in the cleaned comments. **

1. **Exploratory Data Analysis**

****

****

****

****

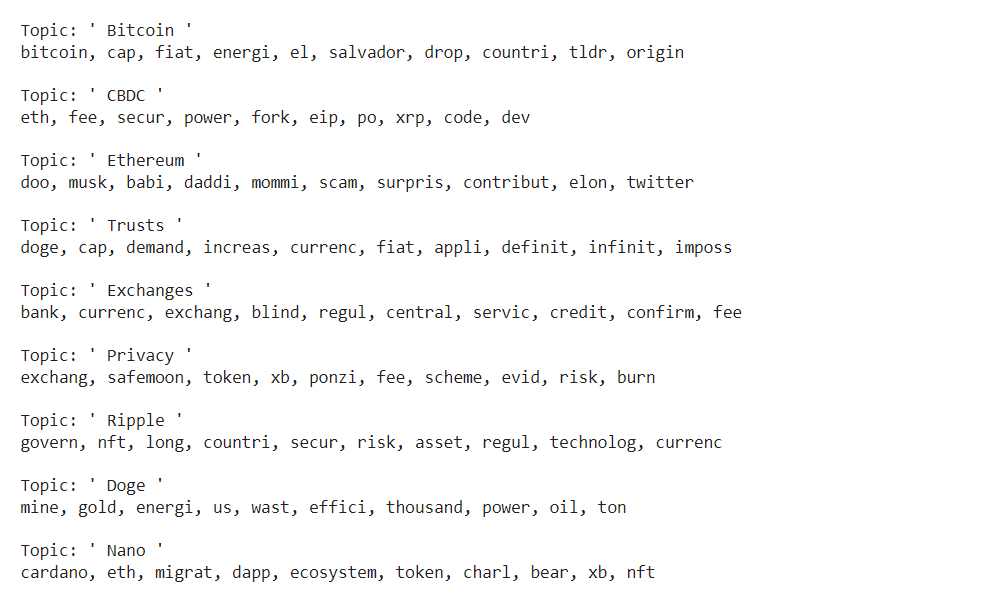
1. **Topic Modelling**

**Count Vectorization**, also known as the Bag of Words (BoW) model, is a technique used in natural language processing (NLP) to convert a collection of text documents into a numerical format. It involves creating a vocabulary of unique words in the corpus and counting the frequency of each word's occurrence in each document. This results in a matrix where rows represent documents, columns represent unique words, and the values in the matrix indicate word frequencies.

1. **Latent Semantic Analysis (LSA)** is a technique that's often applied in combination with Count Vectorization for text analysis. LSA aims to discover the latent semantic structure in a corpus by reducing the dimensionality of the Count Vectorization matrix using singular value decomposition (SVD). This process helps identify relationships between words and documents, making it possible to find semantically related terms and improve information retrieval, topic modeling, and document classification tasks. LSA can uncover hidden patterns in textual data and enhance the understanding of the underlying structure within a document collection.



1. **Non-Negative Matrix Factorization (NMF) with Count Vectorization**: Non-Negative Matrix Factorization (NMF) is a dimensionality reduction technique often applied in conjunction with Count Vectorization for text analysis. Unlike Latent Semantic Analysis (LSA), NMF enforces non-negativity constraints on both the factorized matrices, making the resulting factors more interpretable. NMF decomposes the Count Vectorization matrix into two lower-dimensional matrices representing topics and term distributions. This approach is particularly useful for topic modeling, document clustering, and uncovering meaningful patterns in text data.

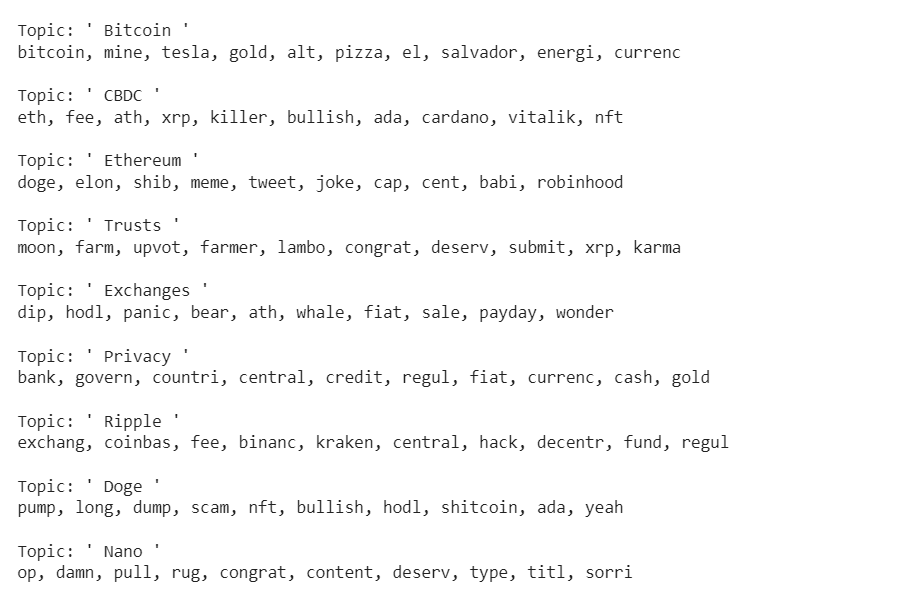


1. **Latent Dirichlet Allocation (LDA)** is a probabilistic topic modeling technique frequently used with Count Vectorization in natural language processing. LDA assumes that each document in a corpus is a mixture of topics, and each topic is a mixture of words. It utilizes Count Vectorization to represent the frequency of words in documents. LDA then aims to identify these underlying topics and their word distributions by iteratively estimating parameters. It's a valuable tool for discovering the thematic structure of a text corpus, aiding in tasks like content categorization, document clustering, and identifying prevalent topics within a dataset.



**TF-IDF (Term Frequency-Inverse Document Frequency) Vectorization** is a text preprocessing technique used in natural language processing. It represents the importance of words in a document within a corpus by considering two factors: term frequency (TF), which measures how often a word appears in a document, and inverse document frequency (IDF), which measures the rarity of the word across the corpus. TF-IDF assigns higher weights to words that are frequent in a document but rare across the corpus, making it effective for information retrieval, text classification, and text mining tasks where you want to prioritize the significance of words in a document.

1. **Non-Negative Matrix Factorization (NMF) in conjunction with TF-IDF** (Term Frequency-Inverse Document Frequency) Vectorization is a powerful combination in natural language processing. NMF, by enforcing non-negativity constraints, decomposes the TF-IDF matrix into two lower-dimensional matrices representing topics and term distributions, providing a more interpretable representation of text data. This approach is particularly valuable for topic modeling, document clustering, and feature extraction from text documents while maintaining non-negativity, which helps in making the factors more understandable and intuitive.



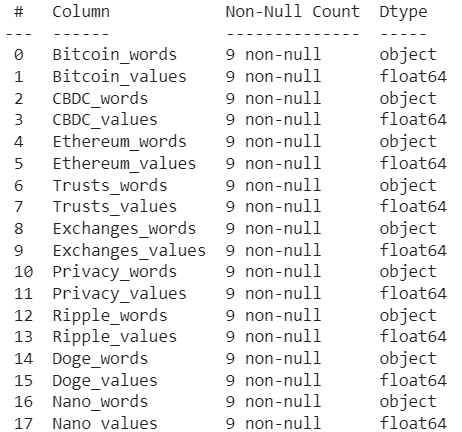
**Final Topic Model:**

**NMF with Count Vectorization got the most clear topics with a topic number of 9**

**Document Matrix:** The document-topic matrix is a numerical matrix where each row corresponds to a document, and each column represents a topic. The values in this matrix indicate the probability or weight of each document's association with each topic. This matrix helps to reveal the distribution of topics across a collection of documents, allowing you to understand which topics are prevalent in each document.

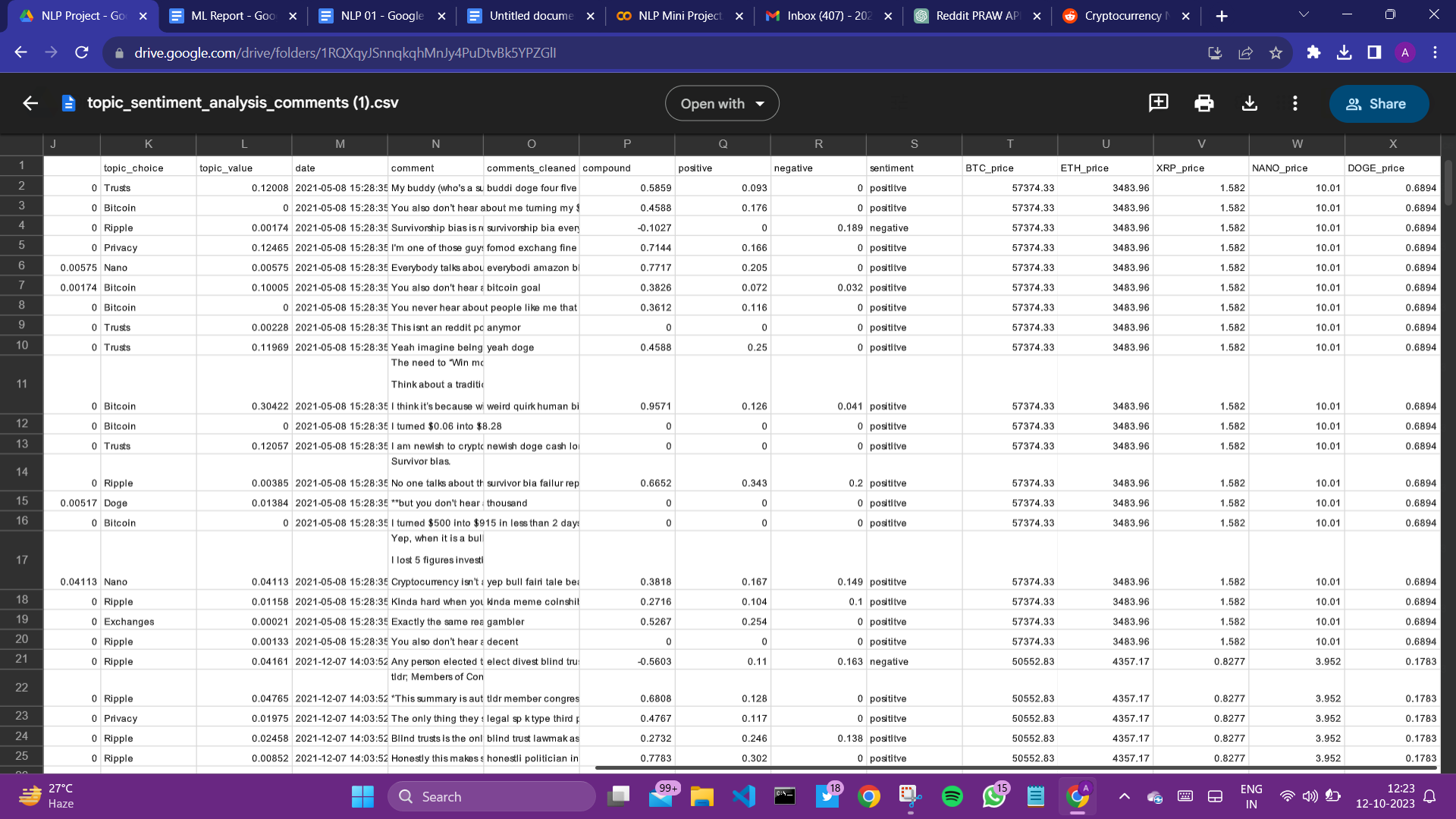


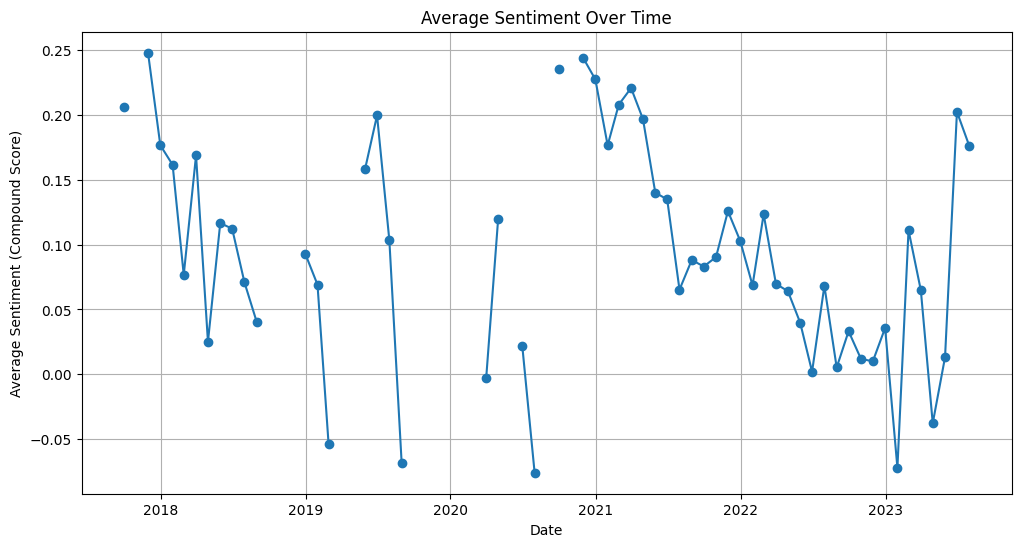
**Word Matrix:**  The word-topic matrix is another numerical matrix where each row corresponds to a word in the vocabulary, and each column represents a topic. The values in this matrix indicate the probability or weight of each word's association with each topic. This matrix allows you to explore how words are distributed across topics, revealing the words most relevant to specific topics in your corpus

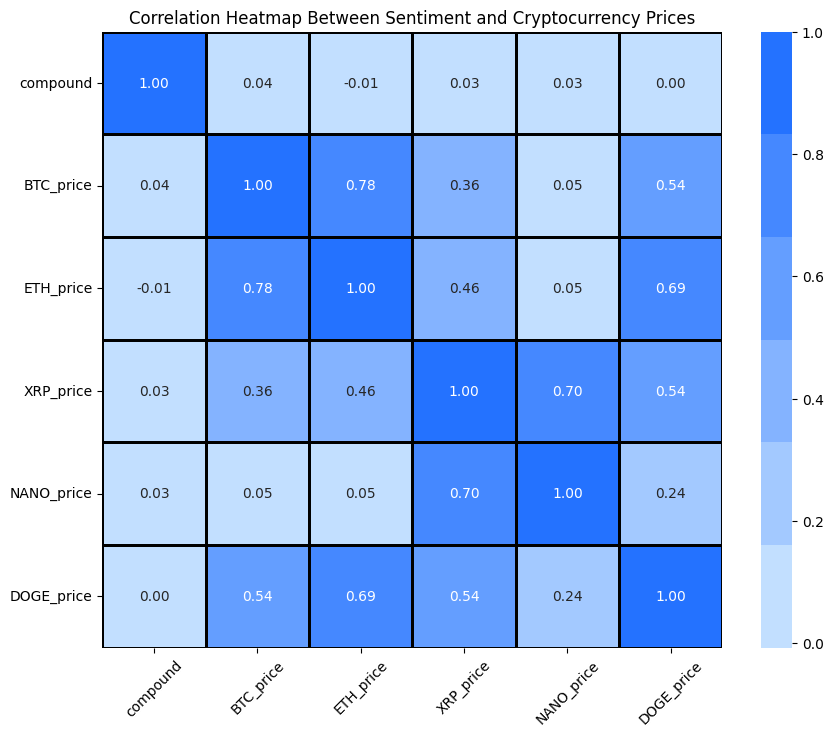


**Sentiment Analysis**

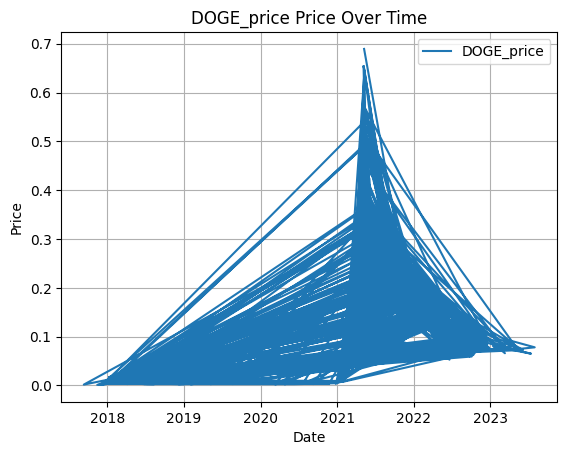
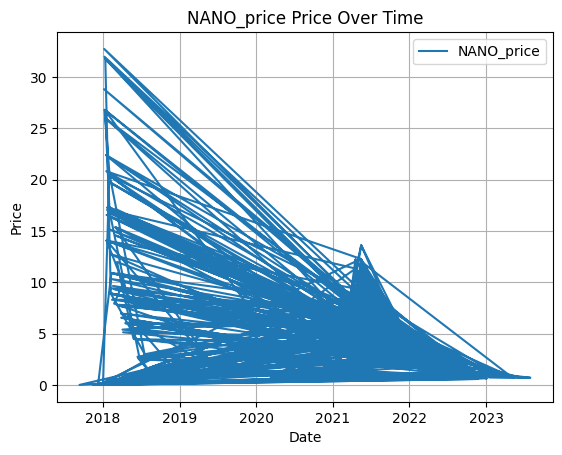
The VADER (Valence Aware Dictionary and sEntiment Reasoner) model is a pre-trained sentiment analysis tool widely used in natural language processing. It's specifically designed for social media text and considers not only polarity (positive, negative, neutral) but also intensity and sentiment compounds. VADER analyzes text by assigning a sentiment score to each word and combines them to determine the overall sentiment of a sentence or document. It's valuable for tasks like social media sentiment monitoring, product reviews, and understanding the emotional tone of text in a fast and context-aware manner.

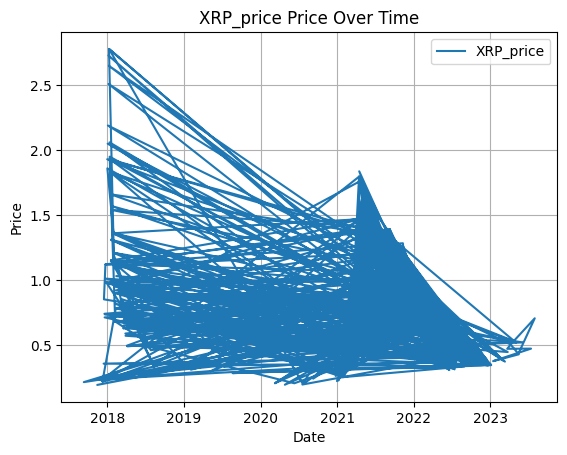
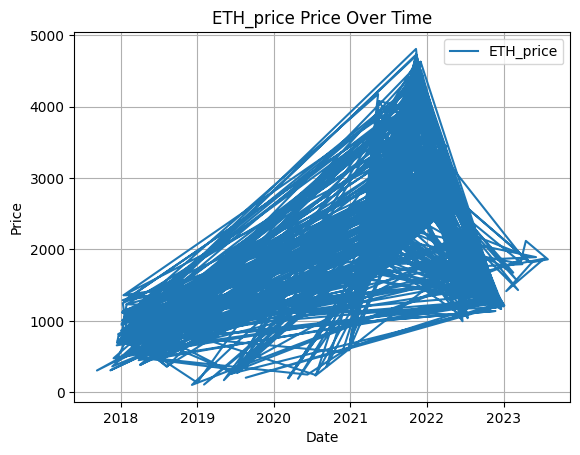


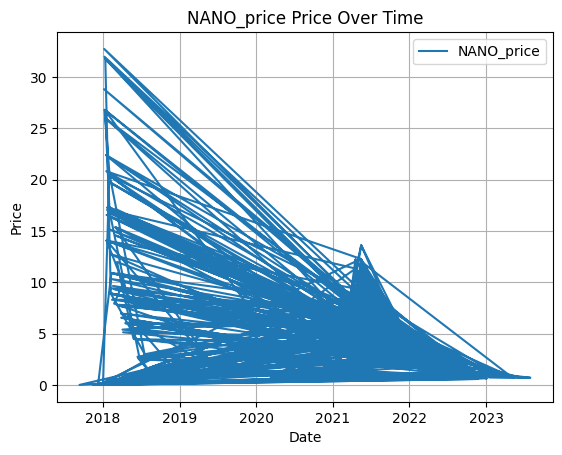


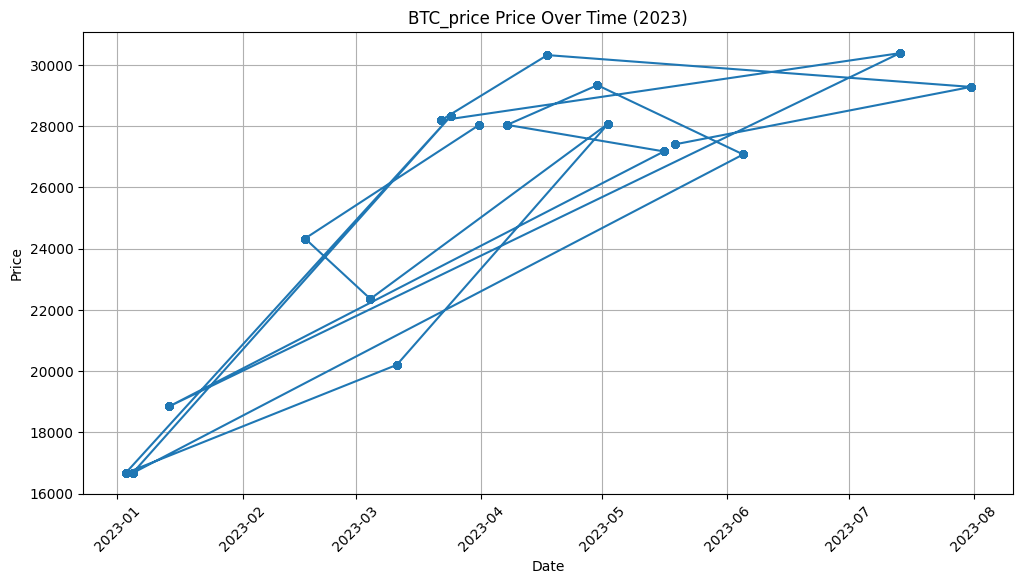


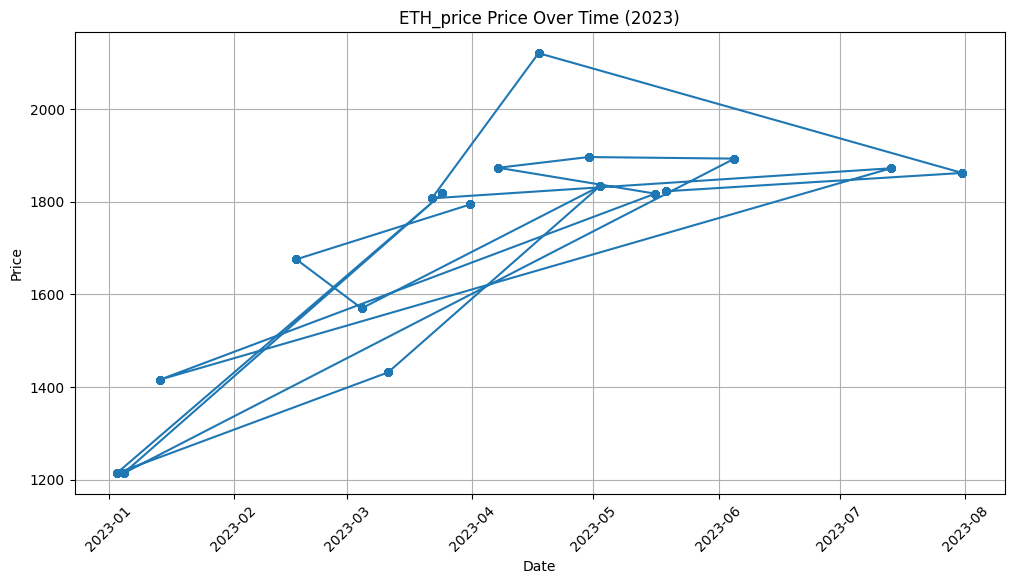
**Price of Cryptocurrencies over time**

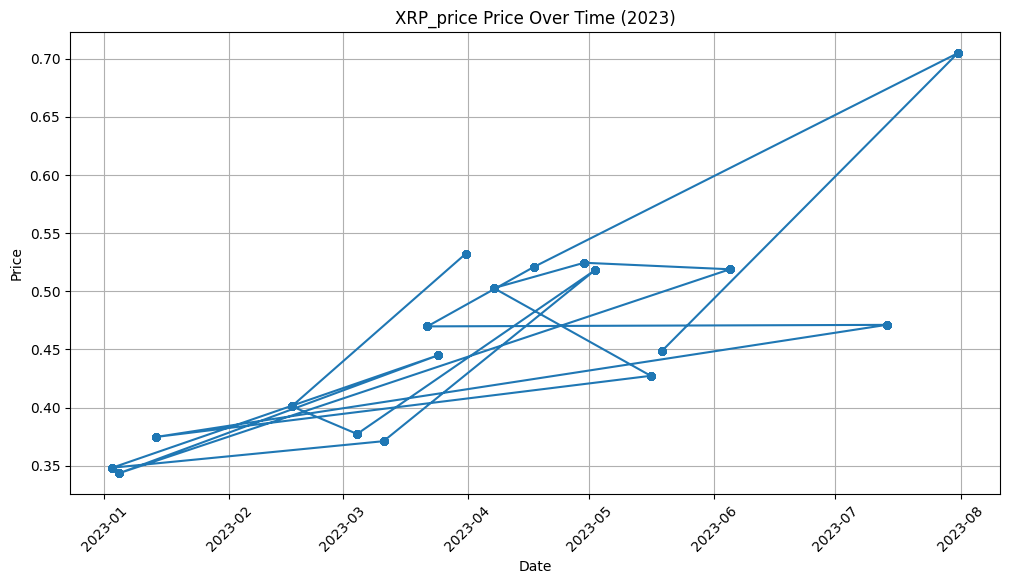
****

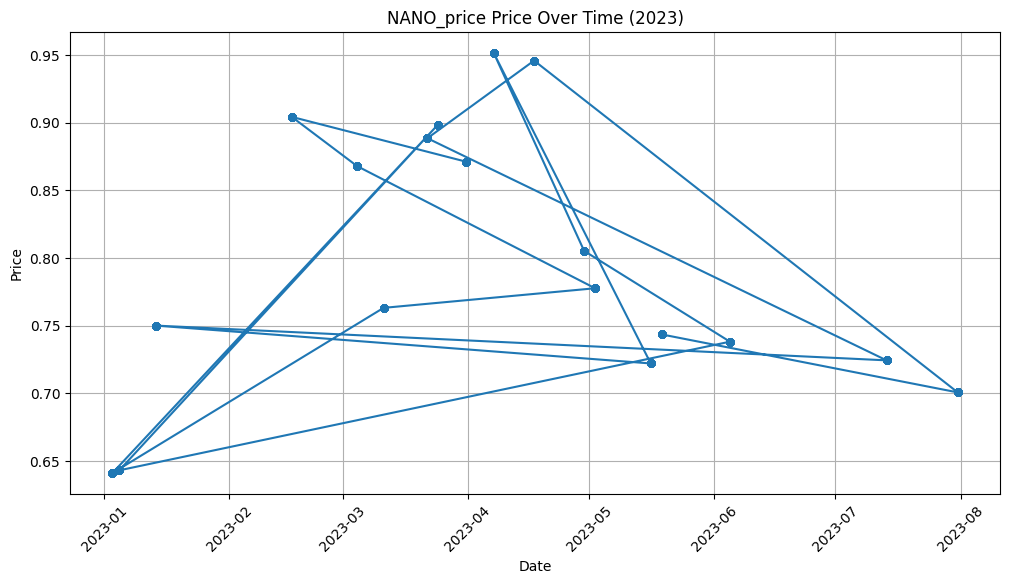
****

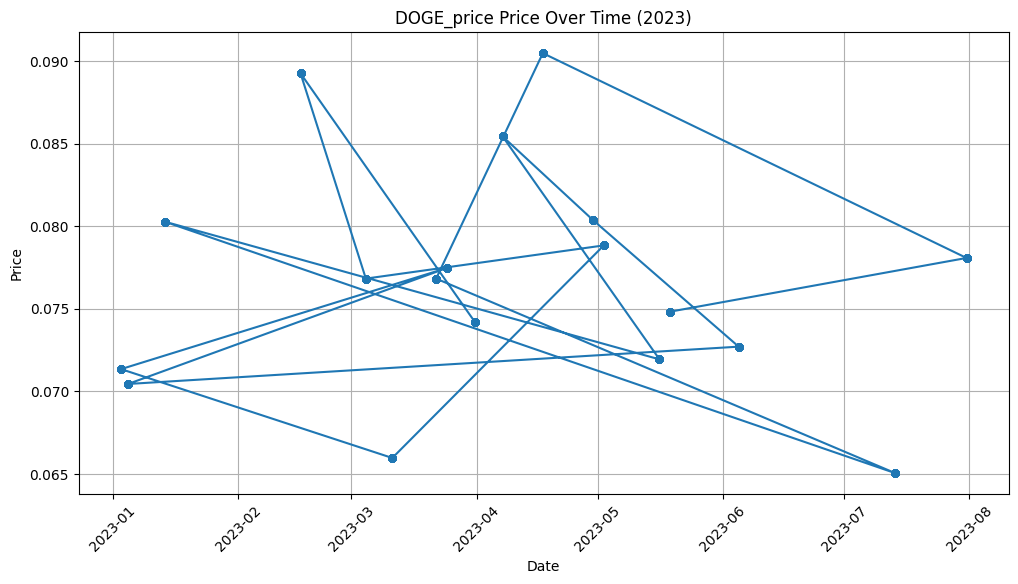
****

****

****

****

****

****

**Conclusion:**

In conclusion, the proposed cryptocurrency sentiment analysis project holds significant promise in uncovering invaluable insights from the dynamic and ever-evolving realm of digital currencies. By harnessing the power of natural language processing (NLP) techniques, advanced analytics, and an interdisciplinary approach, this project aspires to shed light on the sentiments, discussions, and trends within the cryptocurrency community.

The project's novelty is underscored by its deliberate focus on the cryptocurrency community, utilizing over 17,000 Reddit comments from the r/Cryptocurrency subreddit to capture the pulse of the industry. Furthermore, the integration of diverse techniques, from data collection through APIs and rigorous text preprocessing to advanced topic modeling and sentiment analysis, reflects a comprehensive methodology that is well-equipped to tackle the complex nature of cryptocurrency discourse. The temporal analysis of top posts from specific years and the correlation with cryptocurrency prices elevate the project's analytical prowess, enabling a deeper exploration of the interplay between sentiment and market movements.