International Journal of Mathematical, Engineering and Management Sciences
Vol. 9, No. 4, 902-913, 2024
https://doi.org/10.33889/IJMEMS.2024.9.4.047

Ram Arti
Publishers

A Hybrid Model of Integrating Sentiment Analysis and Key Market Indicators for IPO Listing Trend Prediction

Ashish Garg
Department of Computer Science and Engineering,
Graphic Era Deemed to-be University Dehradun, Uttarakhand, India.
Corresponding author: ashish.garg@geu.ac.in

Kamal Kumar Ghanshala
Department of Computer Science and Engineering,
Graphic Era Deemed to be University Dehradun, Uttarakhand, India.

Sachin Sharma
Department of Computer Science and Engineering,
Graphic Era Deemed to be University Dehradun, Uttarakhand, India.

(Received on December 29, 2023; Revised on February 29, 2024 & April 17, 2024; Accepted on April 22, 2024)

Abstract

Initial Public Offerings (IPOs) provide great opportunities for companies to grow and expand, and they allow investors to invest their money wisely to get a decent Return on Investment (ROI) in the short term. Nevertheless, the intricate nature of the stock market is susceptible to several influences such as a company's financial statement, governmental regulations, and public sentiment, which hinders the attainment of a satisfactory ROI. This study aims to address this challenge by developing a model that combines public sentiment analysis and machine learning approaches to optimize the ROI for IPO trends. The study gives a novel approach that uses multiple different features associated with IPOs like the public opinion, grey market price (GMP), issue prices, lot size etc. and leverages the use of various machine learning techniques like Random Forest, Decision Tree, Naive Bayes, and K-Nearest Neighbour (KNN) to make well-informed investment recommendations. The testing results demonstrate that the Decision Tree method surpasses the other algorithms, with an accuracy rate of 82.3%. This discovery emphasizes the effectiveness of our method in forecasting the success of IPOs by utilizing a combination of sentiment analysis and crucial financial indicators.

Keywords- Sentiment analysis, IPO, Machine learning algorithm, GMP, Decision tree, Issue prices.

1. Introduction

An IPO is a significant turning point in a business's life cycle. It happens when a privately owned company offers its shares to the public for the first time, or when the ownership of the company is shifting from private to public ownership. An IPO is a big step as now the company has the chance to raise a large amount of money. Companies generally initiate an IPO to raise huge capital, value assessment, company growth, enhance price transparency credibility, expand their operations and for marketing purposes. Whenever a company goes public or raises an IPO, the previously owned private shares are converted to public ownership and existing stakeholders are allowed to trade with those shares in the stock market with maximum ROI.

Even though an IPO provides an excellent opportunity to invest in a company during its initial growth stage, it also comes with new risks. Newly made public companies may lack financial management and their stock prices may change regularly. Other factors such as industry-specific patterns, economic cycles, seasonal trends of some companies, market sentiments, regulatory changes, and global events may also affect the IPOs which makes it hard for the investors to allocate their resources wisely. Market sentiments, influenced by public perception, plays a significant role in impacting the opening price of an IPO. The timing and valuation of an IPO are also crucial factors affecting the success of an IPO. Thus, stakeholders, including investors, and policymakers must be cautious while investing in any IPOs.

The stock market has an additional complexity that shows a non-linear pattern of the growth and degrowth of any IPO. This added complexity makes it more difficult for the investors to extract the pattern and then invest. Therefore, to effectively negotiate the landscape of an IPO, a careful blend of strategy, educated decision-making, and agility in the face of market volatility is needed. Machine Learning gives us a way to correctly identify the pattern of the stock market, thus aiding the stakeholders to make well-informed decisions. Machine Learning helps in analysing huge amounts of historical data related to IPOs such as financial metrics, market conditions, trends, other details and focuses on identifying the hidden patterns and correlations among them. It helps to identify the most important features that impact an IPO success thus improving the correctness of the predictive model.

This research focus on the fact that market fundamentals alone do not impact the IPO success, public sentiment plays a great role. It focuses on integrating sentiment analysis from various news, articles along with machine learning techniques to understand the factors that impact an IPO success. This approach will help the stakeholders to make informed choices in an unpredictable market and get the maximum ROI.

The contribution of this study includes:-

- Identifying the correlation between sentiment analysis and IPO investment factors.
- Analyzing the effect of the issue price and the GMP on the list price.
- Engineered a hybrid model that integrates sentiment analysis, utilizing neural networks, and various machine learning models for an accurate IPO recommendation system.
- Implemented four diverse machine learning techniques to provide strategic investment recommendations.

The study is split into several sections where first one being the Introduction, which highlights the importance of IPO listing and the complex nature of IPO and stock market. It also highlights how ML is effective in the prediction of an IPO’s listing price trend. The next section discusses the existing literature on IPOs, sentiment analysis and the use of machine learning in this field. After this section, the methodology section follows which consists of the overall proposed model architecture, and different units of the model explained in detail. This section also highlights the details about the dataset, the techniques used and other procedures of data retrieval, sentiment, and quantitative analysis. The next section demonstrates the outcomes of the suggested model and compares the models with each other, providing insights into their efficacy in predicting IPO success. Following the results, the conclusion provides the overview of the main discoveries, and their significance for various stakeholders and then the paper concludes by discussing the potential directions for the future investigation in this area.

2. Literature Review

A model to forecast the values of Google and Nike stock prices was put up by Moghar and Hamiche (2020), who also demonstrated their ability to predict opening prices using Long short-term memory (LSTM) and RNN. Nelson et al. (2017) used LSTM to predict the uptrend of the stock price in the near future with an average accuracy of 55.9%. Tyagi et al. (2020) used a hybrid model of CNN-LSTM to extract sentiment features for the classification into positive or negative polarities with an accuracy of 81.20%. Khare et al. (2017) made comparative research on LSTM and MLP for the short-term stock market price prediction and found that MLP performs better. Nayak et al. (2016) used support vector machine (SVM), logistic regression, and decision-boosted tree to predict the next day’s stock price based on the previous trends and volumes of the stocks. Prabhat and Khullar (2017) explore sentiment classification of Twitter reviews using algorithms of ML, specifically Naïve Bayes and Logistic Regression. Real-time Twitter data. The study evaluates algorithm performance in terms of accuracy, precision, and throughput. Naïve Bayes achieves 66.67% accuracy and 69.23% precision, while Logistic Regression on Hadoop with Mahout attains 76.67% accuracy and 73.57% precision. Logistic Regression outperforms Naïve Bayes with 10.1% higher accuracy and 4.34% higher precision. Velankar et al. (2018) carried out their research to forecast daily Bitcoin prices using various attributes such as the price of the previous day, trading volume, and transaction fees with exceptional precision. To achieve this, they employed two methods: Generalized linear model (GLM)/Random Forest and Bayesian regression. The experimental outcomes showed that the GLM/Random Forest model had a better performance than the Bayesian regression model, with an accuracy of 93% compared to 86%.

Aborisade and Anwar (2018) research classifies tweet authors using LR and Naïve Bayes machine learning methods. With 46,895 tweets, including known authors (politicians, celebrities) and unknown authors (regular users), logistic regression achieved a higher accuracy of 91.1% on the other hand Naive Bayes 89.8%. The study implemented in Python, addresses the challenges of limited tweet characters, emphasizing the flexibility and ease of the approach. Romadhon and Kurniawan (2021) analyze the factors affecting the rehabilitation rate of patients in Indonesia affected by COVID-19, focusing on variables such as age as well as gender. Naïve Bayes, KNN and logistic regression, methods were compared, with KNN showing the highest accuracy (0.750). The research suggests the potential for improvement by incorporating additional patient data variables like comorbidities and travel history. Hasanli and Rustamov (2019) This work presents a roadmap for sentiment analysis of Azerbaijani tweets, employing algorithms-Linear Regression, SVM, and Naïve Bayes-with bag-of-words models. Naive Bayes achieves the highest accuracy (94.00% and 90.00%) in classification using Bag-of-Words and TF-IDF features, respectively. SVM and Logistic Regression both attain an accuracy of 93.00% in Bag-of-Words experiments. Helmi et al. (2018) developed a chatbot system focusing on intent classification using Naïve Bayes and LR. Evaluation results show Logistic Regression outperforms Naive Bayes, achieving higher accuracy (0.727 vs. 0.636), precision, and recall. Both methods are suitable for chatbot systems, but Logistic Regression exhibits superior accuracy and precision. Al Amrani et al. (2018) used a hybrid approach of Random Forest and SVM, utilizing 1000 instances from Amazon, divided into review sentiment positive and negative. Cross Validation fold value of 10 is employed for training and testing. The hybrid method outperforms individual classifiers, achieving an accuracy of 83.4%. Chang et al. (2017) introduced a new approach for the detection of network intrusion, combining random forest and SVM. Highlighting the significance of host-based statistical features in predicting network intrusion. Evaluation of the KDD 99 dataset reveals that utilizing 14 selected features outperforms using the total 41 features, achieving higher attack detection rates. Dhyani et al. (2020) the study examined NIFTY daily data of the Nifty50 index, emphasizing the importance of understanding past behaviour through the ARIMA model's three variants: Basic, Trend-Based, and Wavelet-Based. This research contributes insights into the key components of time series data for enhanced predictive modelling in financial markets.

Ahmad et al. (2018) addressed the challenge of enhancing the performance of intrusion detection systems, by considering large datasets in system and network data analysis. Well-established machine learning algorithms includes SVM, Random Forest, and Extreme Learning Machine (ELM), are applied to the NSL-KDD dataset. The result reveals that on 80% training data and 20% testing samples, ELM has the highest accuracy and outperforms SVM (RBF), Random Forest, and ELM on 1/4 of data samples. Hassonah et al. (2019) study compares Decision Tree and KNN algorithms for churn prediction in Customer Relationship Management. Key findings include Decision Tree's higher accuracy (93%) compared to KNN (87%), a significant difference in F1 score (73% for Decision Tree vs. 33% for KNN), and a slightly better Area Under the Curve (AUC) for it (86% for Decision Tree vs. 82% for KNN). The Lift measure favours

Decision Tree (5.387) over KNN (4.277), indicating its superior responsiveness in predicting churn. Bathla et al. (2023) employed deep learning and machine learning to forecast stock market patterns, particularly focusing on candlestick patterns in technical analysis. It achieves 66% accuracy by considering four specific patterns within a one-day timeframe. Utilizing data from 800 NSE-listed stocks, the research predicts the next trend with 65.64% accuracy, notably identifying 86 out of 131 instances where patterns formed on 11-10-2022. Paramanik and Singhal (2020) applied text-based analysis to assess both negative and positive sentiments in the market. They utilized the autoregressive conditional heteroskedasticity model of conditional volatility, incorporating sentiment data from the preceding 14 years. The results indicated a higher trend in negative sentiments compared to positive sentiments, highlighting the substantial impact of noise in predicting stock market prices. Kumar et al. (2022) emphasized the ubiquity of decision-making and the need for refining conflicting criteria decision-making techniques. Despite the openness of multi-criteria decision-making (MCDM) techniques, their paper specifically focuses on WSM, WPM, and WASPAS models for real-life decision-making problems. Li et al. (2020) suggested a method that utilizes LSTM models to combine technical indicators with news feelings to anticipate stock market movements. Empirical evidence demonstrates that their approach surpasses MKL and SVM models in terms of both prediction accuracy and F-measure. The technique emphasizes the efficacy of using sentiment dictionaries that are specialized in the finance domain, namely the Loughran-McDonald Financial Dictionary. This dictionary leads to a 120% enhancement in predicting accuracy compared to other sentiment dictionaries. Han et al. (2023) presented an analysis of the importance of data labelling in the creation of trading systems in the stock market. The N-Period Min-Max (NPMM) labelling method is suggested as a solution to the limitations of traditional methodologies. NPMM utilizes minimum and maximum values to track stock price trends over a set period (N), to minimize the impact of minor price fluctuations. The paper empirically analyzes the Nasdaq stock market and concludes that NPMM labelling is an effective technique for predicting stock price trends. This method outperforms other labelling methods in terms of trading performance.

3. Proposed Framework

When thinking about making financial and resource investments, choosing the right IPOs is crucial along with maximizing ROI. It has been noted that the general public's perception and behaviour about a certain IPO significantly influence its prospective profitability. Sentiment analysis could not, however, be the only factor in determining a profitable investment. Public opinion, GMP, and the issue price are important basic variables that have a big impact on IPO performance. The utilization of machine learning methods is significant in forecasting market trends. When combined with sentiment research, these approaches provide investors with even more insight into which IPO to invest. Random Forest, Naïve Bayes, and KNN are the few techniques that are used to predict the trend on IPO listing.

3.1 Dataset Description

The IPO-related data scraped from various reliable sources such as Money Control, NSE, Yahoo Finance and The Economic Times, is self-collected and then organized. The dataset is collected and organized into two different categories: Structured and textual data and then pre-processed.

The structured data, collected from the sources mentioned, consists of the key factors affecting the IPO prices such as Listing Date, Offer Price, Lot Size, Amount raised (in Rs. Cr), and P/E ratio, GMP-High & GMP-Low, Closing Price, % change and no. of shares offered as shown in **Table 1**. These attributes like GMP, Lot size, offer price, P/E ratio plays a significant role in predicting the success of an IPO and offer valuable insights for stakeholders. GMP score is used to predict the demand of an IPO before it is officially listed in the market, whereas Lot size shows level of the demand for any share. The P/E ratio is used to assess the company’s valuation assisting the investors to evaluate whether any stock is over-valued, undervalued or correct in its price. The data collected from The Economic Times is used to make the textual dataset consisting of the company's name and news associated with its IPO, as shown in **Table 2**. Overall, these two datasets provide critical information about the company's attributes and public sentiments about the company, which will help conduct financial analysis, enhancing the predictive capabilities for IPO success.

Table 1. Dataset contains fundamental attributes of companies.

Company Name	Listing Date	Offer Price	Lot Size	AmountRaised (Rs Cr)	P/E Ratio	GMP (H)	Gmp(L)	Close Price	% Change	Shares Offered
Shanthala FMCG ProductsLimited	3-Nov-23	91	1200	16.07	25.66	NA	NA	103.55	13.79	1765934
Paragon Fine and Speciality ChemicalsLimited	3-Nov-23	100	1200	51.66	14.56	205	120	213.75	113.75	5165999
On Door Concepts Limited	1-Nov-23	208	600	31.18	4.56	238	218	203.4	-2.21	1499038
Blue Jet Healthcare Limited	1-Nov-23	346	43	840.27	37.49	441	365	395.85	14.41	24285260
Rajgor Castor Derivatives Limited	31-Oct-23	50	3000	47.81	NA	12	4	61.1	22.2	9562000
Woman Cart Limited	27-Oct-23	86	1600	9.56	50.67	30	10	122.85	42.85	1111627
IRM Energy Limited	26-Oct-23	505	29	545.4	24.13	105	18	472.95	-6.35	10800000
Arvind and Company Shipping Agencies Limited	25-Oct-23	45	3000	14.74	9.74	21	8	80.05	77.89	3275555
Committed Cargo CareLimited	18-Oct-23	77	1600	24.98	10.94	30	6	86.1	11.82	3244155
Plada Infotech Services Limited	13-Oct-23	48	3000	12.36	12.31	20	7	56.05	16.77	2575000
Vivaa Tradecom Limited	12-Oct-23	51	2000	7.99	47.66	0	0	42.54	-16.59	1566666

The second dataset is centred around textual data, specifically housing news related to IPOs, as displayed in **Table 2**. This collection encompasses a range of news articles and discussions, providing a qualitative aspect to the evaluation by preserving sentiments and market feedback encompassing the IPOs.

Table 2. Textual dataset on IPO-related news.

Company Name	News
Mamaearth	Mamaearth IPO: Should you subscribe? Here's what top 5 brokerages say on Honasa Consumer's book build issue | Mint
Mamaearth	MFs draw flak on social media for investment in Mamaearth IPO
Mamaearth	Mamaearth IPO: Company reveals the reason for reducing issue size | Mint
Cello	Cello World IPO gets thumbs up from brokerages: Should you buy into Rs 1,900-cr issue?
Cello	Cello World IPO: Firm mobilises â,¹567 crore from anchor investors ahead of issue | Mint
Cello	Consumer ware player Cello World sets IPO price band at Rs 617-648 per share
Arrowhead Seperation Engineering IPO	Arrowhead Seperation Engineering IPO opens November 16, price band set at 1233 apiece. Check GMP, other details | Mint
Arrowhead Seperation Engineering IPO	Arrowhead Seperation Engineering launches 1113 crore IPO
Arrowhead Seperation Engineering IPO	Arrowhead Seperation BSE SME IPO review (Avoid)

Remarkably, the news and basic features for 2021 and 2022 are used as the training sets for the model construction, whilst the data for 2023 is kept for testing. After sentiment analysis, the sentiments computed from the textual dataset are seamlessly merged with the fundamental attribute's dataset as shown in **Table 3**, along with Retail Individual Investor (RII) and Non-Institutional Investor (NII) data. By merging the mathematical elements of fundamental analysis with the qualitative elements gleaned from sentiment research, this integration seeks to produce more insightful results. The combined information serves as the foundation for later predictive modelling, offering a thorough assessment of the investment potential of IPOs.

Table 3. Merged dataset of fundamental attributes and sentiment results.

Company Name	Listing Date	Offer Price	Lot Size	Amount Raised (Rs Cr)	P/E Ratio	NII	RII	GMP (H)	GMP (L)	TOTAL	Sentiments	% Change
Shanthala FMCG Products Limited	3-Nov-23	91	1200	16.07	25.66	4.76	3.05	NA	NA	3.91	1	13.79
Paragon Fine and Speciality Chemicals Limited	3-Nov-23	100	1200	51.66	14.56	419.46	185.28	205	120	205.74	1	113.75
On Door Concepts Limited	1-Nov-23	208	600	31.18	4.56	3.16	7.87	238	218	5.59	1	-2.21
Blue Jet Healthcare Limited	1-Nov-23	346	43	840.27	37.49	13.59	2.24	441	365	7.95	1	14.41
Rajgor Castor Derivatives Limited	31-Oct-23	50	3000	47.81	NA	260.01	80.7	12	4	107.43	1	22.2
Woman Cart Limited	27-Oct-23	86	1600	9.56	50.67	56.3	71.94	30	10	67.48	1	42.85
IRM Energy Limited	26-Oct-23	505	29	545.4	24.13	48.34	9.29	105	18	27.05	-1	-6.35
Arvind and Company Shipping Agencies Limited	25-Oct-23	45	3000	14.74	9.74	436.05	321.97	21	8	385.03	1	77.89
Committed Cargo Care Limited	18-Oct-23	77	1600	24.98	10.94	94.2	78.73	30	6	87.78	-1	11.82

3.2. Methodology

The foundation of the suggested methodology is the knowledge that, when used separately, sentiment analysis and fundamental research would not produce the degree of precision needed to evaluate the profitability of investments in certain IPOs. It has been demonstrated, however, that the complementary use of emotion and fundamental analysis yields more insightful results and increases the precision of IPO investment potential predictions. The model is designed to be a hybrid system that combines technical and fundamental research methods to fully utilise available data and provide a thorough assessment of investment opportunities in the IPO market. The proposed model is divided into 3 different units as shown in **Figure 1**.

3.2.1 Information Retrieval and Aggregation Unit

The most fundamental component of the model is this unit, which consists of gathering and arranging data from various reliable sources. The web scraping tools were used to scrape the data from online sources such as Money Control, NSE, Yahoo Finance and The Economic Times, which consists of qualitative and quantitative data as well. After data acquisition, the pre-processing techniques were applied to clean and organize the data into two different datasets, which is essential for better data analysis. The separation of the whole scraped data into two different sets is fundamentally the most important step in this methodology. The data was divided into two separate sets – one consisting of all the quantitative attributes such as issue price, GMP, P/E ratio, lot sizes etc and the other one consisting of the qualitative textual data extracted from news, articles, videos, and comments from the online sources. Quantitative attributes give insights about the financial analysis and growth prospects of the company, thus helping the stakeholders to know where to invest and where not to. Qualitative data shows how public sentiments can affect the success of an IPO. By organizing the data into these distinct categories, the Information Retrieval and Aggregation Unit lays the foundation for a thorough evaluation of IPO investment potential.

3.2.2 Sentiment Analysis Unit

The sentiment analysis is systematically carried out by the usage of neural networks, in which the first step is to use the textual dataset of news and article headlines to train the neural network architecture. The percentage change between the issue price and closure price on the day of the IPO listing is used to determine the mood of the training dataset. Additionally, results from several pre-trained models such as TextBlob and NLTK, are used to deduce initial feelings. However, it is imperative to acknowledge that these pre-trained models cannot be directly applied in this context due to their general-purpose nature. Specifically, they are not tailored for sentiment analysis about IPO-related news. TextBlob and NLTK are two examples of pre-trained models that are used because of their capability to provide a variety of sentiment evaluations. Our specialised sentiment analysis model for news primarily relevant to IPOs is trained using the combined findings from these models.

To conduct sentiment analysis, all stock-specific news headlines are vectorized, transforming the text into a sequence of integers. The neural networks that were used for training-which included layers like the Embedding layer, 1D Convolution layer, Global Max Pooling, Dropout layer, and Dense layer-are then trained using these vectorized headlines. Each layer was configured with specific parameters to enhance the ability and accuracy of the model to extract intricate and complex features and learn better from the input data. The sequential text classification model is defined by three parameters-Maximum Feature, sequential length, and the embedding value. The maximum feature is set to be 20,000 meaning that the model is trained on the first 20,000 words from the dataset, the sequential length is set to be 40 meaning that each input has this many dimensions, and the embedding value is assigned to 64 indicating that the accepted index is represented by 64-dimensional vector as the output of the first embedding layer. The model can effectively scan and extract relevant information from the words in the input data, thus aiding in better training all due to this vectorized representation.

a) Embedding Layer

The Layer processes the inputs by the sequence length, which is determined by the formula "maximum feature + 1." Because it creates a 64-bit embedding, one way to interpret this is that every approved index corresponds to an output 64-dimensional vector. Through the utilization of its vectorized representation, the model can effectively perform word analysis and extract significant information. Because of this superior information extraction, training is enhanced in terms of overall enjoyment, and it also assists the model in comprehending concepts more effectively.

b) Conv1D Layer

The Conv1D layer, which has '128' different convolution filters, comes after the Embedding layer. It works with 40-length sequences, processing three components at once. Rectified Linear unit (ReLu) is applied to this layer as the activation function, which introduces non-linearity in the model so that the model can learn more intricate and complex relationships in the data. It also helps to reduce overfitting, which improves the model's overall effectiveness.

c) Dense and Dropout Layer

The model's last layer improves overall training by making use of the information learned from the layer before it. It uses the available pool to process three words at once. To avoid overfitting during the activation phase and provide a more reliable and broadly applicable model, the activation function "Sigmoid" is utilised in conjunction with "regulizerL2". The model was trained across several epochs until the current epoch's loss function was equal to the one from the previous epoch. The model that is trained is then used to execute sentiment analysis.

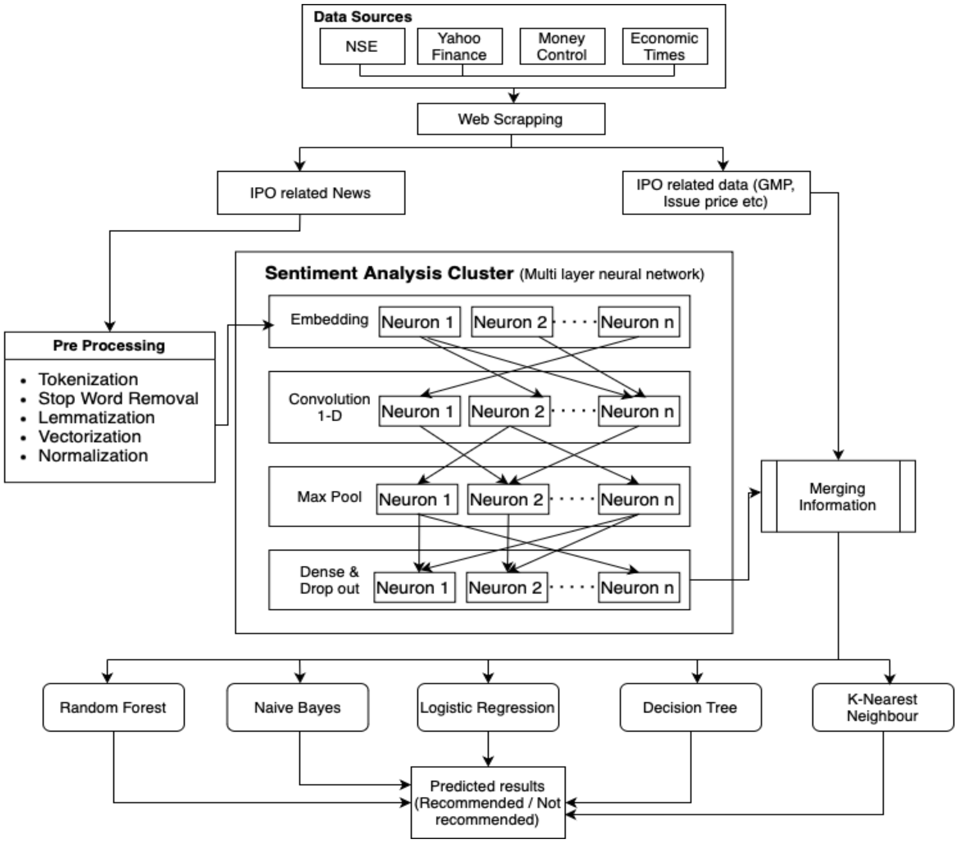

Figure 1. Fundamental and sentiment driven IPOs prediction model.

3.2.3 Quantitative Analytics Segment

In this section, a thorough and hybrid approach is implemented which integrates sentiment analysis along with a variety of machine learning algorithms like Decision Tree, Random Forest, Naïve Bayes, KNN, and logistic regression to predict the success of an IPO. These modules work independently of one another to interpret the combined dataset, which facilitates the data's multidimensional analysis. It is critical to acknowledge the disadvantages associated with each of these machine learning approaches in addition to their advantages. Decision Tree is prone to overfitting, which means they perform very well on the trained dataset but fail to perform well on the unknown data. Then, the Random Forest may be less interpretable even though it is resistant to overfitting due to its ensemble nature. Despite the logistic regression is easy to understand and implement, it might not be able to identify the non-linear correlations between the variables which can thus restrict its power to predict the IPO success correctly. Since Naïve Bayes model has a feature independence assumption that may not be realistic in real-world scenarios, it may have a side effect on the sentiment analysis's accuracy, particularly when the features are dependent on each other. Lastly, KNN does not always guarantee effective results as it is largely dependent on the number of neighbours and the distance metric selected for further calculations. These algorithm-specific drawbacks must be solved to improve the model's performance thus improving this study's reliability.

The goal of this study is to focus on the intricate relationship between people's sentiments and the market fundamentals by combining both and thus providing a detailed understanding of the factors that affect the market strategies and IPO success rate. This process model uses a k-fold cross-validation technique for model evaluation and success. This technique splits the whole dataset into k subsets where (k-1) subsets are used for training purpose and the one left is used for validation purposes. This process is repeated k times to generalize the model so that it learns from unseen data as well. It also ensures that the model is resistant to overfitting, and it improves the overall dependability of the dataset.

4. Results

This study focuses on the complexities of IPOs considering the dynamic market nature, continuously evolving stock market, and ever-changing phase of investments. The paper focuses on the fact that integrating sentiment analysis along with other market indicators can help the model in better prediction in terms of accuracy score. The paper employs a diverse array of machine learning algorithms like Random Forest, Naïve Bayes, Decision Tree, and KNN to examine various factors such as sentiment scores, GMP and the issue prices associated with the IPOs. These factors are used in the evaluation of the classification model thus aiding in the assessment of the model predictions.

The study encompasses the self-made dataset spanning from the year 2021 to 2023 to check the models' effectiveness. **Table 4** and **Figure 2** of the study demonstrate how the Decision Tree surpasses all other algorithms by reaching an astounding accuracy rate of 82.3%. The different machine learning models are also compared based on the various benchmarks such as recall, F-measure, precision, and AUC curve. The Decision Tree algorithm surpasses all other algorithms in most benchmarks due to its resilience to outliers and missing values, as well as its superior ability to handle non-linear relationships.

Table 4. Comparative analysis of machine learning models in IPO trend listing.

S. No.	Model	Accuracy	Precision	Recall	F1 Score	AUC
1.	Decision Tree	0.823	0.923	0.857	0.888	0.761
2.	Random Forest	0.764	0.857	0.857	0.857	0.654
3.	Naive Bayes	0.647	0.689	0.571	0.727	0.857
4.	KNN	0.823	0.866	0.928	0.896	0.630
5.	Logistic Regression	0.588	0.769	0.714	0.740	0.428

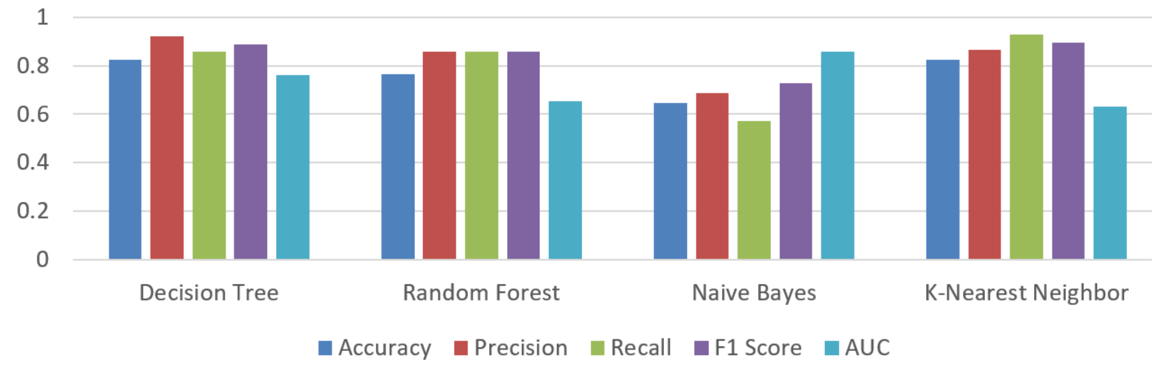

Figure 2. Comparison of performance metrics across machine learning algorithms.

This study not only marks the efficacy of the methods used but also integrates the two main key factors affecting an IPO's success. Thus, the study acts as a robust tool for all the stakeholders to make informed decisions, thus achieving maximum ROI.

5. Conclusion

This paper suggests a hybrid model to forecast the performance of an IPO while addressing several intricate issues, such as dynamic markets, non-linear data, and other associated economic hazards. Sentiment analysis and conventional market indicators are successfully integrated into the model to harness qualitative and quantitative aspects, resulting in a more comprehensive understanding of the market conditions and thus, creating a computationally more robust and efficient model. Insights concerning investor sentiment and public opinion that are typically overlooked by other traditional models are enhanced by the incorporation of sentiment analysis. As a result, our hybrid model produces predictions with greater accuracy, assisting stakeholders and investors in making informed decisions and managing risks.

The study utilizes five machine learning algorithms to forecast the quantitative score of IPO success. The Decision Tree demonstrates the effectiveness of the model by outperforming all other algorithms with an accuracy rate of 82%. The results of this study can be used by corporations, investors, and policymakers to get useful insights and make more informed decisions by considering all available market information.

6. Future Work and Limitations

In the future, there will be various opportunities in the area of IPO Listing and the stock market. We can include multiple data sources such as images, videos, graphs, audio from news reports, and social media channels apart from textual data for a better-trained machine learning model. Enhancement of sentiment analysis can be accomplished by the utilization of transformer-based models such as BERT or GPT. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) values can be utilized to provide an explanation for the predictability of the model and to gain insights into the elements that are driving IPO listing outcomes. A hybrid real-time model can be implemented dynamically to evaluate the fluctuating market conditions, advancing the model's capability to perform better in IPO forecasting.

However, apart from predicting the outcomes, it is significant to address some of the drawbacks associated with the study of IPO listings. The data quality which has noise and outliers, complex market dynamics, non-linear relationships between the data, algorithm-specific problems, and economic conditions can greatly impact the stock market trends which makes it difficult from a study point of view. The legal and ethical concerns of predictive models can lead to unintended consequences. Despite these challenges, our study aims to fortify the predictive capabilities of the proposed hybrid model, advancing its applicability in IPO forecasting and contributing to the ongoing evolution of computational finance.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that helped improve the quality of this work.

References

Aborisade, O., & Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration* (pp. 269-276). IEEE. Salt Lake City, UT, USA. https://doi.org/10.1109/iri.2018.00049.

Ahmad, I., Basheri, M., Iqbal, M.J., & Rahim, A. (2018). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. *IEEE Access, 6*, 33789-33795. https://doi.org/10.1109/access.2018.2841987.

Al Amrani, Y., Lazaar, M., & El Kadiri, K.E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science, 127*, 511-520. https://doi.org/10.1016/j.procs.2018.01.150.

Bathla, D., Garg, A., & Sarika. (2023). Stock trend prediction using candlestick pattern. In: Jain, R., Travieso, C.M., Kumar, S. (eds) *Cybersecurity and Evolutionary Data Engineering*. Springer, Singapore. pp. 235-246. https://doi.org/10.1007/978-981-99-5080-5_21.

Chang, Y., Li, W., & Yang, Z. (2017). Network intrusion detection based on random forest and support vector machine. In *2017 IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing* (pp. 635-638). IEEE. Guangzhou, China. https://doi.org/10.1109/cse-euc.2017.118.

Dhyani, B., Kumar, M., Verma, P., & Jain, A. (2020). Stock market forecasting technique using ARIMA model. *International Journal of Recent Technology and Engineering, 8*(6), 2694-2697. https://doi.org/10.35940/ijrte.f8405.038620.

Han, Y., Kim, J., & Enke, D. (2023). A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost. *Expert Systems with Applications, 211*, 118581. https://doi.org/10.1016/j.eswa.2022.118581.

Hasanli, H., & Rustamov, S. (2019). Sentiment analysis of Azerbaijani twits using logistic regression, naive Bayes and SVM. In *2019 IEEE 13th International Conference on Application of Information and Communication Technologies* (pp. 1-7). IEEE. Baku, Azerbaijan. https://doi.org/10.1109/aict47866.2019.8981793.

Hassonah, M.A., Rodan, A., Al-Tamimi, A.K., & Alsakran, J. (2019). Churn prediction: A comparative study using KNN and decision trees. In *2019 Sixth HCT Information Technology Trends* (pp. 182-186). IEEE. United Arab Emirates. https://doi.org/10.1109/itt48889.2019.9075077.

Helmi Setyawan, M.Y., Awangga, R.M., & Efendi, S.R. (2018). Comparison of multinomial naive Bayes algorithm and logistic regression for intent classification in chatbot. In *2018 International Conference on Applied Engineering* (pp. 1-5). IEEE. Batam, Indonesia. https://doi.org/10.1109/incae.2018.8579372.

Khare, K., Darekar, O., Gupta, P., & Attar, V.Z. (2017). Short term stock price prediction using deep learning. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology* (pp. 482-486). IEEE. Bangalore, India. https://doi.org/10.1109/rteict.2017.8256643.

Kumar, A., Garg, P., Pant, S., Ram, M., & Kumar, A. (2022). Multi-criteria decision-making techniques for complex decision making problems. *Mathematics in Engineering, Science & Aerospace, 13*(2), 791-803.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management, 57*(5), 102212. https://doi.org/10.1016/j.ipm.2020.102212.

Moghar, A., & Hamiche, M. (2020). Stock market prediction using LSTM recurrent neural network. *Procedia Computer Science, 170*, 1168-1173. https://doi.org/10.1016/j.procs.2020.03.049.

Nayak, A., Pai, M.M.M., & Pai, R.M. (2016). Prediction models for Indian stock market. *Procedia Computer Science, 89*, 441-449. https://doi.org/10.1016/j.procs.2016.06.096.

Nelson, D.M.Q., Pereira, A.C.M., & de Oliveira, R.A. (2017). Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks* (pp. 1419-1426). IEEE. Anchorage, AK, USA. https://doi.org/10.1109/ijcnn.2017.7966019.

Paramanik, R.N., & Singhal, V. (2020). Sentiment analysis of Indian stock market volatility. *Procedia Computer Science*, *176*, 330-338. https://doi.org/10.1016/j.procs.2020.08.035.

Prabhat, A., & Khullar, V. (2017). Sentiment classification on big data using Naïve Bayes and logistic regression. In *2017 International Conference on Computer Communication and Informatics* (pp. 1-5). IEEE. Coimbatore, India. https://doi.org/10.1109/iccci.2017.8117734.

Romadhon, M.R., & Kurniawan, F. (2021). A comparison of naive Bayes methods, logistic regression and KNN for predicting healing of Covid-19 patients in Indonesia. In *2021 3rd East Indonesia Conference on Computer and Information Technology* (pp. 41-44). IEEE. Surabaya, Indonesia. https://doi.org/10.1109/eiconcit50028.2021.9431845.

Tyagi, V., Kumar, A., & Das, S. (2020). Sentiment analysis on Twitter data using deep learning approach. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking* (pp. 187-190). IEEE. Greater Noida, India. https://doi.org/10.1109/icacccn51052.2020.9362853.

Velankar, S., Valecha, S., & Maji, S. (2018). Bitcoin price prediction using machine learning. In *2018 20th International Conference on Advanced Communication Technology* (pp. 144-147). IEEE. Chuncheon, South Korea. https://doi.org/10.23919/icact.2018.8323676.

Publisher's Note- Ram Arti Publishers remains neutral regarding jurisdictional claims in published maps and institutional affiliations.