Initial Coin Offerings Success Prediction Using Social Media and Large Language Models

Khouloud Safi Eljil [1,2,*], Essia Hamouda [3], and Farid Nait-Abdesselam [1]

[1] Computer Science Department, Université Paris Cité, France
[2] Higher School of Communications, University of Carthage, SUP'COM, Tunisia
[3] Department of Information and Decision Sciences, California State University San Bernardino, USA
Email: khouloud.safi-eljil@etu.u-paris.fr (K.S.E.); essia.hamouda@csusb.edu (E.H.);
farid.nait-abdesselam@u-paris.fr (F.N.-A.)
*Corresponding author

***Abstract*—Initial Coin Offering (ICO) is a fundraising method utilized by blockchain startups to raise capital by issuing and selling digital tokens to investors. ICOs have become widely popular for cryptocurrency fundraising, often generating millions of dollars, and surpassing traditional crowdfunding methods like Initial Public Offerings. However, ICO is a risky way of investing and raising capital due to the lack of regulations and standardisation. In this research, we delve into the impact of social media and sentiment analysis on the success of ICOs, employing various machine learning models and Large Language Models. Our analysis is based on data from over 1,000 ICOs gathered from diverse ICO information platforms, coupled with a corpus of 910,478 tweets associated with these ICOs. We extend our investigation to include other social media platforms such as BitcoinTalk, Telegram, Facebook, and Medium. Our analysis revealed that valuable insights regarding the success of ICOs can be derived by examining text sentiment and investigating metadata across these diverse social media channels.**

***Keywords*—token sales, web scraping, sentiment analysis, social media, Bidirectional Encoder Representations from Transformers (BERT)**

I. Introduction

Recently, it has become difficult to overlook the role of Artificial Intelligence (AI) and blockchain in the wave of the Fourth Industrial Revolution (4IR) [1]. Marwala and Xing [1] argues that the joint force of these two technologies could determine the depth and breadth of the 4IR. Blockchain technology enables new ways of organizing economic activities, reduces costs and time associated with intermediaries, and strengthens the trust in an ecosystem of actors [2]. Lately, people are becoming accustomed with the idea of digital money in the form of cryptocurrencies like bitcoin, where transactions are recorded on a secure, distributed ledger (blockchain). Along this comes a new concept: the blockchain-based token. A token is a digital representation of some shared value. A considerable number of startups have employed blockchain for fundraising, known as Initial Coin Offerings (ICOs). Ideally, technology startups generate and sell tokens via blockchain in exchange for well established cryptocurrencies, mostly bitcoin and Ethereum. This token sale, as known as ICO, entails a limited period of online sale of a predefined number of tokens to the public. The token generation event is generally a one-time event at the beginning of a project's lifetime. The creation and sale of tokens occur either on an existing blockchain, such as Ethereum, which is most common for ICOs, or on a new blockchain specifically created for a given ICO. The ICO was initiated in 2013, and since then, its volume has increased rapidly. By 2017, its funds had reached \$3.5 B surpassing traditional crowdfundings. This can be primarily attributed to the success of five prominent ICOs; Filecoin \$257 M, Tezos \$232 M, EOS \$185 M, Paragon \$183 M and Bancor \$153 M [3]. These figures highlight the significant interest from both ventures and investors in participating in such projects. Nevertheless, the level of success achieved by ICOs varies significantly across various projects. While certain ICOs manage to attract hundreds of millions of dollars in funding, many others struggle to raise any fund. Moreover, ICOs represent a risky method of investing and raising funds due to the lack of regulations, which increases the chances of scams [3]. According to research conducted by Statis Group, an ICO advisory firm, approximately 11% of investments made in ICOs were victimized by fraudulent projects "scams" [4]. Hence, it is pertinent to explore the fundamental factors that contribute to the success of ICOs.

In this study, we are monitoring the social media sentiment of ICOs from the time an ICO is launched all the way to the day it closes. The success of an Initial Public Offerings (IPO) or ICO is typically assessed after the fundraising period has concluded. However, by sensing the community engagement, and opinion in the different social media platforms during the fundraising, the ICO initiators may take several actions to improve participation and attract more interest in their project to avoid a potential failure. This can involve, intensifying their marketing efforts to raise awareness about their ICO through targeted advertising campaigns, social media promotions, and

Manuscript received June 21, 2024; revised July 10, 2024; accepted August 13, 2024; published January 9, 2025.

doi: 10.12720/jait.16.1.12-20

outreach to potential investors through various channels. Moreover, offering incentives and bonuses to early investors can encourage participation. These incentives may include bonus tokens, discounted prices, or exclusive benefits for early contributors.

However, in the domain of entrepreneurship and early-stage financing, the notion of success lacks a precise definition. Thus, studies use diverse approaches to evaluate funding success. Czaja and Röder [5] use the collected fund during the ICO as the dependant variable. This suggests that the more funds an ICO collects, the more successful it is. However, this perspective is not entirely accurate. For instance, consider a startup with a funding goal of $10,000,000 but manages to raise only $1,000,000. This represents only 10% of their target. On the other hand, another smaller startup with a funding goal of $10,000 and successfully collects it, that is a success irrespective of the small amount, as it fulfills their asserted need to start their venture. This is the second approach of defining success in relation to the funding goal, which will be used in this research. Unfortunately, ICOs tend to disclose either a "soft cap" or a "hard cap". The soft cap represents a minimum threshold, and if this target is not reached during the ICO, it usually results in a complete refund of all investments made. To avoid such a scenario, ICO initiators often set the soft cap at a relatively low level. On the other hand, the hard cap denotes the maximum amount of investment permitted by the ICO's smart contract algorithm. Consequently, both the soft cap and hard cap are insufficient benchmarks for evaluating ICO funding success [5]. A standard ICO usually begins with the creation of a white paper [6]. It typically elucidates the underlying products or services, introduces the project team, and outlines a business plan. It also encompasses details related to the token sale, including the quantity of tokens, the duration of the sale, and how the raised funds will be utilized [7]. Soon after, the initiators of the ICO leverage social media platforms to actively promote their project. The marketing efforts primarily revolve around activities conducted exclusively on social media platforms. Moreover, most startups introduce their project on ICO information platforms. Typically, the entirety of an ICO marketing endeavors is confined to online channels [5]. Furthermore, in a qualitative research study, researchers conducted semi-structured interviews to examine signals affecting ICO success. The findings indicate the significance of social media discussions within the blockchain community, owing to the decentralized nature of the community's ideology. Surprisingly, some investors rely solely on sentiment as their decision-making criterion [8]. For this reason, we believe that by tapping into social media channels and conducting sentiment analysis, we can uncover valuable insights regarding the potential success of ICOs. To that end, we have collected and analyzed data from Twitter, BitcoinTalk, Telegram, Facebook, and Medium, while applying sentiment analysis to predict ICO success.

In this paper, we propose a combined approach based on classification models such as Support Vector Machines (SVM), logistic regression (LR), and random forest, along with sentiment analysis using Large Language Models (LLM), to identify significant features to predict ICO's success. In the following we highlight our main contributions:

- We scraped and analyzed data of over 1,000 ICOs gathered from diverse ICO information platforms, coupled with a corpus of 910,478 collected tweets associated with these ICOs.
- We enhanced our understanding of human opinions and sentiment analysis by fine-tuning Bidirectional Encoder Representations from Transformers (BERT) [9], a groundbreaking LLM, distinguishing our approach from previous studies in ICO literature.
- We predicted ICO success using the extracted sentiments and other attributes from social media channels. Various machine learning models were explored and compared, including SVM, logistic regression, an ensemble learning method, and several boosting algorithms. SVM outperformed all other models.
- We used SHapley Additive exPlanations (SHAP) to explain the SVM model. We found that Twitter and Bitcointalk data exert the greatest influence on ICO success compared to other social channels. This finding is particularly noteworthy since, to our best knowledge, Bitcointalk, despite its prominence in the cryptocurrency domain, has not been previously explored in ICO literature.
- We demonstrate a relative increase in accuracy of 12.5% in our designed model compared to models documented in existing literature that use social channels for ICO success prediction.
- We found that the text analysis of Bitcointalk forums revealed trust as the predominant emotion, highlighting the importance for ICO initiators to cultivate investor trust to enhance token sales.

The paper is organized as follows: Section II discusses related work. Section III describes the proposed methodology. Section IV presents our results. Section V summarizes our work.

II. Related Work

The aim of this literature review is to outline the primary research advancements in ICO, with a particular emphasis on social factors. Dürr *et al.* [10] built a predictive model to identify ICO fraud, compiling data on ICOs in the form of white-papers and implementing both natural language processing techniques and machine learning algorithms. In similar context, Cerchiello *et al.* [11] employed logistic regression and classification tree to highlight the key aspects of an ICO that are significantly related to fraudulent behaviors. For this, they have analyzed a sample of 120 ICOs, they concluded that the absence of a website and of the Twitter account for an ICO is a driver of scam activity suspect. Meoli and Vismara [12] applied machine learning models along text-analysis of white-papers to predict the success or failure of ICOs. Jin *et al.* [3] used social media data in addition to the financial factors collected from ICOs information platform to build their logistic model that predicts the total failure of ICOs (i.e., the start-up could not sell any of their issued tokens). They calculated correlation between the existence or not of a given startup social channel and the ICO result. Their result indicates that the existence of Twitter account for a given ICO has the highest influence on ICO success, when compared with other social channels (Facebook, Youtube, LinkedIn, telegram, blog, and others). For Twitter, they merely analyzed the metadata and surprisingly they found a very week correlation between no. of followers and no. of tweets with ICO result, after reviewing 400 ICO startup Twitter channels. This result contradicts with Albrecht *et al.*'s [13] results. Specifically, Albrecht *et al.* [13] argue that the number of people following a blockchain startup on Twitter (Nfollowers) is highly significant and positively linked to ICOs' success in terms of raised volume, we find their finding persuasive. Moreover, they stated that the total number of tweets (NTweets) is statistically significant. In addition to using startups Twitter's metadata, they have analyzed the sentiment of the Twitter account of the company issuing the cryptographic tokens, but not the public tweets (opinion) regarding a given ICO. Chursook *et al.* [14] studied ICOs from Australia and Singapore, only 87 ICOs were analyzed. They employed sentiment analysis on tweets using the "lexicon" package in R, a dictionary-based method where sentiment is assigned based on predefined words. This method, reported to be less efficient compared to a trained classifier that is able to capture contextual nuances specially the negation. The highest reported accuracy was 57.7% using logistic regression. Chursook *et al.* [15] extended their work and added expert ratings to the sentiment analysis, however their results are questionable, as there is a mismatch between the reported confusion matrices and their performance metrics. For example, they are claiming an accuracy of 91.2% for Naive Bayes while the corresponding confusion matrix shows true positive of 22, true negative of 13, false positive of 10, and false negative of 12, doing the math will lead to 61.40% accuracy, this applies for all the other confusion matrices of the different models.

III. The Proposed Methodology

In this section, we discuss our proposed method as depicted in Fig. 1. We start by web scraping the ICO information platform, and Twitter to get the metadata of startup accounts, and the public tweets regarding each ICO using the hashtag and ICO name. The collected tweets will be automatically annotated to distinguish between positive and negative sentiments using a sentiment analysis model based on BERT. This model, capable of understanding context and complex sentences, is further enhanced through a thorough preprocessing stage. This is particularly crucial for unstructured short text messages, such as those found on the Twitter platform, which pose several challenges. Among these challenges are informal conversational elements (e.g., repeated letters, emoticons and emojis) and extraneous information like noise, links, and user mentions. In addition to that, we scraped the metadata of four other social media platforms, namely, BitcoinTalk, Facebook, Telegram, and Medium. All collected data will be fed to supervised machine learning models to be trained accordingly. Finally, we assess and compare the different models. We have selected those five social media channels, as they are the most used social channels in our dataset as depicted in Fig. 2.

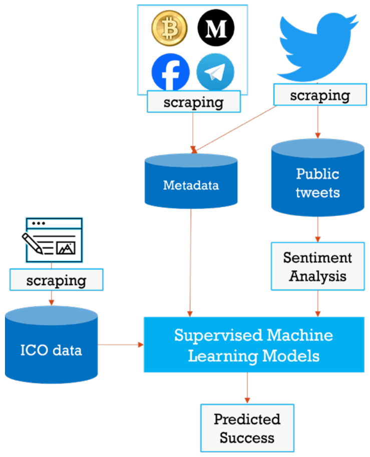

Fig. 1. Research methodology.

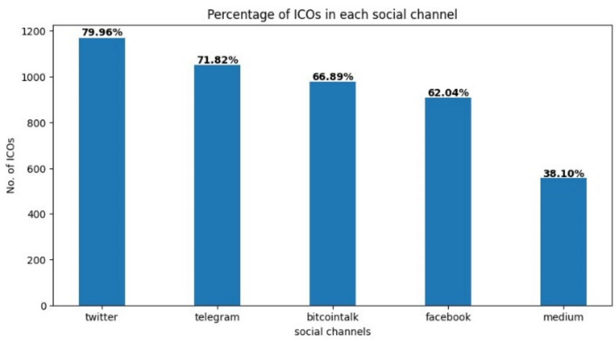

Fig. 2. Top five social channels used by ICOs initiators; ordered from left to right (The values displayed at the top of each bar represent the percentage of ICOs initiators that use a given social channel according to our dataset).

A. Data Collection and Analysis

Our dataset was originated from the icoholder website [16], an online platform providing comprehensive information on ICOs, in the version of October 2022. We have built a web scraper in python using Beautifulsoup [17] and Selenium [18] libraries. Initially the dataset had 10,234 token sales, next we filtered ICO pre-sales and private sales from our sample, as we are only interested in ICO main sale. Additionally, we have excluded any ICOs that failed to provide information about their project's funding goal. As mentioned earlier, most ICO initiators typically disclose the soft cap and/or hard cap, which is not relevant to this research.

Some ICOs report the goal in ethereum or bitcoin instead of USD, so for unification, we did the exchange to USD according to the exchange rate at the time of the ICO event. Additionally, if icoholder.com does not provide the actual amount of funds collected for a specific ICO, we strive to obtain this information from other ICO information platforms. If we are unable to find this information, we exclude the ICO from our study. Finally, we ended up with 1462 ICOs, and the attributes are: ICO name, start date, end date, raised fund, goal, social media channels, and review (number of ICO influencers who reviewed a given ICO). For social media channels, there are 21 alternatives (Twitter, Telegram, Bitcointalk, Facebook, Medium, Youtube, Reddit, etc.). We count for each ICO, how many social media channels it has. Fig. 2 shows the top five used social channels by ICO initiators according to our dataset, ordered from left to right. Where the most used social channel by the different ICO initiators is Twitter, namely 79.96% of ICOs in our dataset have Twitter accounts. Hence, those five social channels are utilized in predicting ICO success. Regarding the output variable "success", it is Boolean. As mentioned earlier we use the goal as the success determinant. If the ICO was reported as scam, it was terminated for some reason, or the raised fund is less than the goal then we consider it as failure "0". It will be "1" only if the collected fund is greater than or equal to the reported goal (see instruction 1). Our dataset is balanced, i.e., 49.6% of the ICOs are success (1), and 50.4% are failure (0). Table I presents a comprehensive overview of the features utilized in our prediction model that will be explained next, sourced from various social media channels, sentiment analysis, and ICO information platforms, alongside summary statistics (mean, Standard Deviation (SD), minimum, and maximum).

$$success = raised_{fund} >= goal \ ? \ 1 : 0 \qquad (1)$$

TABLE I. THE USED FEATURES WITH SUMMARY STATISTICS (NUMBER OF SAMPLES IS 953 ICOS)

Features			Data Type	Mean	SD	Min.	Max.
input	ICO information platform	review	number	0.75	1.55	0	14
		ICO duration (days)	number	63.47	108.17	0	2019
		total_social_channels	number	5.38	2.04	1	11
	Twitter	Twt Followers	number	6818.87	23725.87	0	357753
		Twt Friends	number	502.25	1157.06	0	9702
		Twt Status Count	number	1874.52	27312.08	0	822000
	Sentiment Analysis for Twitter	No. Pos Tweets	number	662.06	2433.07	0	41116
		No. Neg Tweets	number	203.55	844.42	0	14224
		Total No. of Tweets	number	865.61	3017.26	0	41668
	Bitcointalk	BTCTalk number_pages	number	37.88	107.55	0	1993
		BTCTalk number_reads	number	30868.56	132980.04	0	2983605
		BTCTalk Merit	number	127.16	288.49	0	1639
		BTCTalk Activity	number	257.87	476.71	0	3080
		BTCTalk total_Admin_Comments	number	42.67	116.14	0	1838
		BTCTalk total_comments	number	746.87	2155.79	0	40143
	Telegram	Telegram number of members	number	604.13	2990.45	0	69615
		Telegram photos	number	15.54	357.34	0	11000
		Telegram videos	number	0.18	1.52	0	29
		Telegram links	number	19.14	360.34	0	11000
		Telegram files	number	0.13	2.89	0	88
	Facebook	FB likes	number	11263.41	181046.28	0	5500000
		FB followers	number	11393.95	181048.87	0	5500000
		FB rating	number	1.28	2.13	0	5
		FB number of reviewers	number	38.44	178.62	0	4050
	Medium	medium followers	number	323.12	1569.78	0	26000
	Dummy Variables	Bitcointalk dummy	boolean	-	-	-	-
		Telegram dummy	boolean	-	-	-	-
		Facebook dummy	boolean	-	-	-	-
		Medium dummy	boolean	-	-	-	-
Output	Success		boolean	-	-	-	-

- Twitter: to gather Twitter data, we used SNScrape [19], a python library that utilizes web scraping techniques instead of relying on APIs. We eliminated ICOs that either lacked Twitter accounts or had their accounts closed or suspended, resulting in a final selection of 953 ICOs. Subsequently, we performed web scraping to gather metadata from their respective Twitter accounts, including username, display name, number of followers, number of friends (following), and status count. Additionally, in order to evaluate the level of community engagement and opinions regarding each ICO, we conducted three search queries on public tweets, as illustrated in Fig. 3. These search queries involved utilizing the ICO Twitter's username both with and without the "#" symbol, as well as with and without the "ICO" postfix. By specifying the start and end dates of each ICO based on our dataset, we gathered tweets related to a given ICO while it is taking place. The resulting tweets were stored in separate CSV file for each ICO, culminating in a total of 953 CSV files, and a corpus of 910,478 collected tweets. Those collected tweets lack annotation, the details of sentiment analysis are explained in the next subsection.

```
search1 = f'{row["twitter_username"]} until:{row["end date"]} since:{row["start date"]}'
search2 = f'#{row["twitter_username"]} until:{row["end date"]} since:{row["start date"]}'
search3 = f'#{row["twitter_username"]}-ICO until:{row["end date"]} since:{row["start date"]}'
```

Fig. 3. Search queries used to collect public tweets about a given ICO.

- BitcoinTalk: is one of the earliest and most well-known online forums dedicated to discussions about bitcoin and other cryptocurrencies. It was created by Satoshi Nakamoto, the pseudonymous creator of bitcoin [20]. The BitcoinTalk forum has specific sections where cryptocurrency projects can announce and discuss their ICOs. When an ICO announces its project on the BitcoinTalk forum, the corresponding announcement thread becomes a central hub for providing information about the project and engaging with the community. We have extracted some useful data from each thread of our ICOs. Fig. 4 shows the thread of an ICO called 0xcert, we have highlighted and numbered some of the collected data. We have developed a Python web scraper utilizing the Beautifulsoup and Selenium libraries to parse through the forum and extract valuable information. The first attribute that we crawled is the "number of reads" which is the number of times a particular thread has been viewed by users. This metric provides an indication of the level of interest or engagement the ICO thread has generated within the BitcoinTalk community. The second attribute is "total number of pages". Each page typically contains a certain number of posts, and as the discussion progresses, more pages are added to accommodate additional contributions from forum members. The number of pages can provide insights into the level of activity and engagement within the thread. We also have created two more variables that can as well reflect the activity and engagement of Bitcointalk community. Namely, "BTCTalk total_comments" which shows how many posts are there inside the thread. This was calculated by parsing through all the pages and counting in each page how many comments/posts are there and summing them up all, the maximum number of pages we had in our dataset is 1993. And "BTCTalk total_Admin_Comments", specifically tallies the posts related to the thread creator/ICO initiator. This metric is utilized to demonstrate the level of commitment and engagement exhibited by ICO initiators in the blog, including their efforts in sharing updates and addressing the concerns of the community. The last two attributes related to the thread creator, i.e., ICO initiator in our case, are "Activity" and "Merit". Those numbers show the user ranking and reputation in Bitcointalk forum. Namely, "Activity" is a measure of a user's overall participation and contribution to the forum. "Merit", on the other hand, is a measure of the quality and usefulness of a user's posts, as perceived by other community members. Apart from gathering and computing these attributes, we conducted textual analysis on certain ICO threads, which will be elucidated in the following section.
- Telegram: an account in Telegram can be a Simple user, a group, or a channel, which is a one-way communication where the creator broadcasts messages to the audience. For ICO Telegram accounts that are channels, we crawled number of subscribers, photos, videos, links, and files. In cases where a value was unavailable, we recorded zero as the corresponding value. If it is a group, we can only scrape number of members. However, for private groups, this information is not publicly accessible, resulting in the inability to retrieve it. Finally, if the ICO account is simple user, there is nothing to be scraped. Unfortunately, almost half (46.18%) of the ICO telegram accounts in our dataset are either Private Group or Simple User.
- Facebook: the ICO Facebook accounts are either pages or groups. For groups, we scraped number of members. While for pages, number of likes, number of followers, rating, and number of reviewers were gathered.
- Medium: some ICOs use the Medium platform as a means to share informative and engaging articles about their projects. For medium, we scraped number of followers for the official Medium accounts of ICOs. However, it is noteworthy that over half of the ICOs in our dataset (64.84%) do not possess Medium accounts.

Fig. 4. Example of Bitcointalk thread with some important collected attributes.

B. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a natural language processing technique aimed at computationally determining the sentiment or emotion conveyed within a text. In the literature sentiment analysis employs two distinct approaches:

- Dictionary-based approach: determining sentiment involves tallying the matches between a predefined list of positive and negative words and the terms within a given text source (e.g., a tweet, a sentence, or a paragraph).
- Classifier-based approach: a proper classifier is trained on a sufficiently large dataset with pre-labeled examples, enabling it to predict the sentiment class of new examples. Those classifiers can leverage traditional machine learning models or more advanced deep learning models, among which are LLM, like BERT that was pre-trained on vast amounts of text data and can capture complex relationships within human language.

Dictionary-based sentiment analysis is straightforward and easier to interpret but lacks contextual understanding, leading to less accurate sentiment prediction. On the other hand, classifier-based methods, especially those leveraging advanced language models such as BERT, provide superior accuracy and contextual comprehension but are computationally intensive and require annotated datasets for training. We mainly leveraged BERT model to automatically annotate all of the collected tweets of each ICO, benefiting from its exceptional performance, especially given the concise nature of tweets and the availability of annotated Twitter datasets online. However, when dealing with longer texts in BitcoinTalk forums that span multiple pages, we opted for a dictionary-based approach.

- Twitter Textual Analysis: We employed the BERT model to the collected tweets, as detailed in our previous publication [21]. Where, given the inherent noise and peculiarities in Twitter data, we introduced an enhanced pre-processing phase. This phase involved tasks such as data cleaning, removal of user mentions, as well as handling emoticons and emojis. Additionally, we conducted comprehensive training and comparison experiments, encompassing both traditional machine learning models and BERT. Among these, the BERT models demonstrated outstanding performance attaining an accuracy ranging from 92.92% to 94%. Consequently, we preserved the trained model and applied it in the current study. Concretely, we applied the recommended preprocessing techniques outlined in [21] to the gathered tweets. Additionally, in this study, we introduced an additional step, utilizing the Google Translate API to render non-English tweets into English. Due to the presence of non-English tweets, and as our sentiment analysis model was trained on English data, translation was imperative. Following the completion of the pre-processing phase, we proceeded to employ the pretrained BERT model for the automatic annotation of each collected tweet. This entire procedure was iterated for every tweet across the 953 CSV files. After annotation, we have acquired three attributes from each file for each ICO. Namely, number of positive tweets, number of negative tweets, and total number of collected tweets. Where, for "number of positive tweets" attribute, we have counted in every CSV file how many tweets were auto-annotated as positive "1", similar task was done for "number of negative tweets" but this time counting the negative tweets (label 0). Regarding "total number of collected tweets" attribute is just the summation of the two attributes.
- Bitcointalk Textual Analysis: To analyze the text of Bitcointalk forums using dictionary-based model, we employed three vocabularies from the R package "tidytext": AFINN [22], BING [23], and NRC Emotional Lexicon [24]. These three lexicons are based on unigrams (bag of words). Where AFINN lexicon is a list of English words each assigned a numerical sentiment score (−5 to 5). Where negative scores imply negative sentiment and vice versa. The BING lexicon categorizes words in a binary fashion into positive and negative. However, The NRC Emotional Lexicon contains words annotated with their association to eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust).

C. *Supervised Machine Learning Models*

To identify the optimal model for predicting ICO success, we have investigated various machine learning classifiers. We started with exploring the Naive Bayes (NB) model, a linear probabilistic model based on Bayes' theorem [25], as a baseline model. Subsequently, we examined Decision Trees (DT) [26] and evaluated Logistic LR [27] as a generalized linear model. The optimization algorithm used in LR is Limited-memory Broyden–Fletcher–Goldfarb–Shanno "LBFGS" solver, to prevent overfitting, we have used L2 penalty as the regularization technique. Additionally, we explored SVM, a non-probabilistic linear model capable of solving both linear and non-linear problems [28]. We employed a linear kernel for SVM and set the regularization parameter (C) to "1". We also considered Random Forest (RF) [29], an ensemble learning method that combines multiple decision trees and uses voting to determine the final prediction. The number of trees in the forest (n_estimators) is set to "100". Lastly, we evaluated two boosting algorithms [30], namely AdaBoost and Gradient Boosting Classifier (GBC). Boosting algorithms combines multiple weak models to create a stronger model, by iteratively correcting the mistakes or misclassifications made by previous models. For Adaboost, we set the maximum number of estimators (n_estimators) to "100" and the learning rate to "1", using the Stagewise Additive Modeling using a Multiclass Exponential (SAMME) algorithm, which is a discrete boosting method. For GBC, we configured the hyperparameters as follows: the number of boosting stages (n_estimators) is set to "80", the learning rate is "1", and the maximum tree depth (max_depth) is set to "1". It is worth mentioning that we have explored some deep learning models, namely Convolutional Neural Network (NN), Radial Basis Function (RBF) NN, and Multilayer Perceptron (MLP) NN. However, their results were not promising, that is why they weren't reported in the results section. To assess the performance of the used models, we split the dataset into an 80% training set and a 20% testing set. Our finalized dataset has 953 ICOs. The used attributes are as mentioned previously, refer to Table I. Precisely, social media channels, review, ICO duration in days (this was calculated using end date and start date). Twitter related data: No. of followers, No. of friends, status count, No. of positive tweets, No. of negative tweets, and total number of collected tweets. Bitcointalk related data: number_pages, number_reads, author Merit, author Activity, total Admin Comments, and total comments. Telegram related data: No. of members, photos, videos, links, and files. Facebook data: No. of likes, followers, rating, and No. of reviewers. Finally, No. of followers of Medium. Any feature/column that has more than half missing values, was dropped. We have also introduced four dummy variables to signify the presence of the most frequently used social media channels in ICOs—Facebook, Telegram, Bitcointalk, and Medium. A dummy variable for Twitter was omitted since all 953 ICOs in our finalized dataset already have Twitter accounts.

IV. RESULTS AND DISCUSSION

After creating the features for the classification models and choosing the suitable machine learning models, we determine the performance metrics tailored to the specific problem. Accuracy (expressed by Eq. (2)), an oft-cited metric for classification evaluation, reflects the percentage of correctly assigned class labels. Precision, recall, and F1-Score are other commonly used metrics [31]. Where precision (expressed by Eq. (3)) focuses on the accuracy of positive predictions, indicating the proportion of correctly predicted positive instances out of all instances predicted as positive. In contrast, recall (Eq. (4)) also known as sensitivity emphasizes the model's ability to capture all positive instances, indicating the proportion of correctly predicted positive instances out of all actual positive instances. Precision evaluates the accuracy of positive predictions, while recall evaluates the comprehensiveness of positive predictions. F1-Score (Eq. (5)), a harmonic means of precision and recall, balances both metrics and is beneficial for evaluating overall model performance, particularly in imbalanced datasets. In Eq. (2)–Eq. (5) [31], *TP*, *TN*, *FP*, and *FN* represent "True Positive", "True Negative", "False Positive", and "False Negative", respectively. We let $TotalPrediction = TP + TN + FP + FN$.

$$\text{accuracy} = \frac{TP + TN}{TotalPrediction} \tag{2}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

TABLE II. PERFORMANCE COMPARISON OF THE PROPOSED MACHINE LEARNING MODELS ON THE TEST DATASET

Model	Accuracy	Precision	Recall	F1-Score
NB	54.45%	68.00%	32.38%	43.87%
DT	55.50%	60.00%	57.14%	58.54%
LR	63.35%	68.42%	61.90%	65.00%
SVM	64.92%	64.84%	79.05%	71.24%
RF	62.83%	67.35%	62.86%	65.02%
AdaBoost	59.69%	66.28%	54.29%	59.69%
GBC	59.69%	63.92%	59.05%	61.39%

The results are presented in Table II, which compares the performance of the seven ML models on the test dataset. The model with the lowest performance was the baseline model (NB), with an F1-Score of 43.87%. On the other hand, the SVM model had the best performance, with an F1-Score of 71.24%. Following SVM, LR and RF as an ensemble learning method, both exhibited relatively good performance with F1-Scores of 65%. The obtained accuracy of SVM is sufficient for a preliminary assessment, especially when compared with [14], where our designed model demonstrates a relative increase in accuracy of 12.5%. However, we suggest that further research is needed with other factors like team members, white-papers, and others to determine if the accuracy will improve by customising the training model of the machine learning platforms. Inevitably though, insights regarding the success of ICOs can be derived by examining text analysis and other social factors as will be shown next.

A. Model Interpretation

The decision function of an SVM involves complex calculations in high-dimensional spaces, making it less intuitive to interpret and find the feature importance within the model. Thus, we have employed SHAP, a state-of-the-art tool capable of reliably computing and visualizing the impact of features on the output of a machine learning model based on game theory [32]. It aims to provide insights into the contribution of each input feature to the models' predictions, helping to shed light on the black-box model's behavior. To get an oversight of the most important features for our model, we depict the SHAP values of the most influential features in our dataset in Fig. 5.

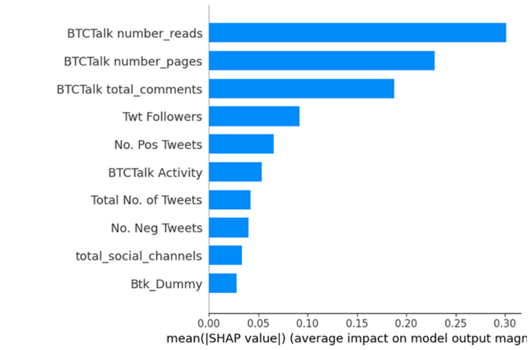

Fig. 5. SHAP bar plot shows the magnitude of top 10 features' effect, sorted by feature importance.

Using SHAP bar plot (Fig. 5), "BTCTalk number_reads" was identified as the most important predictor of ICO success, reflecting the ICO initiators' ability to attract a substantial audience to read or engage in the forum. This finding is in concordance with the results of [13], where they showed that the number of people following a blockchain startup on Twitter (Nfollowers) is highly significant and positively linked to ICOs' success. The two attributes "BTCTalk number_reads", and Nfollowers reflect the overall effectiveness of a company's social media marketing strategy. No. of pages, and total comments of Bitcointalk, along with the No. of positive tweets stand out as key features among the top five. These attributes signify societal engagement and positive feedback concerning ICO initiators. In essence, a higher level of community engagement with extensive comments and positive tweets correlates with an increased likelihood of ICO success. It is worth highlighting that within the top ten important features, only data associated with Twitter and Bitcointalk is present, underscoring the significance of these social channels in comparison to

Facebook, Telegram, and Medium. This observation is further validated by the dummy variables, with only the presence of Bitcointalk proving significant. As mentioned earlier, no dummy variable was employed for Twitter as all ICOs in our finalized dataset have Twitter accounts. This finding is especially significant because, to our best information, Bitcointalk, despite its significance in the cryptocurrency domain, has not been previously examined in ICO literature. This is a call for researchers in ICO literature to conduct further analysis of the Bitcointalk forum.

B. Bitcointalk Textual Analysis

Further analysis has been conducted on the textual part of Bitcointalk forum as detailed in Section III-B. For the sake of simplicity and readability we report an example of ICO for each category: Aeternum as an illustration of a successful ICO, and Shamrock as an illustration of failure. Figs. 6 and 7 compare the most frequent positive and negative words in the success and failed ICOs respectively. It is notable that the proportion of positive words (refer to the percentage) is more than the negative ones in the success ICO, the opposite is true in the failed one. Moreover, we have calculated the average sentiment score based on AFINN vocabularies, 1.08 and 0.4 were the scores for success and failed ICOs, respectively. It appears evident that the successful ICO has a higher score than the failed one. The most three dominant emotions for the success ICO according to NRC Emotional Lexicon, were trust, anticipation and joy. For the failed ICO, trust, fear, and sadness were the most appearing ones. This is enlightening for the ICO initiators where they should gain the trust of the investors in order to sell more tokens. Moreover, it is observable that negative feelings were tied with failed ICO, and happy emotions in the successful ICO.

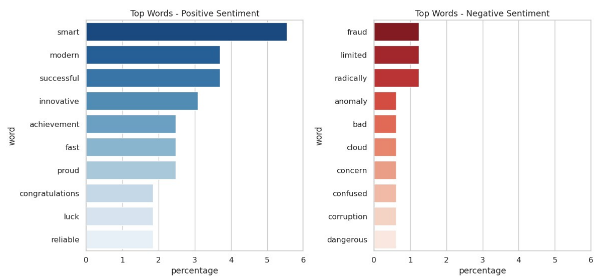

Fig. 6. Success ICO (Aeternum), most frequent positive and negative words.

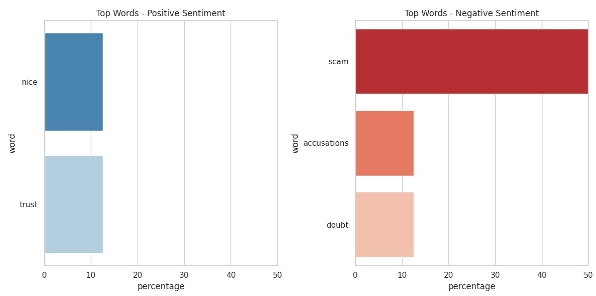

Fig. 7. Failed ICO (Shamrock), most frequent positive and negative words.

V. CONCLUSION

In this paper, we address the issue of discovering the social success drivers of an ICO. For this purpose, we have scraped and mined financial and ICOs' social media data, namely, Twitter, BitcoinTalk, Telegram, Facebook, and Medium. We proposed a combined approach based on classification models and textual analysis to highlight significant attributes in distinguishing success from failure ICOs. Precisely, we leveraged BERT, an advanced language model, to enhance sentiment analysis of collected public tweets to grasp people opinion toward each ICO, along other attributes collected from different social media were fed to different machine learning models. SVM has outperformed other models. To interpret and elucidate the work of SVM model, we employed the SHAP tool for conducting feature importance analysis. We found that the primary social channels associated with the success of ICOs are Twitter and Bitcoinalk. Further analysis has been conducted on the textual part of Bitcointalk forum, where we have illustrated that more negative words appear with high frequency in failed ICO compared to a successful one. This trend is is further confirmed by a sentiment score calculated using AFFIN vocabulary. Moreover, we studied the emotions in Bitcointalk, the most three dominating emotions found in the success ICO were trust, anticipation, and joy. For the failed ICO, trust, fear, and sadness were the most appearing ones.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Khouloud Safi ElJil formulated the research question, designed the methodology, collected and analyzed the data, conducted experiments, and interpreted the results. She also performed a thorough literature review and took the lead in drafting the manuscript, incorporating feedback from the advisors (Dr. Essia Hamouda, and Prof. Farid Nait-Abdessalam) and addressing reviewer comments. The advisors provided guidance, intellectual input, supervision, and critical feedback, helping refine the research objectives, edit the manuscript, and secure necessary resources. All authors approved the final version.

REFERENCES

[1] T. Marwala and B. Xing, "Blockchain and artificial intelligence," arXiv preprint, arXiv:1802.04451, 2018.

[2] J. Weking, M. Mandalenakis, A. Hein, S. Hermes, M. Böhm, and H. Krcmar, "The impact of blockchain technology on business models—A taxonomy and archetypal patterns," *Electronic Markets*, vol. 30, pp. 285–305, 2020.

[3] S. Jin, R. Ali, and A. Vlasov, "Cryptoeconomics: Data application for token sales analysis," in *Proc. 2017 International Conference Information Systems Special Interest Group on Big Data*, 2017.

[4] S. Dowlat and M. Hodapp. (2018). Cryptoasset market coverage initiation: Network creation. Satis Group. [Online]. Available: https://research.bloomberg.com/pub/res/d37g1Q1hEhBkiRCu_ruMdMsbc0A

[5] D. Czaja and F. Röder, "Signalling in initial coin offerings: The key role of entrepreneurs' self-efficacy and media presence," *Abacus*, vol. 58, no. 1, pp. 24–61, 2022.
[6] D. Florysiak and A. Schandlbauer, "The information content of ICO white papers," ERN: Foreign Exchange Models, 2019.
[7] S. Adhami, G. Giudici, and S. Martinazzi, "Why do businesses go crypto? An empirical analysis of initial coin offerings," *Journal of Economics and Business*, vol. 100, pp. 64–75, 2018.
[8] M. Yadav, "Exploring signals for investing in an Initial Coin Offering (ICO)," *SSRN*, 2017. http://dx.doi.org/10.2139/ssrn.3037106
[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," arXiv preprint, arXiv:1810.04805, 2018.
[10] A. Dürr, M. Griebel, G. Welsch, and F. Thiesse, "Predicting fraudulent initial coin offerings using information extracted from whitepapers," in *Proc. ECIS*, 2020.
[11] P. Cerchiello, P. Tasca, and A. M. Toma, "ICO success drivers: A textual and statistical analysis," *The Journal of Alternative Investments*, vol. 21, no. 4, pp. 13–25, 2019.
[12] M. Meoli and S. Vismara, "Machine-learning forecasting of successful ICOs," *Journal of Economics and Business*, vol. 121, 106071, 2022.
[13] S. Albrecht, B. Lutz, and D. Neumann, "How sentiment impacts the success of blockchain startups—An analysis of social media data and initial coin offerings," in *Proc. the 52nd Hawaii International Conference on System Sciences*, 2019.
[14] A. Chursook, N. Naktnasukanjn, S. Chaimaim, P. Udomwong, J. Chatsirikul, and N. Chakpitak, "Can tweets predict ICO success? sentiment analysis for success of ICO whitepaper: Evidence from Australia and Singapore markets," in *Proc. the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2020.
[15] A. Chursook, A. Y. Dawod, S. Chanaim, N. Naktnasukanjn, and N. Chakpitak, "Twitter sentiment analysis and expert ratings of initial coin offering fundraising: Evidence from Australia and Singapore markets," *TEM Journal*, vol. 11, no. 1, 44, 2022.
[16] ICOhodler. [Online]. Available: https://www.icoholder.com/
[17] L. Richardson. Beautiful soup documentation. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[18] Seleniumdocumentation. [Online]. Available: https://www.selenium.dev/documentation/en/
[19] Just Another Archivist: Snscrape: A python library for scraping social media data. [Online]. Available: https://github.com/JustAnotherArchivist/snscrape
[20] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, 2008.
[21] K. S. Eljil, F. Nait-Abdesselam, E. Hamouda, and M. Hamdi, "Enhancing sentiment analysis on social media with novel preprocessing techniques," *J. Adv. Inf. Technol.*, vol. 14, no. 6, pp. 1206–1213, 2023.
[22] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," arXiv preprint, arXiv:1103.2903, 2011.
[23] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.
[24] S. M. Mohammad and P. D. Turney, "NRC emotion lexicon," *National Research Council Canada*, vol. 2, 234, 2013.
[25] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *Proc. the Workshop on Empirical Methods in Artificial Intelligence, IJCAI 2001*, 2001, vol. 3, pp. 41–46.
[26] B. de Ville and P. Neville, "Decision trees for analytics: Using SAS enterprise miner," *SAS Institute Cary*, 2013.
[27] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer, 2001.
[28] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
[29] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*, Springer, 2012.
[30] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Proc. Computational Learning Theory: Second European Conference, EuroCOLT'95*, Barcelona, Spain, Springer, 1995, pp. 23–37.
[31] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press, 2011.
[32] S. M. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.