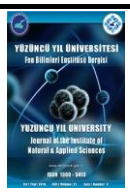
Yuzuncu Yil University
Journal of the Institute of Natural & Applied Sciences
https://dergipark.org.tr/en/pub/yyufbed



Research Article

Machine Learning Approach for Emotion Identification and Classification in Bitcoin Sentiment Analysis

Erol KINA[*1], Emre BİÇEK[2]

[1]Van Yüzüncü Yıl University, Özalp Vocational School, Computer Science Department, 65800, Van, Turkey
[2]Van Yüzüncü Yıl University, Engineering Faculty, Computer Engineering Department, 65100, Van, Turkey
Erol KINA, ORCID No: 0000-0002-7785-646X, Emre BİÇEK, ORCID No: 0000-0001-6061-9372
*Corresponding author e-mail: erolkina@yyu.edu.tr

Article Info

Received: 13.08.2024
Accepted: 23.09.2024
Online December 2024

DOI:10.53433/yyufbed.1532649

Keywords
Algorithms,
Bitcoin,
Machine learning,
NLP,
Sentiment analysis

Abstract: Bitcoin is the most valuable cryptocurrency and is renowned for its rapid and volatile price fluctuations in comparison to other currencies. This offers potential for the prediction of Bitcoin prices and has attracted the interest of researchers. Twitter (X) is one of the most widely used social media platforms. The aim of this study is to analyse the sentiment expressed in comments about bitcoin on the social media platform X using a variety of machine learning algorithms. A variety of machine learning techniques are used to classify user sentiment towards bitcoin. Moreover, the efficacy of standard bag-of-words and term frequency-inverse document frequency (TF-IDF) methods is evaluated in comparison with machine learning approaches for the purpose of expressing text as numerical vectors. Finally, a keyword ranking was performed to determine the importance of each sentiment in the development of cryptocurrencies. The bag-of-words and TF-IDF methods were used, which facilitate the representation of text-based data. The best result was obtained with the decision trees algorithm (98.74% accuracy) using the TF-IDF method. The bag-of-words method was found to produce better results in general.

Bitcoin Duygu Analizinde Duygu Tanıma ve Sınıflandırma için Makine Öğrenmesi Yaklaşımı

Makale Bilgileri

Geliş: 13.08.2024
Kabul: 23.09.2024
Online December 2024

DOI:10.53433/yyufbed.1532649

Anahtar Kelimeler
Algoritmalar,
Bitcoin,
DDİ,
Duygu analizi,
Makine öğrenmesi

Öz: Bitcoin en yüksek piyasa değerine sahip kripto para birimidir ve diğer para birimlerine kıyasla hızlı ve değişken fiyat dalgalanmalarıyla bilinir. Bu durum Bitcoin'in fiyat tahmini için fırsatlar sunmakta ve araştırmacıların ilgisini çekmektedir. Twitter (X), en yaygın kullanılan sosyal medya platformlarından biridir. Bu çalışma kapsamında, makine öğrenimi algoritmalarını kullanarak Bitcoin ile ilgili X yorumlarının duyarlılığı analiz edilmiştir. Bitcoin'e yönelik kullanıcı duyarlılığını sınıflandırmak için spesifik makine öğrenimi teknikleri kullanılmış ve metni sayısal vektörler olarak ifade etmek için standart kelime torbası ve terim frekansı-ters belge frekansı (TF-IDF) yöntemleri makine öğrenimi yaklaşımlarıyla karşılaştırılmıştır. Son olarak, kripto para birimlerinin gelişiminde her duygunun önemini belirlemek için anahtar kelime sıralaması yapılarak, metin tabanlı verilerin temsilini kolaylaştıran Bag-of-words ve TF-IDF yöntemleri kullanılmıştır. En iyi sonuç TF-IDF yöntemi kullanılarak karar ağaçları algoritmasıyla (%98.74 doğruluk) elde edilmiş, çalışmada Bag-of-words yönteminin genel olarak daha iyi sonuçlar ürettiği görülmüştür.

1. Introduction

Bitcoin, the cryptocurrency with the highest market value, is actively traded on over 40 exchanges across more than 30 countries and facilitates transactions in a remarkably diverse range of fiat currencies and digital assets (Vumazonke & Parsons, 2023). Due to its relatively young age as a currency, Bitcoin exhibits significantly higher price volatility compared to conventional fiat currencies. This heightened volatility presents an intriguing opportunity for exploring the potential of price forecasting in the context of Bitcoin (Dutta et al., 2020; Gozbasi, 2021). Bitcoin's transparency is unparalleled compared to fiat currencies, as the latter only provides limited information on monetary activities and cash transactions. In today's social media ecosystem, the abundance of data enables us to extract users' perspectives on events, goods, supply, and desires. X, being one of the most widely accessed online social media platforms, serves as a communication channel for over 100 million active users each month, facilitating the exchange of ideas among individuals. Researchers have recently leveraged the power of machine learning to analyze sentiment in Bitcoin-related tweets as a means of predicting price volatility (Bulu et al., 2019; Sallis et al., 2021). Sentiment refers to the attitude or emotion conveyed through a statement or position. Sentiment analysis, as a branch of text analytics, delves into the underlying motivations behind an individual's feelings about a product or service. Textual emotions analytics leverages natural language processing (NLP) and machine learning (ML) techniques to assign numerical ratings to various entities, topics, themes, and categories present in a given phrase or sentence. Over the years, researchers have developed different approaches to sentiment analysis in Bitcoin sentiment analysis. X, with its extensive user base, generates a wealth of data that holds valuable insights to enhance market dynamics. Consequently, employing sentiment analysis is vital to comprehend users' requests, whether they are positive or negative. As the pioneering force behind the blockchain-based monetary revolution, Bitcoin exerts a disproportionate influence on the overall valuation of the cryptocurrency market. Hence, there is considerable interest within the machine learning and data mining communities in forecasting fluctuations in the price of bitcoin and providing direct explanations for these changes (Avcı & Koca, 2023). Such information can aid in the better assessment of risks associated with the cryptocurrency industry. Machine learning approaches for emotion identification and classification in Bitcoin sentiment analysis have gained traction as a promising method to predict price volatility (Pradana & Hayaty, 2019). Many studies have been conducted on machine learning algorithms and social media sentiment analysis. People’s methods of expressing themselves have shifted since the advent of the Internet. Blogs, internet forums, review sites, etc., have become the new standard for this. This user-created content is vital to people’s daily lives. People who are considering making a purchase will typically read customer feedback about the product on the internet. Users can’t possibly sift through all of the user-generated content. The use of sentiment analysis and similar methods allows this process to be automated. Due to their low correlations with other major financial assets and the advantages brought by blockchain technologies, cryptocurrencies have developed as a new class of financial assets in the asset management business in recent years, typically comprising a portion of riskier investment portfolios (Li et al., 2019).

Despite only 92% of the total supply of Bitcoins currently in circulation, the market capitalization of this cryptocurrency has surpassed $588 Million according to www.coinmarketcap.com data, July - 2023. This unprecedented growth and interest in Bitcoin has led researchers to explore the connection between sentiment analysis and Bitcoin prices. Sentiment analysis primarily utilizes two methodologies: symbolic approaches, also referred to as the Knowledge base approach and machine learning techniques (Georgoula et al., 2015; Alasmari & Dahab, 2017). To accurately identify sentiments, a knowledge-based system requires a comprehensive repository of predefined emotions and a robust knowledge representation (Fakieh et al., 2023). Based on a comprehensive review of the existing literature, it is evident that there has been a scarcity of research dedicated to the analysis of sentiment toward Bitcoin utilizing X as a data source. Furthermore, to the best of our knowledge, no prior investigations have been conducted to establish a systematic ranking of keywords based on their association with specific sentiments. In light of this research gap, the present study attempts to fill this void by introducing a novel machine learning model for predicting user sentiment towards Bitcoin by harnessing X data. The results obtained from our model show marginal improvements over previous research efforts in this area. This scholarly contribution not only advances the field's understanding of sentiment analysis concerning Bitcoin using X data but also introduces an innovative approach to

keyword ranking about sentiment classification. Furthermore, we conducted keyword ranking to identify the most influential terms associated with each sentiment, which may impact the development of cryptocurrencies. This approach involves utilizing natural language processing techniques and machine learning algorithms to analyze sentiment in tweets related to Bitcoin (Tanwar et al., 2021).

To classify emotions, a machine learning approach employs a training set to build a sentiment classifier. Machine learning techniques offer advantages over the Knowledge base method as they do not necessitate a pre-established database encompassing all potential emotions. In our research, we employed multiple distinct machine learning methods to classify tweets. It is customary in sentiment analysis to consider various granularities, ranging from broad to specific. Sentiment analysis can be divided into two distinct levels of granularity: the coarse level, which focuses on identifying the overall tone of a document, and the fine level, which examines individual attributes. Hence, it is common practice to conduct sentiment analysis at different granularities, encompassing both broader and more specific aspects (Dass et al., 2020).

In classification tasks, machine learning approaches necessitate the utilization of both a training set and a test set. The training set comprises feature vectors that are assigned categorical labels and serve as input for the model construction. The objective of constructing a classification model is to accurately assign classes to input feature vectors based on the information gleaned from the training set. Predictions are then generated for the class labels of feature vectors that have not been previously observed, thus facilitating the evaluation and validation of the model's performance using a separate test set (Sami et al., 2021). In cases involving features with high interdependencies, previous research by Domingos & Pazzani (2017) revealed that naive Bayes showed strong performance while contradicting the basic assumption of naive Bayes, which assumes feature independence. The application of machine learning techniques in the context of the Bitcoin market represents a novel approach that has not been previously investigated. In a span of fifty days involving the buying and selling of Bitcoins (Shah & Zhang, 2014) employed Bayesian regression and achieved an 89% profit. Neethu & Rajasree (2013) introduced a novel feature vector to effectively classify tweets as either positive or negative and to extract individuals' opinions regarding products. Time series analysis provides a valuable methodology for examining the relationship between Bitcoin prices and a range of factors, including economic conditions, technological developments and even the sentiment expressed on X (Cambria et al., 2014). Despite the existence of numerous online platforms dedicated to sentiment analysis, the availability of a publicly accessible interactive online platform for dynamic adaptive sentiment analysis remains limited.

In his study Ibrahim (2021), the author employs the analysis of a corpus of tweets, which he manipulates and interprets, in order to predict the early movements of the cryptocurrency market. The primary objective is to showcase the efficacy of deep learning architectures in addressing challenges associated with sentiment analysis. Machine learning users frequently express challenges in determining the optimal size of the training data set, which can induce significant stress (Quinlan, 1992). To resolve such conflicts, Wang et al. (2020) propose a systematic approach through the analysis of real-world scenarios encompassing forecasting the popularity and sentiment of tweets, Facebook posts, Mashable blogs, Google News, Yahoo News, the US housing survey, and Bitcoin pricing.

The objective of this study is to conduct a sentiment analysis of comments related to Bitcoin on the social media platform X, employing a range of machine learning algorithms. Various machine learning techniques are applied to classify user sentiment regarding Bitcoin. In addition, the effectiveness of conventional bag-of-words and term frequency-inverse document frequency (TF-IDF) methods is assessed in comparison to machine learning approaches for converting textual data into numerical vectors. Furthermore, a keyword ranking analysis was performed to evaluate the significance of different sentiments in the development of cryptocurrencies. Both the bag-of-words and TF-IDF methods, which enable the representation of textual data, were utilized.

2. Material and Methods

The workflow diagram for this study is presented below (Figure 1). The word clouds, created with the 50 most frequently used words, are presented in Figure 2.

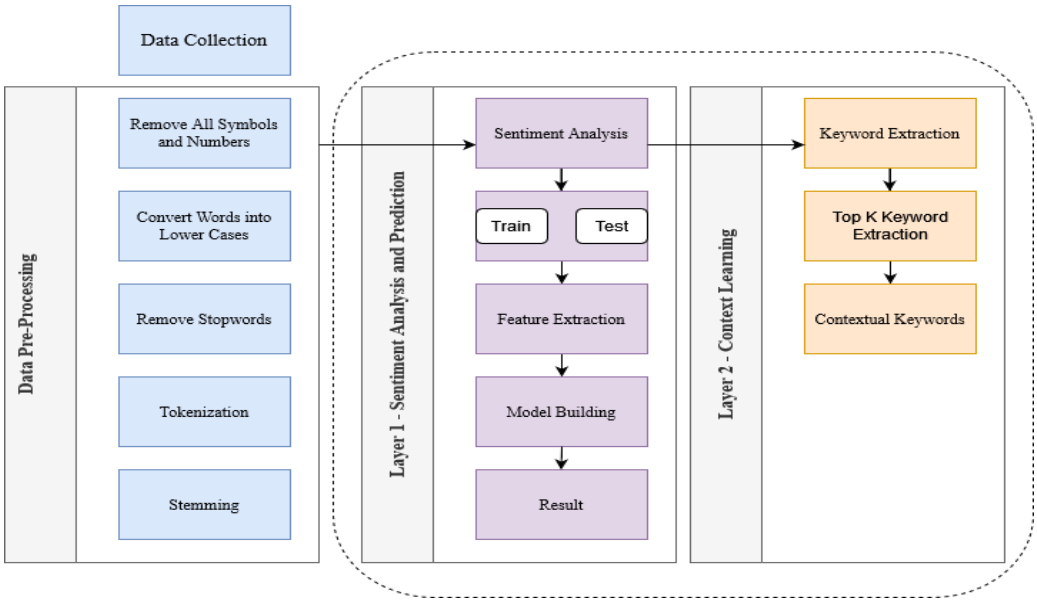


Figure 1. The workflow is as follows: collection of data, data preparation, pre-processing and feature extraction.

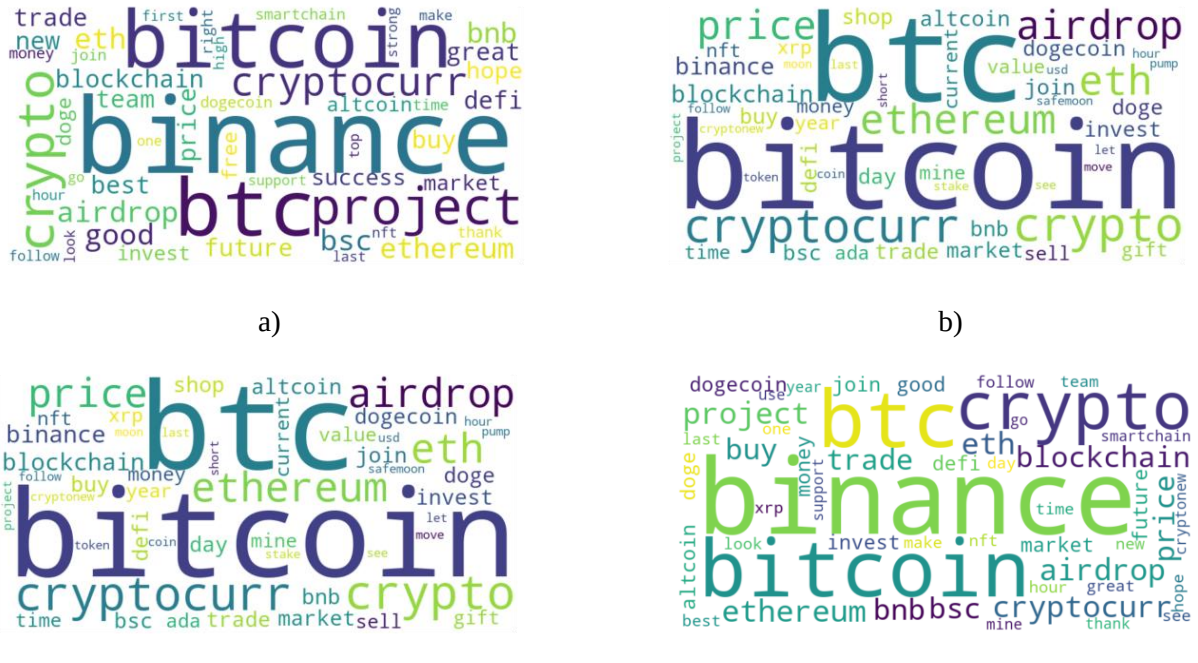


Figure 2. Opinions expressed by the top 50 keywords in a word cloud. a) Positive Sentiments b) Negative Sentiments c) Neutral Sentiments d) Overall Sentiments.

2.1. Acquisition and preparation of data

The primary source of tweets relevant to Bitcoin sentiment research is X, from which Bitcoin-related tweets were collected for multiple months in 2022 using the Kaggle platform (Anonymous, 2023a). The dataset comprises one million Bitcoin-related tweets, which underwent preprocessing steps, including the exclusion of non-English tweets, removal of emoji, non-printable characters, and common punctuation symbols (#, !, &, %, $). In most cases, the more complex X handle "@person" was replaced with the term "username". Tokenization, a subsequent step, was performed to segment a text block into

individual words using a predetermined delimiter (Hasan et al., 2021). The final preprocessing step involved stemming, where all possible affixes were removed from index terms to assign a phrase to the appropriate index entry (Larkey et al., 2007; Alnaied et al., 2020). The data preparation process involved a combination of regular expressions (Elbagir & Yang, 2019) and the Natural Language Toolkit (NLTK) (Yogish et al., 2019).

2.2. Layer 1: Prediction and sentiment analysis

Utilizing sentiment analysis (Feldman, 2013), we employed a methodology that calculates the degree of engagement with each topic by initially determining the average number of retweets, likes, and followers associated with that topic, subsequently ranking the topics based on their level of engagement. To conduct sentiment analysis on the text corpus, we utilized the textblob package in Python (Loria, 2018). The sentiment index ranged from -1.0 to 1.0, with a value of -1.0 representing the most negative content and a value of 1.0 indicating the most positive text. It is worth noting that NLP algorithms yield numerical outputs. However, due to the current limitations, direct input of our text into the algorithm was not feasible. Consequently, we employed the bag-of-words and TF-IDF approaches to extract features from the text by converting them into numeric vectors. Bag of words: When attempting to explain text corpora, information extraction researchers primarily utilize the bag of words presentation. An error-free transformation from unstructured to structured data can be achieved with this straightforward approach (Zhang et al., 2010).

TF-IDF: The significance of a term in a document collection can be measured using the term frequency-inverse document frequency (TF-IDF) metric, which combines the term frequency (TF) that quantifies the frequency of a term in a document and the inverse document frequency (IDF) that quantifies the importance of the term (Hakim et al., 2014). In our study, we aim to train our proposed model using various machine learning techniques, including Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, XgBoost, and Support Vector Machine algorithms. Logistic Regression: Logistic Regression is a preferred regression analysis method when the dependent variable being studied can be categorized into two distinct categories, making it suitable for binary classification tasks (LaValley, 2008). The logistic function, also known as the sigmoid function, plays a critical role in the logistic regression methodology. This function transforms a real-valued input into a bounded range between 0 and 1, following an S-shaped curve, although it never exactly reaches the limits (Kleinbaum & Klein, 2002).

Support Vector Machine: The Support Vector Machine (SVM) is a classification algorithm that constructs a discrete hyperplane in the feature space of the training data and its samples to separate them (Suthaharan, 2016). SVM formulates an optimization problem to identify the optimal hyperplane that maximally separates the two sets of data in the training dataset (Joachims, 1999).

Naïve Bayes is a classification method that applies Bayes' Principle and assumes the independence of predictors to classify data (Sammut & Webb, 2011). The posterior probability is calculated using the likelihood, which represents the probability of the predictors, the prior probability of the class, and Bayes' theorem (Murphy, 2006).

Decision Tree is a problem-solving technique that utilizes a tree structure, where the leaves represent classes and the internal nodes denote attributes (Quinlan, 1992). The concept of information gain is employed to evaluate the impact of input attributes on the entropy change resulting from dataset partitioning (Nowozin et al., 2011).

Random Forest is an ensemble learning method comprising multiple classification or regression trees, each trained on analogous datasets obtained through bootstrapping. The predictions produced by these trees are then combined to yield a more accurate result (Rigatti, 2017). As processing power becomes more distributed, there is an increasing trend towards utilising advanced mathematical techniques in various fields. This is particularly evident in the era of "big data" and machine learning, where survival analysis has gained prominence as a methodology (Breiman, 2001). XGBoost, an advanced implementation of gradient boosting machines, has demonstrated its ability to surpass the computational limitations of enhanced tree methods, establishing itself as a versatile and state-of-the-art approach in various domains (Chen & Guestrin, 2016; Mitchell & Frank, 2017).

In the second phase of the analysis, a keyword ranking algorithm is employed to assess the relative significance of each keyword present in the text. Subsequently, the keywords are ordered based on their relative importance. From this ranked list, the top k keywords, where k is set to 50 in this study, are selected to provide an evaluation of the overall contextual representation.

In the initial experiment, the distribution of samples across different sentiment categories is determined by calculating the fraction assigned to each category. Subsequently, a subsequent experiment is conducted to evaluate the performance of different machine learning strategies on an independent test dataset, thereby demonstrating their effectiveness. Subsequently, the machine learning models are ranked based on their predictive accuracy, and the most promising model is chosen for sentiment analysis based on this ranking.

2.3. Preparing an experiment and deciding on parameters

The cleaned-up text underwent bag-of-words and TF-IDF transformations using the CountVectorizer and TfidfVectorizer modules from the scikit-learn library (version: 0.22.2) (Anonymous, 2023b) in Python (version: 3.7) (McGrath, 2023). Leveraging the Scikit-Learn module in Python facilitated the efficient development of a robust machine learning model. Both TF-IDF and bag-of-words methodologies were employed as feature extraction techniques, with a filter applied to exclude words that occurred fewer than 500 times in the texts.

2.4. Sentiment analysis

I have compiled a comprehensive Bitcoin X dataset extracted from the Kaggle database, consisting of 1,000,000 opinions specifically related to Bitcoin. Upon conducting an analysis of the sentiment expressed in these tweets, the findings reveal that out of the total dataset, 380.044 tweets are categorized as positive, 509.108 tweets exhibit a neutral sentiment, and 110.848 tweets convey a negative sentiment, as visually presented in Figure 3.

2.5. Evaluation matrix

In order to assess the accuracy of our predictions, we conducted an evaluation based on several performance metrics, including accuracy, precision, recall, and F1-score. Accuracy (A) represents the ratio of correctly predicted instances to the total number of training samples, providing an overall measure of prediction correctness. Precision measures the proportion of true positive predictions over the total predicted positive instances, reflecting the classifier's ability to minimize false positive results. Recall (R), on the other hand, quantifies the proportion of true positive instances out of all actual positive samples, indicating the classifier's ability to identify positive instances correctly (Raaijmakers & Shiffrin, 1992). F1-score (F1), often referred to as the harmonic mean, offers a balanced measure between precision and recall, effectively capturing the trade-off between them. It is particularly useful in cases where the dataset is imbalanced and can handle false positive and false negative instances simultaneously (Chicco & Jurman, 2020; Narkhede, 2018).

3. Results and Discussion

Utilizing the available data, I constructed a model capable of distinguishing between positive, neutral, and negative sentiment labels. The initial step in dataset preparation involved eliminating potential sources of confusion. Following data cleaning and organization, the dataset was divided into training (80%) and testing (20%) sets. Feature extraction was performed using the bag of words and TF-IDF techniques, converting textual content into numerical vectors. Classification algorithms based on machine learning were then applied. The outcomes of various classification techniques are presented in Table 1 and Figure 5. Specifically, the Decision Tree algorithm achieved an accuracy of 98.74% with TF-IDF and 97.83% with bag-of-words, while SVM achieved accuracies of 93.09% and 97.38%, respectively. Naïve Bayes attained accuracies of 86.89% (TF-IDF) and 85.18% (bag-of-words), and Random Forest achieved accuracies of 97,55% (TF-IDF) and 97.02% (bag-of-words). Xgboost yielded an accuracy of 84.72% with bag-of-words and 84.24% with TF-IDF. It is notable that Xgboost exhibits

lower accuracy than Naive Bayes, particularly in the context of TF-IDF. With regard to TF-IDF and Bag-of-words, Decision Tree demonstrated the highest accuracy. Additionally, other performance indicators such as recall and F1-score were evaluated, as they provide insights into the model's ability to identify each sentiment accurately.

In terms of precision, recall, and f1-score, the decision tree model utilizing bag-of-words exhibits superior performance compared to other models, achieving precision, recall, and f1-score values of 97.71%, 97.57%, and 97.68%, respectively. The random forest model employing bag-of-words demonstrates a precision score of 97.45%, a recall score of 93.19%, and an f1-score of 94.76%. In the context of TF-IDF classification, Random Forest showcases notable performance with a precision score of 97.74%, a recall score of 93.84%, and f1-score of 95.66% (Figure 6). Figure 4 illustrates the log loss curve and corresponding loss values for each model, with Naïve Bayes exhibiting the highest loss value of 0.55 among the tested models. In the bag-of-words scenario, SVM attains the lowest loss value of 0.13. SVM emerges as the most effective model in discerning between positive, negative, and neutral classifications, boasting the lowest loss value among the evaluated models. Regarding accuracy, precision, recall, and f1-score, the decision tree model surpasses all other models, while bag-of-words demonstrates superior performance compared to TF-IDF in the majority of algorithms.

The word cloud representation in Figure 2 illustrates the prevalence of certain words in the sampled data. In order to identify the most commonly used terms associated with each sentiment, we applied context learning to the dataset. Following the sentiment identification process, we extracted the top 50 frequently occurring words through a keyword analysis. The resulting lists consist of the top 50 terms for positive, negative, and neutral sentiments, respectively. The top 50 most frequently occurring terms for each emotion are presented in Table 2.

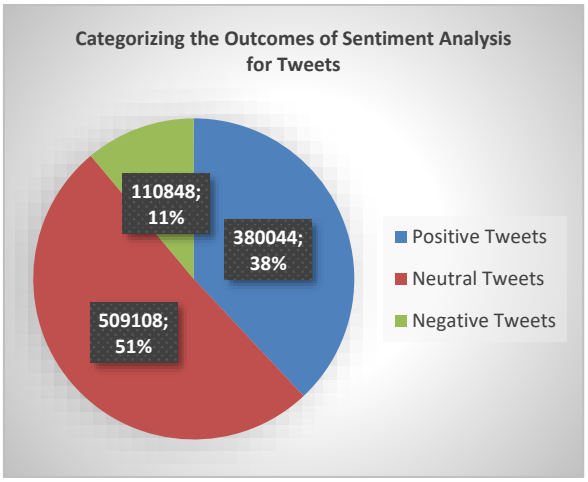


Figure 3. Sentiment analysis result.

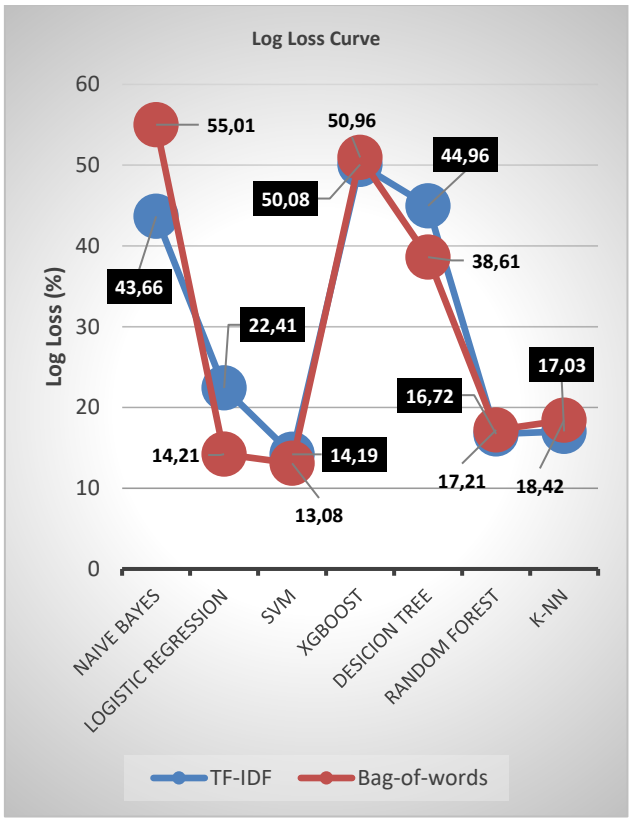


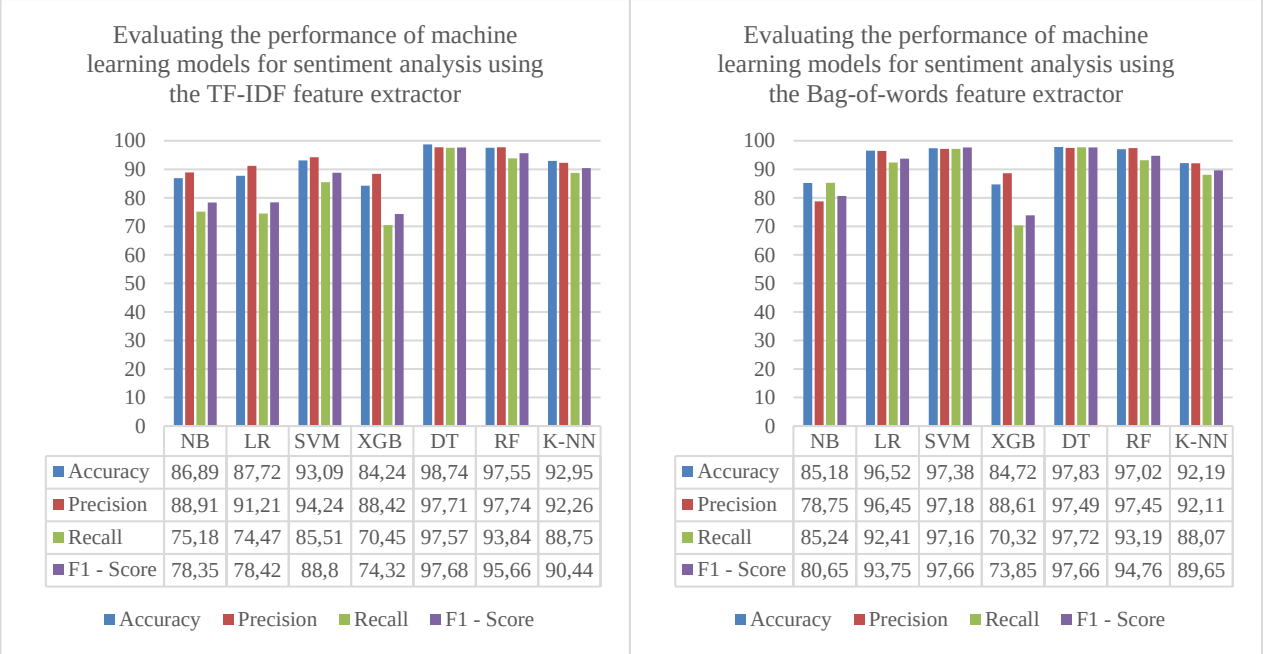
Figure 4. Analyzing the Bag-of-words and TF-IDF log loss curves side by side.

Table 1. Evaluating the performance of several machine learning models for sentiment analysis with Bag-of-words and TF-IDF as feature extractors

Feature Extraction	Algorithms	Accuracy	Precision	Recall	F1-Score
TF-IDF	Naïve Bayes	86.89	88.91	75.18	78.35
	Logistic Regression	87.72	91.21	74.47	78.42
	SVM	93.09	94.24	85.51	88.8
	XGBoost	84.24	88.42	70.45	74.32
	Decision Tree	**98.74**	97.71	**97.57**	**97.68**
	Random Forest	97.55	**97.74**	93.84	95.66
	K-NN	92.95	92.26	88.75	90.44
Bag-of-words	Naïve Bayes	85.18	78.75	85.24	80.65
	Logistic Regression	96.52	96.45	92.41	93.75
	SVM	97.38	97.18	97.16	**97.66**
	XGBoost	84.72	88.61	70.32	73.85
	Decision Tree	**97.83**	**97.49**	**97.72**	**97.66**
	Random Forest	97.02	97.45	93.19	94.76
	K-NN	92.19	92.11	88.07	89.65

Table 2. The top fifty terms encompass a diverse range of sentiments

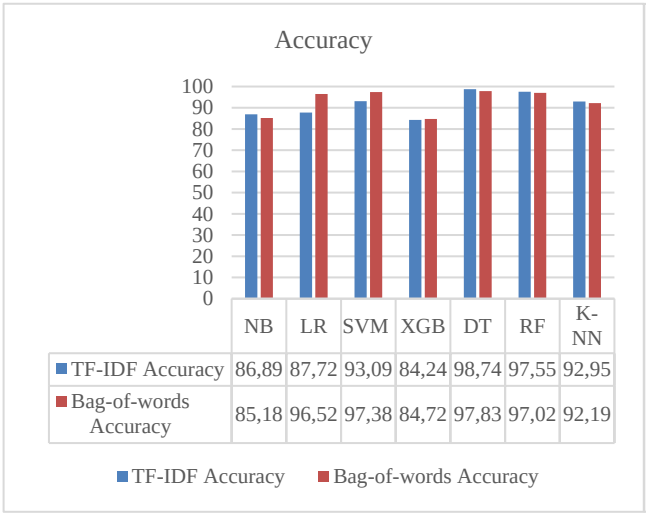
Sentiment Type	Description	Top-50 Keywords
Positive	Discussions pertaining to mutual aid, social consciousness, and moral reinforcement are commonly observed among individuals. The analysis of Bitcoin's overall appeal is a topic of discussion, whereby people have received an overwhelming influx of expressions conveying gratitude and admiration, including from investors, developers, and other relevant stakeholders.	'bitcoin', 'btc', 'project', 'crypto', 'cryptocurr', 'airdrop', 'good', 'bsc', 'eth', 'ethereum', 'bnb', 'great', 'blockchain', 'binance', 'price', 'new', 'future', 'best', 'team', 'get', 'defi', 'trade', 'buy', 'hope', 'success', 'market', 'invest', 'free', 'doge', 'altcoin', 'like', 'dogecoin', 'time', 'binance', 'smartchain', 'join', 'money', 'thank', 'right', 'support', 'top', 'last', 'strong', 'one', 'look', 'nft', 'make', 'hour', 'follow', 'high', 'go', 'first',
Negative	The majority of individuals exhibit a negative sentiment, expressing concerns regarding family issues and displaying reluctance towards investment. Their primary attention is directed towards the most economically disadvantaged workers. Coins are being traded at prices lower than their face value. The involvement of influential investors, known as whales, intensifies when individuals experience losses in their coin holdings.	'bitcoin', 'btc', 'crypto', 'long', 'cryptocurr', 'buy', 'blockchain', 'ethereum', 'market', 'get', 'eth', 'unknown', 'wallet', 'binance', 'price', 'game', 'like', 'money', 'base','time', 'defi', 'people', 'make', 'trade', 'invest', 'use', 'green', 'mean', 'see', 'move', 'usd', 'one', 'current', 'day', 'play', 'signal', 'bnb', 'look', 'dogecoin', 'dont', 'altcoin', 'hard', ' fuck', 'still', 'doge', 'go', 'mine', 'think', 'year', 'follow', 'nft'
Neutral	Following the market conditions observed this year, maintaining a neutral stance can potentially contribute to the restoration of confidence, as it signifies the absence of expectations for further decline in the asset.	'bitcoin', 'btc', 'crypto', 'cryptocurr', 'ethereum', 'eth', 'airdrop', 'price', 'blockchain', 'binance', 'dogecoin', 'bnb', 'buy', 'trade', 'bsc', 'doge' , 'market', 'day', 'altcoin', 'get', 'join', 'invest', 'defi', 'shop', 'money', 'gift', 'like', 'xrp', 'current', 'year', 'nft', 'valu', 'mine', 'ada', 'time', 'sell', 'cryptonew', 'stake', 'follow', 'project', 'moon', 'last', 'see', 'coin', 'let', 'hour', 'safemoon', 'usd', 'short', 'pump', 'move', 'token',
Overall	The collective sentiment of the general populace.	'bitcoin', 'btc', 'crypto', 'cryptocurr', 'project', 'airdrop', 'ethereum', ' eth', 'blockchain', 'price', 'binance', 'bsc', 'bnb', 'buy', 'trade', 'get', 'dogecoin', 'good', 'market', 'defi', 'invest', 'doge', 'altcoin', 'join', 'future', 'like', 'money', 'great', 'day', 'time', 'new', 'nft', 'team', 'hope', 'last','xrp', 'follow', 'one', 'see', 'look', 'best', 'hour', 'make', 'binance', 'smartchain', 'mine', 'go', 'support', 'use', 'year', 'thank', 'cryptonew',



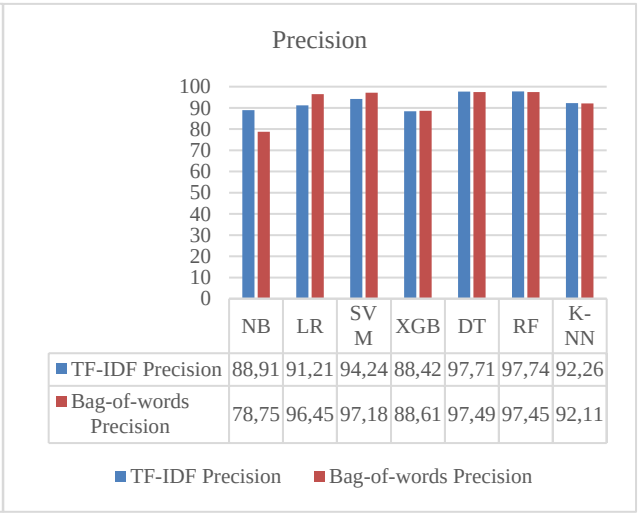
a) b)

Figure 5. a) TF-IDF feature extractor b) Bag-of-words feature extractor.

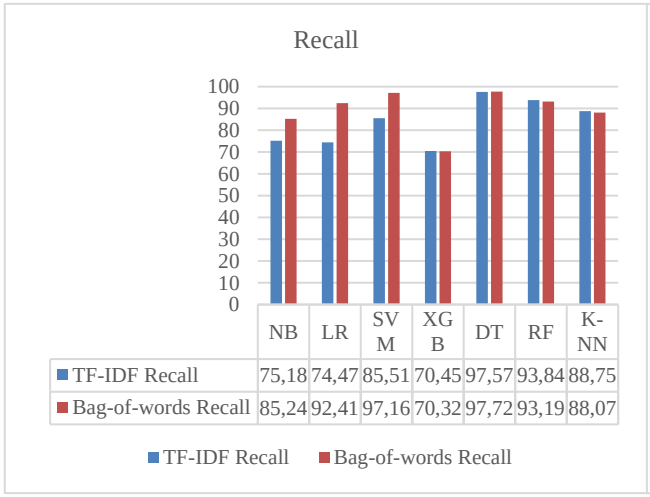
The absence of such a platform is attributed to the requirement of adaptability to emerging data streams, which necessitates a comprehensive solution capable of accommodating evolving data sources. ClowdFlows, as an open-source cloud-based scientific workflow platform, offers additional functionalities through its add-ons, enabling efficient analysis of data streams and facilitating machine learning processes (Kranjc et al., 2015). In aspect-based sentiment analysis, a cascaded framework based on particle swarm optimization (PSO) is employed to perform feature selection and classifier ensemble, aiming to enhance the accuracy of sentiment analysis by focusing on specific aspects or attributes (Akhtar et al., 2017). This approach leverages the optimization capabilities of PSO to iteratively refine the selection of features and create an ensemble of classifiers, thereby improving the overall performance of sentiment analysis in capturing nuanced sentiments toward different aspects (Barros et al., 2020). The distinctive approach to currency transfer and the potential hedging capabilities have propelled cryptocurrencies to the forefront of public interest in recent times. In order to ascertain the relationship between textual information derived from sources such as news articles and tweets and the direction of cryptocurrency prices, data mining techniques are employed. The objective is to utilize these techniques to infer patterns and correlations within the text documents, enabling a better understanding of the factors influencing cryptocurrency price movements (Patel et al., 2022). This study aims to investigate the relationship between fluctuations in Bitcoin price and the sentiment expressed by its users through the utilization of machine learning techniques. By analyzing user-generated data, the objective is to identify patterns and correlations between sentiment and price changes in order to gain insights into the dynamics of Bitcoin market (Suthaharan, 2016; Kinderis et al., 2018). Despite the utilization of various machine learning strategies and tools in sentiment analysis during elections, there is a critical need for an advanced method that remains at the forefront of the field. To address these challenges, the development of hybrid approaches, such as incorporating machine learning into sentiment analyzers, is underway to tackle these issues directly (Greaves et al., 2013; Rahman et al., 2018). In this study, a representative 10% random sample of X users will be employed to illustrate the application of deep learning for assessing market-level sentiment. The primary objective of this approach is to showcase the utilization of deep learning techniques in capturing and analyzing the expressed sentiments of users, thereby facilitating insights into the dynamics of sentiment at the market level (Andhale et al., 2021).



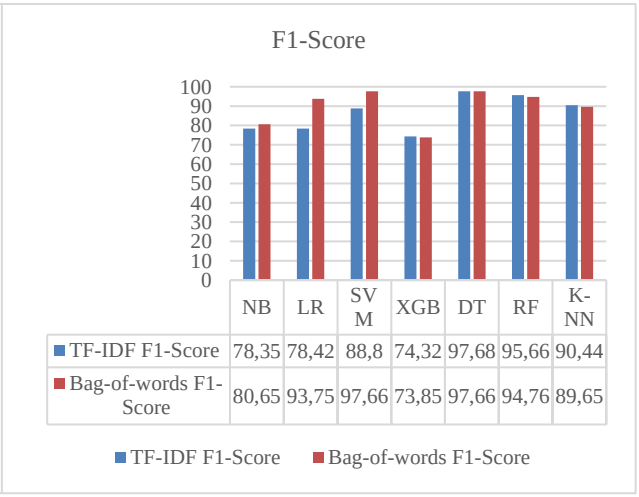
a)



b)



a)



b)

Figure 6. a) Accuracy b) Precision c) Recall d) F1-Score.

4. Conclusion

This research paper presents a novel approach for forecasting the future trajectory of Bitcoin prices through the integration of sentiment analysis and machine learning models. The exponential growth of social media has created a vast amount of data that necessitates effective evaluation and extraction of insights. Various techniques based on Symbolic and Machine Learning methodologies have been developed to infer an author's emotional state from text. Machine Learning strategies, in particular, offer enhanced productivity and reduced complexity compared to Symbolic methods. One promising application of these techniques is sentiment analysis on X data. However, challenges arise when attempting to identify emotionally impactful keywords within tweets containing multiple keywords. Despite these challenges, microblogging platforms and other social media channels serve as valuable sources of information on public opinion. However, it is important to acknowledge the existence of opportunities for the spread of disinformation. In this study, several machine learning approaches were used to analyse sentiment, and the proposed model was implemented after establishing the context of the dataset. In the future, this research can contribute to a comprehensive study of the evolving emotions and attitudes of individuals across different currencies.

References

Akhtar, Md. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect-based sentiment analysis. *Knowledge-Based Systems, 125*, 116–135. https://doi.org/10.1016/j.knosys.2017.03.020

Alasmari, S. F., & Dahab, M. (2017). Sentiment detection, recognition and aspect identification. *International Journal of Computer Applications, 177*(2), 31-38. https://doi.org/10.5120/ijca2017915675

Alnaied, A., Elbendak, M., & Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for Arabic information retrieval. *Egyptian Informatics Journal, 21*(4), 209–217. https://doi.org/10.1016/j.eij.2020.02.004

Andhale, S., Mane, P., Vaingankar, D. C., Karia, K., & Talele, K. (2021). Twitter sentiment analysis for COVID-19. In *2021 International Conference on Communication Information and Computing Technology (ICCICT)* (pp. 1-12), Mumbai, India. https://doi.org/10.1109/iccict50803.2021.9509933

Anonymous. (2023a). Bitcoin sentiment analysis | Twitter data. *Kaggle*. Access date: July 23, 2023. https://www.kaggle.com/datasets/gautamchettiar/bitcoin-sentiment-analysis-twitter-data

Anonymous. (2023b). Scikit-learn/scikit-learn: Scikit-learn 0.22.1. Access date: July 23, 2023. https://doi.org/10.5281/zenodo.3596890

Avcı, İ., & Koca, M. (2023). Predicting DDoS attacks using machine learning algorithms in building management systems. *Electronics, 12*(19), 4142. doi: https://doi.org/10.3390/ELECTRONICS12194142

Barros, D. P., Moura, J., Freire, C. R., Taleb, A. C., De Medeiros Valentim, R. A., & De Morais, P. S. G. (2020). Machine learning applied to retinal image processing for glaucoma detection: Review and perspective. *Biomedical Engineering Online, 19*(1), 20. https://doi.org/10.1186/s12938-020-00767-2

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Bulu, B., Yağar, F., Kopmaz, B., Şişman Kitapçı, N., Kitapçı, O., Aksu Kılıç, P., Köksal, L., & Mumcu, G. (2019). The content of Twitter messages of different health groups: The role of social media in health. *International Journal of Health Management and Tourism, 4*(3), 228–236. https://doi.org/10.31201/ijhmt.644197

Cambria, E., Olsher, D., & Rajagopal, D. (2014). SENTICNeT 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence, 28*(1). https://doi.org/10.1609/aaai.v28i1.8928

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1). https://doi.org/10.1186/s12864-019-6413-7

Dass, S., Kannan, V. K., & Shyamala, K. (2020). Sentiment severity on location-based social network (LBSN) data of natural disasters. *International Journal of Recent Technology and Engineering, 8*(5), 6–12. https://doi.org/10.35940/ijrte.e6631.018520

Domingos, P., & Pazzani, M. (2017). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103-130. https://rdcu.be/dgWb2

Dutta, A., Kumar, S., & Basu, M. (2020). A Gated Recurrent Unit approach to Bitcoin price prediction. *Journal of Risk and Financial Management, 13*(2), 23. https://doi.org/10.3390/jrfm13020023

Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the International Multiconference of Engineers and Computer Scientists* (Vol. 122, p. 16)

Fakieh, B., Al-Ghamdi, A. S. A.-M., Saleem, F., & Ragab, M. (2023). Optimal machine learning driven sentiment analysis on COVID-19 Twitter data. *Computers, Materials & Continua, 75*(1), 81–97. https://doi.org/10.32604/cmc.2023.033406

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89. https://doi.org/10.1145/2436256.2436274

Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D. N., & Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of Bitcoin prices. *Social Science Research Network*. https://doi.org/10.2139/ssrn.2607167

Gozbasi, O. (2021, July 12). Is Bitcoin a safe haven? A study on the factors that affect Bitcoin prices. *International Journal of Economics and Financial Issues, 11*(4), 35-40. https://econjournals.com/index.php/ijefi/article/view/11602

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research, 15*(11), e239. https://doi.org/10.2196/jmir.2721

Hâkim, A., Erwin, A., Eng, K., Galinium, M., & Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In *6th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1–4).

Hasan, K. M. A., Shovon, S. D., Joy, N. H., & Islam, S. (2021). Automatic labeling of Twitter data for developing COVID-19 sentiment dataset. In *2021 5th International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). https://doi.org/10.1109/eict54103.2021.9733548

Ibrahim, A. (2021). Forecasting the early market movement in Bitcoin using Twitter's sentiment analysis: An ensemble-based prediction model. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-5). https://doi.org/10.1109/iemtronics52119.2021.9422647

Joachims, T. (1999). Svmlight: Support vector machine. *SVM-Light Support Vector Machine, University of Dortmund, 19*(4), 25. http://svmlight.joachims.org/

Kinderis, M., Bezbradica, M., & Crane, M. (2018). Bitcoin currency fluctuation. In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (pp. 31-41). https://doi.org/10.5220/0006794000310041

Kleinbaum, D. G., & Klein, M. (2002). Analysis of matched data using logistic regression. In *Logistic Regression* (pp. 227–265). Springer eBooks. https://doi.org/10.1007/0-387-21647-2_8

Kranjc, J., Smailović, J., Podpečan, V., Grčar, M., Žnidaršič, M., & Lavrač, N. (2015). Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. *Information Processing and Management, 51*(2), 187–203. https://doi.org/10.1016/j.ipm.2014.04.001

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007). Light stemming for Arabic information retrieval. In: Soudi, A., Bosch, A.v., Neumann, G. (eds) *Arabic computational morphology. Text, speech and language technology*, vol 38. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6046-5_12

LaValley, M. P. (2008). Logistic regression. *Circulation, 117*(18), 2395–2399. https://doi.org/10.1161/circulationaha.106.682658

Li, T., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R., & Fu, F. (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *Frontiers in Physics, 7*. https://doi.org/10.3389/fphy.2019.00098

Loria, S. (2018). *Textblob Documentation* (Release 0.15, 2[8], 269)

McGrath, M. (2023). *Python in easy steps*.Access date: 23.07.2023. https://openlibrary.org/books/OL26976831M/Python_in_easy_steps

Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ, 3*, e127. https://doi.org/10.7717/peerj-cs.127

Murphy, K. P. (2006). Naïve Bayes classifiers. *University of British Columbia, 18*(60), 1–8

Narkhede, S. (2018). Understanding AUC-ROC curve. *Towards Data Science, 26*(1), 220–227.

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in Twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1-5). https://doi.org/10.1109/icccnt.2013.6726818

Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., & Kohli, P. (2011). Decision tree fields. In *International Conference on Computer Vision* (pp. 1668–1675). https://doi.org/10.1109/iccv.2011.6126429

Patel, N., Parekh, B., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., & Sharma, R. (2022). Fusion in cryptocurrency price prediction: A decade survey on recent advancements, architecture, and potential future directions. *IEEE Access, 10*, 34511–34538. https://doi.org/10.1109/access.2022.3163023

Pradana, A. T., & Hayaty, M. (2019). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on Indonesian-language texts. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 4(4), 375–380. https://doi.org/10.22219/kinetik.v4i4.912

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning*. https://cds.cern.ch/record/2031749

Raaijmakers, J. G., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology, 43*(1), 205–234.

Rahman, S., Hemel, J. N., Anta, S. J. A., & Muhee, H. A. (2018). Sentiment analysis using R: An approach to correlate Bitcoin price fluctuations with change in user sentiments. *BRAC University Institutional Repository*. http://dspace.bracu.ac.bd/xmlui/handle/10361/10163

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine, 47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Sallis, J., Gripsrud, G., Olsson, U., & Silkoset, R. (2021). *Research methods and data analysis for business decisions*. Springer International Publishing. https://doi.org/10.1007/978-3-030-84421-9

Sami, O., Elsheikh, Y., & Almasalha, F. (2021). The role of data pre-processing techniques in improving machine learning accuracy for predicting coronary heart disease. *International Journal of Advanced Computer Science and Applications, 12*(6). https://doi.org/10.14569/ijacsa.2021.0120695

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of Machine Learning*. Springer Science & Business Media.

Shah, D., & Zhang, K. (2014). Bayesian regression and Bitcoin. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (pp. 409-414). https://doi.org/10.1109/allerton.2014.7028484

Suthaharan, S. (2016). *Machine learning models and algorithms for big data classification*. Springer Nature. https://doi.org/10.1007/978-1-4899-7641-3

Tanwar, S., Patel, N. A., Patel, S., Patel, J., Sharma, G., & Davidson, I. E. (2021). Deep Learning-Based Cryptocurrency Price Prediction Scheme with Inter-Dependent relations. *IEEE Access, 9*, 138633–138646. https://doi.org/10.1109/access.2021.3117848

Vumazonke, N., & Parsons, S. (2023). An analysis of South Africa's guidance on the income tax consequences of crypto assets. *South African Journal of Economic and Management Sciences, 26*(1). https://doi.org/10.4102/sajems.v26i1.4832

Wang, H., Yao, Y., & Salhi, S. (2020). Tension in big data using machine learning: Analysis and applications. *Technological Forecasting and Social Change, 158*, 120175. https://doi.org/10.1016/j.techfore.2020.120175

Yogish, D., Manjunath, T. N., & Hegadi, R. S. (2019). Review on Natural Language Processing Trends and Techniques using NLTK. In *Communications in Computer and Information Science* (pp. 589–606). https://doi.org/10.1007/978-981-13-9187-3_53

Zhang, Y., Jin, R., & Zhou, Z. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0