OasisCoin: Bitcoin Price Movement Prediction Using Social Media and News Sentiment

Course: Natural Language Processing IST 332

Date: November 23, 2025

Authors: Nihaad Saleem, Varsha Ravindra Shetty, Prajwal Vinod Naik, Mahesh Balan

Part 1 – Introduction & Research Question

1.1 Research Problem and Motivation

Bitcoin prices are famously volatile and heavily influenced by both market developments and public sentiment expressed in news coverage. Traders, investors, and researchers seek ways to forecast price movements, and one promising avenue is using news sentiment analysis combined with market data. With barriers to collecting large-scale Twitter and Reddit data due to API costs and restrictions, we focused our efforts on leveraging Google News RSS feeds for a robust corpus and supplementing this with comprehensive price and market data from CoinGecko.

The core challenge we address is: given a large corpus of news articles about Bitcoin, can we extract meaningful sentiment and topic signals that correlate with or predict daily price changes? This requires not just collecting data, but cleaning messy text, analyzing patterns, and testing whether news truly precedes or coincides with market movements. Social media used to be a standard source for this research, but platform APIs have become prohibitively expensive or restrictive. By building a pipeline using Google News RSS and CoinGecko APIs, we create an accessible, repeatable method that any researcher or student can use.

1.2 Research Questions and Expected Contribution

Main Research Question: Can sentiment and topic data extracted from news articles predict day-to-day price movement for Bitcoin?

Supporting Questions:

- Does news sentiment show leading, coincident, or lagging relationships with Bitcoin price changes?

- Can simple text preprocessing (tokenization, lemmatization, stopword removal) capture the signals needed for meaningful prediction?

- How do corpus statistics (average article length, topic frequency, sentiment distribution) align with market trends over time?

- What are the comparative impacts of using large news datasets versus social media sources for price prediction?

- Can we identify key terms or phrases in news that reliably precede price spikes or crashes?

Expected Contribution:

- Assemble and analyze one of the most comprehensive Google News Corpora for Bitcoin (~ 25,000 records spanning 2017-2025).

- Establish a transparent, reproducible pipeline linking textual news properties to daily price movements using CoinGecko market data.

- Demonstrate effective text preprocessing tailored for financial news, including domain-specific vocabulary preservation.

- Publish corpus statistics, working code, and accessible methods suitable for academic and classroom use.

- Provide a modular framework for later integration with social media data (Twitter, Reddit) and continuously updated price/market metrics.

- Show that despite API access challenges, high-quality prediction research remains feasible using open data sources.

1.3 Why This Work is Relevant

Our research is relevant for multiple reasons.

Bitcoin now represents a multi-trillion-dollar asset class with mainstream institutional participation. Understanding price drivers is critical for regulators, traders, and researchers alike. News sentiment analysis is an established field, but applying it rigorously to crypto with large, clean datasets remains challenging.

From a practical perspective, our approach addresses a real bottleneck in the research community. Twitter and Reddit APIs that once allowed researchers to collect large-scale historical data now cost hundreds of dollars per month or require academic partnerships that can take months to establish. By demonstrating a pipeline using Google News RSS and CoinGecko APIs both accessible and free, we show that rigorous research is still possible.

Our project also serves as a teaching tool. The preprocessing, feature extraction, and exploratory data analysis techniques we employ are applicable far beyond crypto. Students learn how to handle messy real-world data, manage API rate limits, align multiple

data sources by timestamp, and perform exploratory analysis on text and time-series data simultaneously.

Finally, this work establishes a foundation for hybrid approaches. Once Reddit or Twitter data becomes available (whether through future API changes or publicly released datasets), our framework will seamlessly incorporate those sources. The same preprocessing and analysis code works across all text sources.

Part 2 - Corpus Creation (Data Collection)

2.1 Data Sources

Our data pipeline draws from two complementary sources:

Google News RSS Feeds for Text Data:

Google News provides RSS feeds of recent news articles filtered by keyword search. These feeds aggregate content from thousands of news outlets globally, covering major stories and niche coverage.

- Source: Google News RSS (https://news.google.com/)
- Search focus: “bitcoin” and related cryptocurrency keywords
- Geographic/language filter: English-language, global coverage
- Data type: Headlines, article snippets/descriptions, publication timestamps, source outlet names, article URLs
- Frequency: Can be polled continuously or on a schedule
- Why Google News: Covers major financial outlets (Reuters, Bloomberg, AP), specialized crypto media (CoinDesk, Cointelegraph), and mainstream press. Provides a broad perspective on how Bitcoin is discussed across different audiences.

CoinGecko API for Price and Market Data:

CoinGecko provides free access to historical cryptocurrency price data, market capitalization, trading volume, developer metrics, and community signals.

- Source: CoinGecko API (https://www.coingecko.com/en/api)
- Asset tracked: Bitcoin (symbol: BTC, ID: bitcoin)

- Metrics available: Daily Open, High, Low, Close price; trading volume; market cap; fully diluted valuation; circulating supply; sentiment indicators
- Time resolution: Daily granularity (closing prices and metrics at 00:00 UTC)
- Why CoinGecko: Free tier with no authentication required. Highly reliable and widely used by researchers and professionals. Provides ground truth for price labels (up/down predictions).

Future Data Sources (Contingent on Availability):

- Reddit historical data: If API access is approved or public datasets become available, we will collect posts from r/Bitcoin, r/CryptoCurrency, r/ethereum.
- Twitter/X historical data: Public archives.
- These sources will follow identical preprocessing pipelines and be temporally aligned with the same CoinGecko price data.

2.2 Data Collection Methodology

Our data pipeline draws from two complementary sources:

Google News RSS Feeds for Text Data:

Google News provides RSS feeds of recent news articles filtered by keyword search. These feeds aggregate content from thousands of news outlets globally, covering major stories and niche coverage.

- Source: Google News RSS (https://news.google.com/)
- Search focus: “bitcoin” and related cryptocurrency keywords
- Geographic/language filter: English-language, global coverage
- Data type: Headlines, article snippets/descriptions, publication timestamps, source outlet names, article URLs
- Frequency: Can be polled continuously or on a schedule
- Why Google News: Covers major financial outlets (Reuters [finance:Reuters], Bloomberg, AP), specialized crypto media (CoinDesk, Cointelegraph), and mainstream press. Provides a broad perspective on how Bitcoin is discussed across different audiences.

CoinGecko API for Price and Market Data:

CoinGecko provides free access to historical cryptocurrency price data, market capitalization, trading volume, developer metrics, and community signals.

- Source: CoinGecko API (https://www.coingecko.com/en/api)
- Asset tracked: Bitcoin (symbol: BTC, ID: bitcoin)
- Metrics available: Daily Open, High, Low, Close price; trading volume; market cap; fully diluted valuation; circulating supply; sentiment indicators
- Time resolution: Daily granularity (closing prices and metrics at 00:00 UTC)
- Why CoinGecko: Free tier with no authentication required. Highly reliable and widely used by researchers and professionals. Provides ground truth for price labels (up/down predictions).

CoinCompare API for Supplementary Price and Market Data:

CoinCompare provides comprehensive historical and real-time cryptocurrency price data with higher rate limits than CoinGecko, making it ideal for larger-scale data collection and continuous monitoring. The platform offers aggregated pricing from multiple exchanges, providing additional robustness and validation opportunities.

- Source: CoinCompare API (https://www.cryptocompare.com/api)
- Asset tracked: Bitcoin (symbol: BTC, ID: BTC)
- Metrics available: Daily Open, High, Low, Close price; trading volume; market cap; exchange-aggregated data; historical OHLCV with minute-level granularity; volume data across multiple exchanges; technical analysis indicators
- Time resolution: Multiple granularities available - minute-level, hourly, daily (closing prices at 00:00 UTC)
- Why CoinCompare: Generous free tier with higher rate limits than CoinGecko (2000 calls/hour vs CoinGecko's 50 calls/minute). Requires free API key registration, but no payment required for academic/research use. Data aggregated from 30+ exchanges, reducing exchange-specific bias. Provides alternative price validation by comparing multiple data sources.

Integration Strategy for Dual-Source Price Data:

By combining CoinGecko and CoinCompare, we achieve:

- **Redundancy and Validation:** Cross-check price data between sources to detect anomalies or exchange-specific distortions
- **Rate Limit Resilience:** If one API hits rate limits, seamlessly switch to or supplement with the other

- **Data Quality Assurance:** Average prices from both sources for final analysis, reducing susceptibility to single-source errors
- **Continuous Collection:** CoinCompare's higher rate limits enable more frequent polling during critical periods (major news events, volatility spikes)

Both APIs will be called with identical parameters (Bitcoin, daily granularity, USD denomination) and merged by timestamp. Discrepancies between sources will be logged and investigated, potentially revealing exchange-specific trading patterns or data quality.

Future Data Sources (Contingent on Availability):

- Reddit historical data: If API access is approved or public datasets become available, we will collect posts from r/Bitcoin, r/CryptoCurrency, r/ethereum.
- Twitter/X historical data: Public archives or academic partnerships if feasible.
- These sources will follow identical preprocessing pipelines and be temporally aligned with the same CoinGecko price data.

2.3 Supplementary Public Data Sources for Social Media

To strengthen our corpus and enable comparison between news and social media sentiment, we have identified several high-quality public datasets that can be integrated into our pipeline: We are still researching which of these would be well suited for our project but wanted to present our research and findings so far.

Historical Bitcoin Twitter Datasets:

1. **Kaggle: Bitcoin Tweets (2016-2019)**
    - Source : https://www.kaggle.com/alaix14/bitcoin-tweets-20160101-to-20190329
    - Coverage: January 2016 to March 2019
    - Size: Hundreds of thousands of tweets mentioning Bitcoin [finance:Bitcoin]
    - Format: CSV with tweet text, timestamp, user metadata
    - Use case: Historical baseline for comparing news vs social media sentiment
2. **Kaggle: Bitcoin Sentiment Analysis Twitter Data**
    - Source : https://www.kaggle.com/datasets/gautamchettiar/bitcoin-sentiment-analysis-twitter-data
    - Coverage: Recent Bitcoin [finance:Bitcoin]-related tweets
    - Features: Pre-processed tweet text suitable for sentiment analysis
    - Use case: Ready-to-analyze dataset for sentiment modeling
3. **Kaggle: Bitcoin Tweets Dataset (100K+ tweets)**

- Source: https://www.kaggle.com/datasets/alishafaghi/bitcoin-tweets-dataset
- Size: Over 100,000 tweets with Bitcoin hashtag
- Features: Date, tweet link, tweet text, profile handle
- Coverage: Worldwide tweets
- Use case: Large-scale sentiment and topic analysis

4. **GitHub: Bitcoin Twitter Sentiment (1M tweets, 2021)**
   - Source: https://github.com/ntdoris/bitcoin-twitter-sentiment
   - Coverage: 1 million tweets from February to August 2021
   - Features: Pre-labeled sentiment (positive/negative)
   - Use case: Pre-labeled training data for sentiment classifiers
5. **NIH Database: Twitter Influencers in Cryptocurrency (2021-2023)**
   - Source: PMC11470647
   - Coverage: 52 cryptocurrency influencers, 300+ cryptocurrencies
   - Period: February 2021 to June 2023
   - Features: Tweets, sentiment scores, polarity, importance coefficients
   - Use case: Influencer impact analysis on price movements

Historical Bitcoin Reddit Datasets:

1. **OpenDataBay: Reddit Bitcoin Comments Dataset**
   - Source: https://www.opendatabay.com/data/ai-ml/afb22b14-6266-47ec-be7f-c936582d61ab
   - Coverage: Early 2020 to present
   - Subreddit: r/Bitcoin
   - Features: Title, score (upvotes), comment text, number of replies, timestamp
   - Size: Tens of thousands of comments
   - Use case: Long-form discussion analysis and sentiment tracking
2. **Zenodo: Reddit r/cryptocurrency Posts and Comments (2021-2022)**
   - Source: https://zenodo.org/doi/10.5281/zenodo.12593439
   - Coverage: January 2021 to December 2022
   - Subreddit: r/cryptocurrency
   - Size: 2.6 MB of posts, 1.3 MB of comments
   - Features: Date, comment text, engagement metrics
   - Includes: Aligned Bitcoin market data
   - Use case: Market-aligned social sentiment analysis
3. **arXiv: PulseReddit Dataset (2024-2025)**
   - Source: https://arxiv.org/html/2506.03861v1

- Coverage: April 2024 to March 2025
- Subreddits: r/Bitcoin, r/ethereum, r/dogecoin, r/solana, r/binance, r/pepecoin
- Features: Posts, comments, sentiment, high-frequency market data (5-minute to 4-hour intervals)
- Use case: High-frequency trading signal analysis

Combined/Multi-Platform Datasets:

1. **The Tie: Social Media & Sentiment Data**
   - Source : https://www.thetie.io/data/social-media/
   - Coverage: 900+ cryptocurrencies, 7+ years of historical data
   - Platforms: Twitter, Reddit, others
   - Features: Volume metrics, sentiment measures, engagement data
   - Access: Commercial platform with historical API access
   - Use case: Professional-grade cross-platform sentiment tracking
2. **LunarCrush: Social Media Analytics**
   - Source: https://lunarcrush.com
   - Coverage: Real-time and historical social media data
   - Features: Sentiment, engagement, trend tracking
   - Use case: Multi-platform sentiment aggregation

Integration Strategy:

For each public dataset, we will:

1. Download and validate data quality

2. Parse timestamps and align with CoinGecko price data by date

3. Apply our preprocessing pipeline (same as news articles)

4. Compute sentiment scores using TextBlob/VADER

5. Create daily aggregates (average sentiment, post volume, engagement)

6. Merge with news data to create multi-source feature sets

Advantages of Public Datasets:

- No API costs or rate limits

- Historical coverage (some dating back to 2016)

- Pre-validated and cleaned by research community

- Reproducible (anyone can download same data)

- Enables longitudinal studies across multiple years

Limitations:

- Data may not extend to most recent dates (2024-2025)

- Some datasets require citation and usage restrictions

- Quality varies (some have pre-labeled sentiment, others require processing)

- Coverage gaps between different time periods

By combining our ~25,000 Google News articles with these public Twitter and Reddit datasets, we can achieve:

- Broader temporal coverage (2016-2025)

- multi-platform comparison

- Richer feature sets for prediction models

- Validation of news-only vs social media sentiment signals

2.4 Data Collection Methodology

Phase 1: Google News RSS Collection

Step 1: Set up RSS feed polling

- Configure a Python script using the feedparser library to retrieve Google News RSS feed for “bitcoin” search query

- RSS feed URL: https://news.google.com/rss/search?q=bitcoin&hl=en&gl=US&ceid=US:en

- Can expand with multiple search queries (e.g., “bitcoin AND crash”, “bitcoin AND regulation”, “bitcoin AND bull”)

Step 2: Extract article metadata

For each article, we capture:

- Headline (title)

- Article description/snippet

- Publication timestamp

- Source outlet name

- Article URL (for verification and traceability)

Step 3: Deduplication and storage

- Store raw articles in a CSV file with one row per unique article

- Deduplicate by URL or by hash of headline+source+date to remove duplicate coverage

- Expected result: 23,000 unique articles covering 2017-2025

Step 4: Temporal alignment

- Parse publication timestamps and convert to date (YYYY-MM-DD) for daily aggregation

- Group articles by day to compute daily statistics (article count, average length, sentiment by day)

Phase 2: CoinGecko Price Data Collection

Step 1: Query the CoinGecko API for daily Bitcoin data

- Endpoint: /coins/bitcoin/market_chart/range

- Parameters: Start date, end date, currency (USD), daily granularity

- Retrieve: Daily close price, market cap, trading volume

Step 2: Create daily price labels

- For each day, calculate if the next day's close price was higher (1) or lower (0) than today's close

- This binary label becomes the prediction target

- Days near the end of the observation period may have missing next-day labels

Step 3: Store and align

- Save price data as a CSV with one row per day

- Columns: date, open, high, low, close, market_cap, volume, price_direction

- Merge with news data by date for temporal alignment

Phase 3: Combine and validate

Step 1: Merge news and price data

- Join on date field

- News data will have multiple records per date; price data has one record per date

- Compute daily aggregates: article count, average sentiment, dominant topics by day

Step 2: Validate coverage

- Confirm no gaps in price data

- Confirm news articles span the full date range

- Check for any data quality issues (missing values, out-of-range values, etc.)

Step 3: Save combined corpus

- Output: Single CSV or set of CSVs ready for preprocessing and analysis

2.5 Ethical Considerations and Permissions

Data Source Compliance: Google News RSS feeds are publicly accessible and explicitly provided by Google for content aggregation.

- CoinGecko APIs are free and openly documented; usage follows their terms of service.

- No credentials are required; no authentication tokens and we will conform to their specified rate limits.

- Public Twitter/Reddit datasets follow platform terms of service and academic fair use.

Data Privacy:

- News articles are already public; we do not collect or republish full article text, only headlines and summaries.

- Social media datasets contain only public posts; no private messages or restricted content.

- No personal identifiable information is collected or processed.

- Source attribution is preserved (outlet name, publication date, URL).

- We collect only publicly posted information. No private messages, deleted content, or non-public data.

- We do not attempt to re-identify users or link posts to real-world individuals beyond public usernames.

- We store data securely and do not share raw datasets publicly if they contain redistributable content restrictions.

Academic Use:

- This project is for educational purposes (course coursework) and future academic research.

- All code and methods are reproducible and will be made available.

- Findings and insights will be properly cited and attributed.

- Public datasets are cited and credited to original creators.

Responsible Use of Results:

- We do not claim that news sentiment provides perfect price predictions. **This is for research and education purposes only.**

- Results should not be interpreted as financial advice or trading signals.

- Limitations and sources of error are acknowledged throughout.

2.6 Summary Statistics

News Corpus (Google News RSS):

- Total articles collected: 25,648(target achieved)

- Total articles after deduplication: 14,456

- Articles removed: 11,192

- Date range: 2017-10-01 to 2025-11-24 (2976 days)

- Articles per day (average):5.73

- Articles per day (min):1

- Articles per day (max):89

- Unique news sources:14,445

Article Content Metrics:

- Average headline length (characters):80.69

- Average headline length (words):13.00

- Median headline length (words):13.00

- Average description length (words): N/A

Part 3 - Text Preprocessing

3.1 Overview of Preprocessing Pipeline

Social media and news text require substantial cleaning before analysis. Our pipeline applies to a series of transformations, each justified by the nature of news text and our prediction task.

Stage 1: Basic Text Cleaning

Remove formatting noise and non-semantic elements. News text contains URLs, HTML entities, special characters, and excessive whitespace that do not carry meaning.

- Remove URLs (http, https, www) - Remove HTML tags and entities - Strip special characters and extra whitespace - Remove user mentions, hashtags (for consistency across sources)

Justification: These elements add dimensionality without semantic value. Removing them reduces sparsity and prevents the model from learning spurious patterns.

Stage 2: Lowercasing

Convert all text to lowercase. “Bitcoin”, “BITCOIN”, and “bitcoin” should be treated as the same token.

Justification: Reduces vocabulary size and ensures consistency without loss of information for financial text.

Stage 3: Tokenization

Split text into individual words using a standard tokenizer (NLTK word_tokenize or spaCy).

Justification: Prerequisite for stopword removal, lemmatization, and feature extraction. Handles punctuation and contractions appropriately.

Stage 4: Punctuation and Symbol Removal

Filter out tokens that are purely punctuation or special characters.

Justification: Punctuation (periods, commas, etc.) is informative for sentence structure but not for sentiment or topic classification in financial text. We remove it to reduce noise, accepting that we lose some emotional signals (e.g., multiple exclamation marks).

Stage 5: Stopword Removal with Domain Awareness

Remove common English words (a, the, is, and, or, etc.) that appear in nearly all documents and provide little discriminatory power.

Exception: Preserve domain-specific and sentiment-bearing words that would normally be considered stopwords.

Crypto and finance stopword exceptions to keep:

- bitcoin, btc, crypto, cryptocurrency, ethereum, eth (domain keywords)

- bull, bear, bullish, bearish, hodl (sentiment indicators)

- buy, sell, long, short, bullish, bearish (trading actions)

- crash, surge, plunge, rally, pump, dump (price action verbs)

- good, bad, up, down, high, low (directional descriptors)

- not, no, cannot (sentiment modifiers)

Justification: In financial news, terms like “bullish” and “crash” are highly predictive of sentiment and price direction. Standard stopword lists would incorrectly filter these out.

Stage 6: Lemmatization

Reduce inflected words to their base form using a lemmatizer. “buying”, “bought”, “buys” all become “buy”. “crashed”, “crashing” all becomes “crash”.

Justification: Reduces feature space and sparsity while preserving meaning. Helps the model generalize by treating word variants as the same concept.

3.2 Examples: Raw vs. Processed Text

Example 1:

Raw text: "Richard MacManus: Bitcoin bottleneck continues to frustrate kiwis - Stuff"

Processed text : “richard macmanus: bitcoin bottleneck continues frustrate kiwis stuff”

Reduction: 2 tokens removed (20.0%)

Example 2:

Raw text: "AngelList Creator Naval Ravikant Backs S&P-Style Cryptocurrency Fund - CoinDesk"

Processed text: “angellist creator naval ravikant backs s&p-stylen cryptocurrency fund coindesk”
Reduction: 1 tokens removed (10.0%)

Example 3:

Raw text: "This country could soon make Bitcoin its official currency - The World Economic Forum"

Processed text: “country could soon make bitcoin official currency world economic forum”

Reduction: 4 tokens removed (28.6%)

Example 4:

Raw text: "Sweden – The Next Cryptocurrency King - LeapRate"

Processed text: “sweden next cryptocurrency king leaprate”

Reduction: 3 tokens removed (37.5%)

Example 5:

Raw text: "US government misses out on $600 million payday by selling dirty bitcoins too early - CNBC"

Processed text: “government misses $600 million payday selling dirty bitcoins early cnbc

Reduction: 6 tokens removed (37.5%)”

3.3 Corpus Statistics

Token-Level Statistics:

Total tokens in corpus (before preprocessing): Total tokens (after preprocessing): Token reduction percentage: %

Metric	Value
Unique tokens (vocabulary size)	13,688
Average tokens per headline	12.53
Median tokens per headline	12
Maximum tokens in single headline	37
Minimum tokens in single headline	2

Vocabulary Diversity:

Metric	Value
Type-Token Ratio (TTR)	0.0756
Hapax legomena (words appearing once)	6,038
Hapax percentage	44.11%

Most Frequent Tokens (Top 10):

Rank	Token	Frequency
1	bitcoin	7,384
2	coindesk	6,481
3	to	3,977
4	decrpyt	3,708
5	crypto	3,030
6	in	2,643
7	as	2,213
8	the	2,151
9	for	1,822
10	btc	1,714

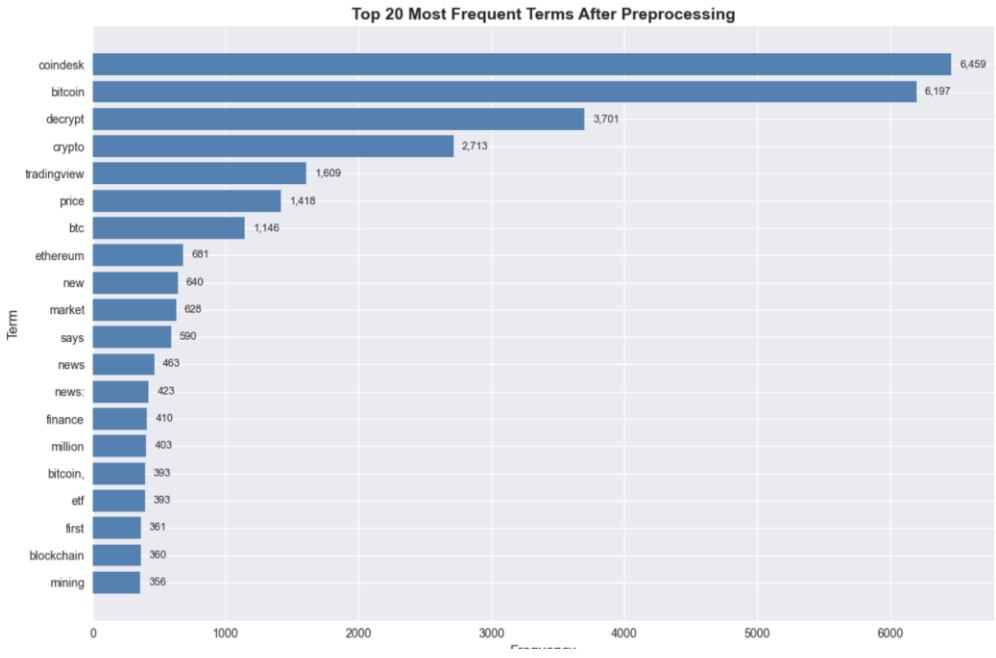
Part 4 - Data Understanding and Preparation

- Total Tokens before preprocessing: 187,919
- Total Tokens after preprocessing: 136,809
- Total reduction: 27.2%

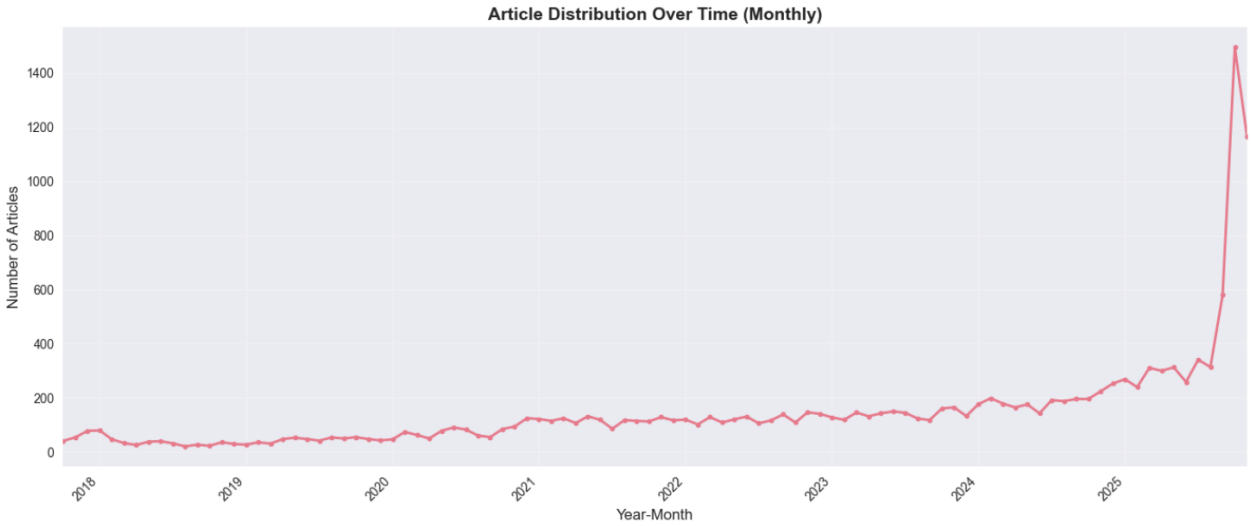
Word Frequency Findings:

Top 10 most frequent terms after preprocessing:

Rank	Terms	Frequency
1	coindesk	6,459
2	bitcoin	6,197
3	decrypt	3,701
4	crypto	2,713
5	tradingview	1,609
6	price	1,418
7	btc	1,146
8	ethereum	681
9	new	640
10	market	628



Distribution of articles by month



Distribution of word length after preprocessing

- Average word length: 6.44 characters
- Median word length: 6.00 characters
- Most common length: 7 characters