Technology in Society 63 (2020) 101409

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Issues and opinions

Twitter data analytical methodology development for prediction of start-up firms' social media marketing level

Sang Hoon Jung *, Yong Jin Jeong

Kwangwoon University, Nowon-Gu, Seoul, P.C. 01897, South Korea

ARTICLE INFO

Keywords:
Social media
Marketing
Data analysis
Machine learning
Start-ups

ABSTRACT

Social media marketing is an essential and important tool for start-up firms, which can help start-up firms remedy the marketing limitations through ease and relatively low costs. Predicting start-up firms' social media engagement level can allow them to gauge the effectiveness of their social media marketing efforts and can provide numerous benefits related to strategic marketing processes. This study focuses on developing a methodology involving data science processes and machine learning models to account for the ongoing advancement of business intelligence methodologies. This study gathered data of 8,434 start-up firms from Twitter, generated social media-based features, and created machine learning models to predict the social media engagement level of each firm. The results show that deep learning provides the best accuracy in predicting the engagement levels. The results also show that the number of tweets by the firms, the number of retweets received, and the number of likes received have the most significance in determining the effectiveness of social media marketing activities.

1. Introduction

Social media marketing has become one of the effective tools for firms to utilise. With the emergence of social media, people from all over the world can voice their opinions and spread awareness on particular issues. Social media has become an inevitable platform of communication, and firms have to adapt its concepts to their business philosophy to prevent possible decay [1]. Furthermore, firms can now communicate directly with their customers and stakeholders, and promote their brand, products, and services [2]. For all firms, social media can bring numerous benefits, such as the ability to reach out to huger audiences, the ability to drive sales through social commerce, and the ability to establish trust and reputation [3]. The benefits of social media marketing are greater for start-up firms. Social media marketing incurs relatively low costs compared to other traditional forms of marketing, and the Internet has opened up a basic tool of business communication through its elements of broadcasting and openness [1,2]. According to Corvera and Johansson [4], the main advantages of social media platform usage among firms are reduced cost and closer interactions with their customer. Start-up firms utilising social media marketing is provided with "effective online marketing tools that enable them to present their products and services in the same way that many large corporations do" [2], allowing start-up firms to perform marketing activities of similar effectiveness as those of large corporations. Social media marketing can also solve the more common problems of marketing activities of start-up firms, such as limited budget, lack of expertise, and positioning against larger competitors [5]. Bearing these advantages in mind, it is quite obvious that start-ups must utilise social media marketing methods to improve their business model.

In order to fully reap the benefits of any business processes, it is important for firms to predict and make forecasts to plan and make necessary adjustments for the future of the business. Even in marketing, sales forecasting is regarded as one of the most important factors to a firm's success, as it allows a firm to gauge their financial performance in the future, allowing firms to make necessary adjustments and to improve their implemented strategies [6,7]. Market size and potential growth of the market are also often predicted to allow the firms to be ready for potential market trends shifts [8]. Such forecasts lead to firms being able to manage costs, being prepared for possible future events, driving future events, and forecasting profits and other financial outcomes. As such, to evaluate their current social media marketing efforts and to determine the future impacts of their actions, it is necessary and beneficial for start-up firms to predict the effects of their social media marketing activities.

Among the tools used for prediction, machine learning has surfaced as one of the most accurate method. Machine learning in business applications has introduced data mining approaches to business analytics in order to identify useful patterns and meanings, allowing predictions according to the patterns found [9]. With recent advancements

* Corresponding author.
E-mail addresses: brianjung21@kw.ac.kr (S.H. Jung), yjjeong@kw.ac.kr (Y.J. Jeong).

https://doi.org/10.1016/j.techsoc.2020.101409
Received 28 November 2019; Received in revised form 8 September 2020; Accepted 12 September 2020
Available online 23 September 2020

in machine learning and continuous rise in gathered and generated data, machine learning methods are being implemented in business processes [10]. Recent advancements in deep learning and big data have allowed larger scales of business process innovation. In sales forecasting, deep learning has outperformed traditional machine learning methods by 89.42%, showing tremendous improvements [11,12]. As such, utilising machine learning methods in predicting various business processes is necessary for firms, and more so for start-ups that can attain competitive advantages through timely analysis and accurate predictions.

As such, given the emergence and benefits brought forth by both social media marketing and machine learning methods, this study seeks to develop a data analytical methodology for start-up firms to predict their social media engagement level as a proxy for their social media marketing activity effectiveness. Through such prediction, start-up firms can make necessary amends and strategical implementations to improve their chances of success. This study utilised a data scientific approach of gathering and cleaning data and several machine learning models for predictive analysis. Data has been gathered from Twitter, one of the most active social media platforms, and from 8434 start-up companies in the cities of USA listed in startup-list.com, a website listing start-up firms in each city. The results show that deep learning is the best performer of such predictive analysis, with an accuracy score of 73.42%, showing the much improved ability of predicting a start-up firm's social media marketing effectiveness using various features describing a firm's efforts. This further shows the relationships between the social media activity features and the effectivity of the activities, implying the importance of such features in maximising the effects of their social media marketing activities. Furthermore, interpretation of permutation importances in deep learning revealed that the number of tweets by the firms, the number of retweets received, and the number of likes received are the most significant features in determining the effectiveness of social media marketing activities. This shows that start-up firms should focus on their number of tweets sent, how much retweets they are getting, and how much likes they are receiving.

The rest of the study is as follows. First, the related literature in regard to social media marketing, the metrics used to determine its effectiveness, and its benefits to start-up firms are explored. A brief explanation of machine learning and its application in business processes are also explained elaborated, then a conceptual framework combines such concepts together to form the basis of this study. Second, the methodology of this study is elaborated, which will be the main focus of this study. Third, the results of the experiments are explained and analysed. Fourth, the conclusions and discussions about the implications of such methodology construction and the results are elaborated.

2. Literature review

2.1. Social media marketing and start-up firms

According to a statistical study by Statusbrew in 2018, there are approximately 3.04 billion active social media users worldwide, and 90% of social media users try to reach out to brands or retailers. This illustrates the importance of using social media platforms by firms, especially start-up firms, to improve their business strategies and models. There are several ways that social media marketing can benefit start-up firms: (1) social media allows firms to communicate directly with consumers and stakeholders, and (2) social media can improve venture financing of start-up firms, and many other benefits that can improve a start-up firm's competitive advantage.

Utilisation of social media can allow firms to enhance communication with consumers and stakeholders. According to Dziadkiewicz, et al. [1], social media users that are involved in business is more prone to "support, give advice, and share their own experience with the newcomers", as they want to be included in the business and are more likely to take on the responsibility of the outcomes of firms' developments without a real will to be rewarded in return. Through this, firms can gain advertisers who will spread the business ideas, reflect the firms' personalities, and lure more people to be included in the brand community within the social media context. Such will, in turn, create a brand community that can contribute to a firm's brand loyalty level among its consumers, level of interaction with the consumers, the level of feedback a firm can gather from consumers, and greater effects of word-of-mouth advertising [3]. Such word-of-mouth advertising and direct interactions with consumers can result in customer relationship building and means for creating viral marketing among a certain group of people that are interested in a common topic [1]. Even if social media users are not part of the brand community, anyone can share their knowledge and experiences about the firm and their products or services. As such, start-up firms can now have the effective tools for presenting themselves the same way as numerous large corporations [2]. Social media platforms allow firms to listen to what is being said about themselves, and they can interact with the consumers directly [13], resulting in better perceived service quality. Firms can furthermore present a positive first impression in social media platforms, highlighting that they are trustworthy, knowledgeable, and approachable through their posts and other social media activities [14]. Furthermore, social media platforms allow global interactions, therefore, not being limited geographically [15]. As such, start-up firms must endeavour to engage in social media marketing activities to gain such advantages and to further improve their business strategies.

Social media marketing activities can also improve a start-up firm's venture financing and engage in successful crowdfunding. Social media platform is a powerful platform for advertising and building brand awareness among users within a strict advertising budget [16]. According to Aggarwal, et al. [17], start-up firms' social media platform usage is linked with increased venture financing. Greenwood and Gopal [18] found that the emergence of social media platforms has led an increase in firm funding. A study by Kuruzovich and Lu [16] found that a start-up firm's social media marketing activities such as the number of posts, the number of replies, etc., can improve the engagement of consumers and venture financing of the start-up firms. This is mainly due to the engagement of customers generated through word-of-mouth marketing. Engagement from word-of-mouth advertising and other social media marketing activities can be regarded as a type of unobtainable financial and fundraising data, and therefore, can help start-up firms evaluate different ventures [19]. Furthermore, weak tie theory states that social media engagement of consumers assists start-up firms to increase the number of weak ties in their social networks and access resources (social capital) necessary for their success [20,21]. Start-up firms' evaluation of venture financing is mainly contributed by engaging in different social capital [22] as stated by social capital theory [23]. Within the context of social media, where network is binding users and firms together, the level of engagement, or interaction, between consumers and firms is an important contributor to a firm's competitive advantage [24]. On the other hand, a study by Zhang et al. [25] found out that active user engagement on social media platforms with firms is highly correlated to crowdfunding success. Crowdfunding is when start-up firms utilise third-party platforms to establish a fundraising campaign. Using start-up firms that are fundraising on AngelList, a third-party fundraising platform, the study analysed the crowdfunding level with the firms' social media activities. From the analysis, it was found that the effects of social media marketing do play a role in the abilities of firms to crowdfund. With such studies proving the success and benefits of social media marketing among start-up firms, the importance of engaging in proper social media marketing efforts is highlighted.

Other than the benefits of social media marketing mentioned above, there are several other benefits that come from engaging in social media marketing activities. One such benefit is the relatively low cost that social media marketing incurs [1–3,26]. With reduced cost, start-up firms can engage in marketing activities that are of equivalent impact to the marketing efforts of large corporations. Reducing cost is also essential for start-up firms that often face financial constraints due to their sizes. Start-up firms can also utilise social media marketing to generate possible ideas about creating the company's mission and vision, promoting a strategy, increase brand exposure, gain market insights, drive sales through social commerce, and establish trust and reputation [1–3]. Today, the most reliable source of advertising is a recommendation from another person that are similar to a user. Utilising such segmentation can show where a firm's audience is, and the possible subjects and issues that they are often interested in [27]. Social media marketing can also allow firms to overcome limited budget and lack of expertise problems [5]. As such, as elaborated above, such benefits and more can be gained through the use of social media platform to take part in social media marketing activities.

Therefore, start-up firms are highly encouraged to utilise social media marketing for reasons explained above. However, there is still room for developments in analytical and evaluation of social media marketing efforts of firms. Most importantly, firms have to determine how to measure and evaluate the effectivity of their social media marketing efforts, and the factors that influence their effectiveness. In doing so, machine learning methods can provide the means for developing abovementioned points.

2.2. Machine learning in business processes

Machine learning, when utilised in business sense, is the "study of computational methods to automate the process of knowledge acquisition from examples" [9]. More elaborately, computers are not programmed step-by-step, but instead, tries to learn from a given data without a programmer or any outside influence assisting the machine [28]. Use of machine learning in businesses have enriched the analysis of business data to make way for data mining, which utilises machine learning, statistical analysis, and visualisation techniques, with domain knowledge of an analyst to find out useful patterns and meanings [9]. Such phenomenon arose as business analytics innovation is both wanted and necessary for a firm's performance [29–31]. Major categories of machine learning application are classification, prediction, association, and detection. Classification tasks can refer to categorising stock risk-return characteristics, and prediction tasks can refer to fault forecasting in telecommunication networks as some examples. Several fields in business can utilise data mining applications, such as finance (forecasting bankruptcies and defaulting loans), telecommunication (call tracking and fraud detection), marketing (market segmentation and new opportunity analysis), and web analysis (similarity assessment of user browsing patterns and searching for specific web pages) [9]. Such advancements and growing necessity of machine learning implementation in business processes are what necessitates needs for more relevant studies.

Due to the recent advancements in machine learning, coupled with continuous rise in data gathered, and the implications of data gathering and analysis in various fields, businesses have also sought to implement machine learning processes to analyse the data that they have accumulated. What was once considered to be non-applicable to businesses are now being implemented as the forefront roles in business [32]. Traditionally, business analysts utilised various statistical techniques to solve various business problems, but machine learning techniques have changed the way data can be analysed to discover useful patterns [9]. Several examples as highlighted by Stratlytics [32] is utilisation of face recognition, risk management, and fraud detection in finance and insurance services, and optical character recognition. Moreover, businesses can apply machine learning in banking and financial sectors to aid in loan approval and customer segmentation, manufacturing and supply chain to aid in demand forecasting and telematics, healthcare, and medicines research to aid in diagnosis and personalised medicine prescription and patient monitoring and predictive alerts, and travel to aid in aircraft scheduling and dynamic pricing. Also, as mentioned in Stratlytics [32], implementing machine learning processes to a firm's operation is dependent on the technical feasibility that the firm has and the managerial enthusiasm regarding the use of machine learning. Many businesses are now leaning more towards to implementing machine learning than before, and it has been utilised in various different fields of business.

Importantly, recent advancements in deep learning and big data have brought about a new scale of improvement and innovations in business analytics. Deep learning refers to deep neural networks, with emphasis on deeper dimension of networks, which can contain up to hundreds of layers, millions of neurons, and very complicated structures and connections between neurons [33,34]. Through this, utilisation of deep learning can introduce flexibility in modelling highly complex and non-linear relationships between variables, which can far outperform traditional machine learning methods and other business analytical methods [35]. Aside from deep learning, big data concepts have emerged in firms especially where traditional machine learning methods are not able to analyse well, but deep learning optimisation routines can scale effectively for analysing big data [36,37]. An example of deep learning and big data utilisation in business is focusing on operations management, where a study by Kraus, et al. [35] tried to predict the number of incoming service requests for an IT department to better adapt available capacities to the load. The results showed that the deep learning model has outperformed all other traditional machine learning methods. Another example is in the field of sales forecasting, which can aid in decision-making and strategic planning of firms [11,12]. Deep learning has outperformed the traditional machine learning methods by up to 89.42% and has shown the effectiveness of using deep learning for such analyses. Therefore, more and more firms are trying to utilise machine learning and deep learning efficiently and are trying to implement them to business operations. Start-up firms should also endeavour to utilise such methods to compete with other firms.

As such, incorporating machine learning methods in analysing and evaluating the social media marketing efforts of start-up firms can bring about new developments in the field and provide start-up firms with the competitive advantage that they need while allowing them to make use of the several benefits brought by the use of social media platforms.

2.3. Conceptual framework

As such, through this study, we seek to determine if machine learning methods can aid in social media marketing development and evaluation through their predictive capabilities. More specifically, we seek to determine if the effects of a firm's social media marketing activities can be predicted using data and features representing the social media platform's activities and machine learning methods.

As highlighted by Sadeque and Bethard [38], several social media features have been identified and used for predicting engagement levels in online social networks. Different studies in social media marketing effects prediction have used different types of features, generally grouped into demographic, linguistic, activity, and interpersonal relationship. Demographic features refer to a user's demographic information including age, sex, location, etc. [39–41]. Linguistic features refer to the content properties of a user's social media platform activities [42–44]. Activity features refer to the activity properties of a user. Interpersonal relationship features refer to the links between users [41,44–46]. For this proposed methodology, linguistic and activity features are extracted and utilised as they are the only available ones easily accessed through the social media platforms without incurring additional costs.

Firms need a metric to gauge the effectiveness of their social media marketing activities, and such metric can be the engagement level of users with the firms' social media activities. According to Perreault and Mosconi [47], "social media engagement can be expressed by

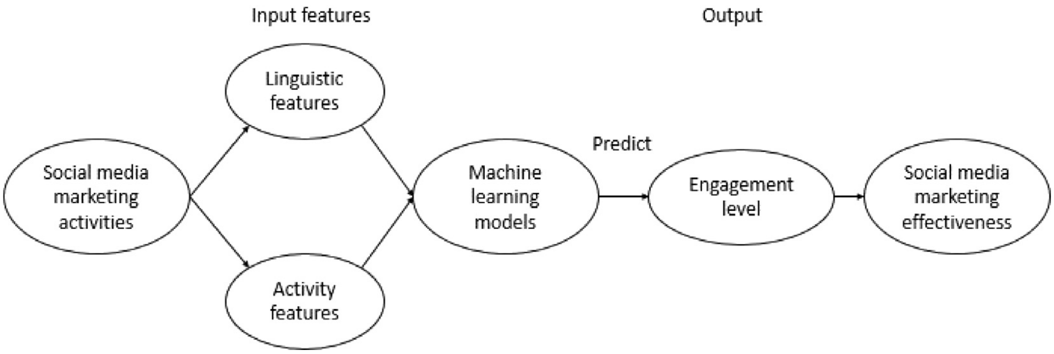


Fig. 1. Conceptual model of this study.

Table 1
Summary of literature review.

Literature area	Authors	Description
Social media marketing benefits to start-up firms	Dziadkiewicz, et al.	• Social media users involved in business more welcoming to newcomers
	Aggarwal, et al.	• Start-up firms' social media platform usage increases venture financing
	Greenwood and Gopal	• The emergence of social media platforms has led an increase in firm funding
	Kuruzovich and Lu	• Social media marketing of start-up firms improves engagement of consumers and venture financing
	Zhang, et al.	• Active user engagement on social media platforms with firms is highly correlated to crowdfunding success
Machine learning in business processes	Stratlytics	• Machine learning incorporated into business processes:
		• Utilisation of face recognition
		• Risk management and fraud detection in finance
		• Optical character recognition
	Kraus, et al.	• Predicted number of incoming service requests for IT department using machine learning
		• Deep learning performed the best
	Boone, et al. Lau, et al.	• Deep learning outperformed other methods by 89.42% in sales forecasting
Social media features	Sadeque and Bethard	• Social media features grouped into demographic, linguistic, activity, and interpersonal relationship
Social media marketing effectivity metric		• Social media engagement expresses:
	Perreault and Mosconi	• Interactions and participation
		• Conversation and eWOM (e-Word-Of-Mouth)
	Rahman, et al.	• Concluded that social media engagement level and purchase intention has positive correlation
	Blaschke and Veh	• Social media engagement is correlated with a brand's financial performance
	Misirlis and Vlachopoulou	• Engagement is one of the most commonly used metrics for social media activities evaluation

interactions [48,49], participation [50,51], conversation [10], eWOM (e-Word-Of-Mouth) [52,53], and other offline and online actions". Customer engagement, in terms of marketing, can mean a psychological process that can eventually lead to loyalty formation, manifestation of behaviour, and psychological state of vigour, dedication, absorption, and interaction of the customers [54]. As such, customers in social media platforms with low level of engagement are those that only consume content, while those with high level of engagement are customers that generate content in regard to the firms [50,55]. Some measures of social media engagement of customers can be done through number of visits, purchasing behaviour, and intended behaviours (sharing, commenting, and liking the firm's posts) [56]. Objectives of social media marketing can fall under two categories: short-term and long-term. Short-term objective is to generate revenue, while long-term objective is to create brand equity and build brand relationships [57]. Engagement can measure the short-term objective as a study by Rahman, et al. [58] concluded that there is a positive correlation between social media engagement level and purchase intention of consumers. Another research by Blaschke and Veh [59] shows that social media engagement is correlated with a brand's financial performance. Engagement is also one of the most commonly used metrics for social media activities evaluation as denoted in Misirlis and Vlachopoulou [60]. As such, social media engagement level can be directly attributed to the level of success of a firm's social media marketing efforts.

Incorporating the abovementioned concepts, the overall conceptual model highlighting the objective of this study is illustrated in Fig. 1. The summary of the literature review is illustrated in Table 1 for clarity.

3. Methodology

In order to determine if social media activities features can be used in conjunction with machine learning methods to predict and analyse the social media marketing effectivity, this section will elaborate on the formulation of methodology to achieve the said objective. First, the data gathering step will be explained. Second, the feature engineering and cleaning methods will be explained, with emphasis on the types of features engineered and used. Third, the use of machine learning methods will be elaborated, with emphasis on model parameter tuning steps.

3.1. Data gathering

Twitter has been chosen for this study's social media platform. Twitter is one of the most popular social media platforms and is the most prominent social media platform for interacting with other users [61]. Twitter has been proven to be beneficial for start-up firms as it is a good platform to relay their messages, engage in two-way communications, hook up with important people, and to establish a position for themselves [1]. Twitter is a social media platform that works in real-time, where a user can post a question and get an almost instant response to those posts [3]. Furthermore, in Twitter, almost two thirds of the users rely on the brand accounts they follow in Twitter to make a transaction [62,63]. Twitter allows direct communication and interaction with other users [15], and it has been shown that 65% of researched companies utilise Twitter, making it as the most commonly used social media platform among firms according to Burson–Marsteller

Table 2
Features from tweets and profiles.

Types of features	Features	Description
Linguistic	Length of tweets	Refers to the total average length of tweets for each company
	Sentiment analysis of replies and regular tweets	Refers to the total average positive, negative, neutral, and compound scores of sentiment analysis of replies and regular tweets of each company
Activity	Number of tweets	Total number of tweets posted by each company
	Likes received	Total number of likes received by each company
	Retweets received	Total number of retweets received by each company
	Favourites count	Total number of favourites received by each company
	Favourites per tweets	Number of favourites per tweets of each company

Table 3
Sentiment-related features.

Features	Description
Positive replies	The total average of positive scores for tweets that are replies for each company
Negative replies	The total average of negative scores for tweets that are replies for each company
Neutral replies	The total average of neutral scores for tweets that are replies for each company
Compound replies	The total average of compound scores for tweets that are replies for each company
Positive regular	The total average of positive scores for regular tweets for each company
Negative regular	The total average of negative scores for regular tweets for each company
Neutral regular	The total average of neutral scores for regular tweets for each company
Compound regular	The total average of compound scores for regular tweets for each company

report in 2010. Kuruzovich, et al. [16] further stated that "more Twitter activity imply that there will exist more possibilities for the level of engagement to be enhanced".

Twitter accounts of 8434 start-up firms were collected from startups-list.com, a website for investors showing all start-up firms (established within the last 10 years with small to medium revenue size) and newly established companies (established less than 5 years ago) within a particular city in the USA. Through Python programming language and Beautifulsoup API, Twitter accounts of start-up firms listed in Atlanta, Austin, Boston, Chicago, Houston, Los Angeles, New York, San Francisco, Seattle, Washington D.C., Dallas, Denver, Detroit, Las Vegas, Miami, New Orleans, Oakland, Oklahoma, Philadelphia, Phoenix, Portland, Sacramento, Salt Lake City, San Diego, and San Jose were gathered. Using Python's Twitter API, the Twitter profiles of start-up firms containing account IDs, number of tweets, number of followers, etc., were gathered. Using the same Twitter API, a maximum of 200 tweets for each start-up firm were collected. Regular Twitter API only allows a maximum of 200 tweets to be collected for each account. Some companies have fewer than 200 tweets while others had more than 200 in total.

3.2. Feature engineering

As explained above, the social media features used in this study are linguistic and activity features, with linguistic features including the length of content and psycholinguistic features, while activity features including the number of activities performed. A more detailed description of the social media features extracted are presented in Table 2.

Features reflecting the sentiment scores of tweets were created using VADER sentiment analysis API for Python. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a "lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media" [64]. This tool was chosen as it describes how positive or negative a sentiment is rather than simply showing positive and negative scores, does not require training of data, analysis is much easier, and handles emojis, slangs, and acronyms in a sentence, making it ideal for analysing tweets. Four different scores are returned, positive, negative, neutral, and compound. Positive, negative, and neutral scores reflect how positive, negative, or neutral a sentence is, respectively, while compound score reflects the overall tone of the sentence, with an example score of 0.16 meaning the tone to be quite negative. As according to Antretter and Arnaboldi [65], tweets were divided into replies and regular tweets to segregate the different conditions and situations of tones of tweets, depending on whether a tweet is directly communicating with another person, or if it is a general message for the general public. Several sentiment-related features were created as shown in Table 3.

Table 4
Engagement level categories.

Engagement rate	Category (numerical assignments)
0%–0.02%	Low (0)
0.02%–0.09%	Good (1)
0.09%–0.33%	High (2)
0.33%–1%	Very high (3)
More than 1%	Highest (4)

The target feature of this study is the engagement level of each start-up firms. This study utilises the definition of engagement as certain amounts of activity performed. More specifically, engagement refers to the culmination of all interactive activities performed within the platform, as defined by Twitter. A more specific method of calculating engagement is shown in Eq. (1). To gauge the level of engagement of each firm, this study utilises the methods used by Scrunch, a consulting firm dedicated to social media marketing consulting [66]. In Scrunch, the level of engagement is classified through engagement rates. Engagement rate is a numerical representation of how involved an audience of a user is. The formula for calculating engagement rate as defined by Scrunch is shown in Eq. (2).

$$\text{Engagement} = \text{total likes received} + \text{total retweets received} + \text{total number of favourites received} \tag{1}$$

$$\text{Engagement rate} = (\text{engagement}/\text{total number of followers})/\text{total number of tweets} \tag{2}$$

Scrunch has also classified engagement rates to different levels of engagement. By using engagement levels, it will be easier to gauge how engaged the audience of a user is, rather than trying to analyse it through vague numerical representations. The engagement level classification is defined in Table 4. The complete list of features generated are presented in Table 5.

The main metric used is accuracy, as accuracy refers to the percentage of data that the models predicted correctly. This refers to the percentage of engagement levels of start-up firms that the model predicted correctly. Accuracy is also the best simplistic metric to be used

Table 5
All features generated.

Types of features	Features	Description
Linguistic	Length of tweets	Refers to the total average length of tweets for each company
	Positive replies	The total average of positive scores for tweets that are replies for each company
	Negative replies	The total average of negative scores for tweets that are replies for each company
	Neutral replies	The total average of neutral scores for tweets that are replies for each company
	Compound replies	The total average of compound scores for tweets that are replies for each company
	Positive regular	The total average of positive scores for regular tweets for each company
	Negative regular	The total average of negative scores for regular tweets for each company
	Neutral regular	The total average of neutral scores for regular tweets for each company
	Compound regular	The total average of compound scores for regular tweets for each company
Activity	Number of tweets	Total number of tweets posted by each company
	Likes received	Total number of likes received by each company
	Retweets received	Total number of retweets received by each company
	Favourites count	Total number of favourites received by each company
	Favourites per tweets	Number of favourites per tweets of each company
Target	Engagement level	Level of engagement of each company

for determining the performance of multiclass classification models in such analysis. For further comparison between each model, other metrics such as precision, recall, and F-1 score are calculated for each engagement level to further analyse the classification performed.

3.3. Data cleaning

To get rid of anomalies in the data, several data cleaning steps were taken. First, start-up firms with 0 tweet lengths were excluded as this meant that the start-up firms did not take part in any social media marketing activities even though they had Twitter accounts. Start-up firms with less than 50 tweets were also excluded as they do not contribute much to the experiment. Start-up firms that do not have any number of following users or followers were also excluded since they do not have any audiences to communicate with. Start-up firms with engagement rates that exceed 1 were also excluded since according to the definition by Scrunch, engagement rate should exist between 0 and 1, and engagement rates above 1 are considered as outliers.

3.4. Machine learning models

Machine learning models used in this study are capable of handling multiclass classification problems, since the target feature, engagement level, is comprised of five classes. The machine learning methods used here are decision trees, random forest, deep learning, nearest neighbour, logistic regression, and gradient boosting. Decision trees, as defined by [67] is “a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree”. Information gain is utilised to determine the separation paths, using impurity measures such as entropy, Gini impurity, etc. Random forest is an ensemble method that involves random generations of decision trees during training, and the classification of an input is done through averaging the outputs of the decision trees. This solves the problem of overfitting in decision trees. Deep learning has been explained in the related literature above. The nearest neighbour algorithm, specifically K-Nearest Neighbours, is a model that assumes that similar objects can be found in close proximity, using similarity metrics such as distance between each object [68]. Nearest neighbour algorithms are used well for identifying objects or classes of same nature being near each other. Logistic regression is a simple machine learning model that is useful when the prediction classes are discrete. It calculates the probability of the input being in whichever class using a decision function called Logistic Function. Gradient boosting models, more specifically XGBoost, is to improve the limitations of decision trees by applying gradients in the loss function, a function that shows how well a model’s coefficients fit with the given data [69]. XGBoost is one such gradient boosted decision trees method to improve speed and performance, as often shown in Kaggle challenges [70].

3.5. Hyperparameter tuning

With machine learning methods, optimisation is key in obtaining the best performance. Various hyperparameters of machine learning models have to be adjusted manually to find the optimal model solutions. Through cross validation, different values of hyperparameters can be tested simultaneously, allowing users to observe and compare the different performances with different hyperparameters. As such, all machine learning models utilised in this study are cross-validated and tested several times to obtain the hyperparameters for best performances.

For decision trees, the hyperparameter to be found is the maximum depth of the trees. Random forest’s hyperparameters consist of number of estimators, maximum features, maximum depth, minimum samples per split, minimum samples per leaf, and bootstrapping. Number of estimators refer to the number of trees in the random forest, maximum features refer to the number of features to consider at each split, maximum depth refers to the maximum number of levels in the tree, minimum samples per split refers to the minimum number of samples required to split a node, minimum number of samples per leaf refers to the minimum number of samples required at each leaf node, and the bootstrapping refers to selecting samples for training each tree. For deep learning, the hyperparameters are numerous. One hyperparameter is the optimiser to use. Optimisation algorithms seek to minimise or maximise the error function, optimising the model parameters for learning. The learning rate and momentum are also hyperparameters to consider when the optimisation algorithm is determined. Momentum accelerates the stochastic gradient descent by controlling the direction of descent. Learning rate simply refers to the rate of learning for optimisation algorithms. Another hyperparameter is the neuron activation function. In deep learning, there are several activation functions that can be used for each neuron to determine if the neurons should be activated or not. Another possible hyperparameter is the dropout rate. Dropout rate refers to randomly dropping out some layer outputs. Another hyperparameter is the weight initialisation. Weight initialisation for deep learning refers to the initial assignments of weights of nodes. Other important hyperparameters are the number of neurons and the number of hidden layers. For nearest neighbour, the hyperparameters are the number of neighbours, the weight function for prediction, the algorithm to be used to find the nearest neighbours, the leaf size that can control the speed of the model construction, and the power parameter for the Minkowski metric. In logistic regression, since this problem is a multiclass classification problem, the parameter “multi_class” has to be set to “multinomial”. This allows the loss to be the multinomial loss for each class. The hyperparameters to be found are penalty for the norm used in penalisation, the solver to use in optimisation problem, the C parameter for regularisation strength, and the maximum iteration for convergence. For XGBoost, the hyperparameters to be determined are minimum child weight to determine

Table 6
Machine learning hyperparameters.

Machine learning models	Hyperparameters
Decision trees	Maximum depth
Random forest	Number of estimators Minimum samples per split Minimum samples per leaf Maximum features Maximum depth Bootstrap
Deep learning	Optimiser Learning rate Weight initialisation Activation function Dropout rate Number of hidden layers Number of neurons
Nearest neighbour	Algorithm Leaf size Number of neighbours P Weights
Logistic regression	Multi_class C Maximum iteration Penalty Solver
Gradient boosting	Subsample ratio of columns Gamma Maximum depth Minimum child weight Subsample ratio

how conservative the algorithm is to be, gamma value indicating the minimum loss reduction to partition a leaf node of the tree, subsample ratio of training, the subsample ratio of columns when building each tree, and the maximum depth of the trees. All hyperparameters to be determined for each model are illustrated in Table 6.

4. Results and discussion

4.1. Hyperparameter values

For decision trees, the maximum depth is found to be 8. The best hyperparameters for random forest are 200 number of estimators, 5 minimum samples per split, 4 minimum samples per leaf, automatic determination of maximum features, 100 maximum depth, and not using bootstrapping methods. The deep learning model performed best when it utilised Adamax for optimisation algorithm, learning rate of 0.001, uniform weight initialisation, neuron activation function of “softplus”, dropout rate of 0.1, and 7 hidden layers with 100 neurons each. Nearest neighbours performed best when the algorithm used is “auto”, leaf size is 10, number of neighbours is 10, power parameter is 1, and the weight function as “uniform”. For logistic regression, the C value of 2, maximum iteration of 200, penalty of “l2”, and solver of “newton-cg” are the optimal values. XGBoost performed best when the subsampling ratio of columns is 0.8, gamma value is 0.5, the maximum depth is 3, the minimum child weight of 5, and the subsampling ratio of 0.8. The optimum hyperparameters for each machine learning model are illustrated in Table 7 below. This illustrates the method of searching for the hyperparameters depending on both the machine learning model and the task to be solved. It is optimal to look for corresponding hyperparameters whenever a new task is being done.

4.2. Accuracy results

The accuracy results for the implemented machine learning models are shown in Table 8. Among all machine learning models utilised, deep learning has shown the highest accuracy level, with 73.42% accuracy. The machine learning method with the lowest accuracy level is logistic regression with 65.95%, which is expected since the feature dimension is vast and logistic regression is the most simplistic method among all other models used. Decision tree has an accuracy score of 69.21%, and random forest improving to 71.06%. Since random forest is an ensemble method of numerous decision trees to optimally minimise bias and variance errors, it makes sense that random forest will perform better than regular decision trees method. Both the nearest neighbour and gradient boosting had a score of 71.40% and 72.12%, respectively, which shows the predictive abilities of both models. These results show that deep learning performs the best, with the best predictive accuracy, allowing numerous implications for both academic and start-up firms. Firstly, with deep learning providing better predictive accuracy, this calls for start-up firms to invest more time and effort in utilising more complex analytical tools for business intelligence purposes. Deep learning may not necessarily provide better results than other classical machine learning methods when the data being utilised is too simplistic. However, in the case of social media marketing features, the data complexity is sufficient and more elaborate features can be added to allow more insightful and elaborate prediction tasks.

4.3. Comparison with traditional methods

As to gauge the predictive performance of this study’s machine learning methods, they have been compared with more traditional methods of marketing forecasts. Social media marketing forecasting is not a widely studied field, therefore making direct comparisons difficult. This study serves as the first contributor of advanced social media marketing forecasting method. Furthermore, in considering the domain of marketing forecasting and the effects of marketing efforts of start-up firms, the accuracy scores of machine learning models are compared with traditional sales forecasting.

Traditionally, sales forecasting techniques have revolved around judgemental and statistical analyses [71]. For start-up firms that face financial constraints, they do not have the luxury to participate in such analyses as they often involve third-party analytics firms and experts, incrementing costs. Furthermore, based on a study by Gartner and Thomas [72], out of the 103 companies tested to identify the factors affecting a start-up firm’s ability to predict new product sales accurately, only 47 of them predicted accurately, with the remaining 56 of them predicting incorrectly. Furthermore, using the traditional sales forecasting methods, Tull [73] found out that firms had an average of 65% mean error, leaving mean accuracy of 35%. Therefore, although the comparison is between prediction of engagement level and sales, in comparison with such traditional methods of marketing forecasting analysis, it can be seen that utilising machine learning methods can improve the forecasting abilities. This is also reflected in Table 9. Fig. 2 illustrates the overall accuracy score comparison of all models with the previous sales forecasting accuracy scores.

Some implications for start-up firms are as follows. Traditional marketing forecasting methods rely on intuition and simple statistical analysis that do not result in great accuracy results. Even with easily attainable data from social media platforms, machine learning methods outperform traditional methods of forecasting in a field more extensively being studied than social media marketing forecasting. Greater accuracy in such prediction can result in more clairvoyant insights into a firm’s real-time or current performance of its marketing efforts and the characteristics of the market that they are participating in. If start-up firms can gather bigger and more complex data for analysis, machine learning methods can bring about more impressive accuracy results, unlike the traditional methods of analysis. Furthermore, this study’s methodology most likely does not require any third-party analytic firms due to the availability of data, reducing cost of analysis and time spent. Therefore, through the results of this methodology, it can be argued that prediction of social media marketing effects can be done to fulfil its growing necessity by incorporating data science and machine learning methods.

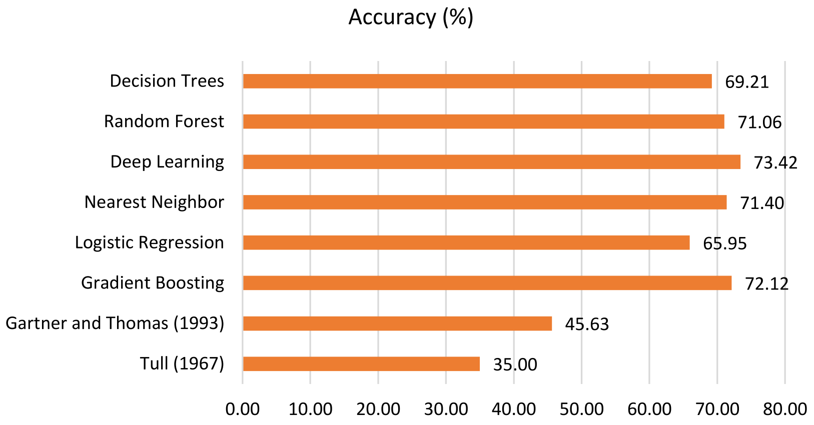


Fig. 2. Accuracy comparison of all models.

Table 7
Optimum machine learning hyperparameters.

Machine learning models	Hyperparameters	Optimum hyperparameters
Decision trees	Maximum depth	8
Random forest	Number of estimators	200
	Minimum samples per split	5
	Minimum samples per leaf	4
	Maximum features	'auto'
	Maximum depth	100
	Bootstrap	False
Deep learning	Optimiser	Adamax
	Learning rate	0.001
	Weight initialisation	'uniform'
	Activation function	'softplus'
	Dropout rate	0.1
	Number of hidden layers	7
	Number of neurons	100
Nearest neighbour	Algorithm	'auto'
	Leaf size	10
	Number of neighbours	10
	P	1
	Weights	'uniform'
Logistic regression	Multi_class	'multinomial'
	C	2.0
	Maximum iteration	200
	Penalty	'l2'
	Solver	'newton-cg'
Gradient boosting	Subsample ratio of columns	0.8
	Gamma	0.5
	Maximum depth	3
	Minimum child weight	5
	Subsample ratio	0.8

Table 8
Accuracy scores of each model.

Model	Accuracy (%)
Decision trees	69.21
Random forest	71.06
Deep learning	**73.42**
Nearest neighbour	71.40
Logistic regression	65.95
Gradient boosting	72.12

4.4. Additional evaluation metrics

Since this problem is a multiclass classification problem, precision, recall, and F1-score of each engagement level can be determined to examine the classification performance in more detail. This analysis is shown in Table 10 below. The distribution of each class is also illustrated in Table 11. The test data utilised is the same for all models, resulting in the same distribution of classes for each model prediction.

Table 9
Sales forecasting accuracy scores of previous studies.

Reference	Accuracy (%)
[72]	45.63
[73]	35.00

From the above analysis, it can be observed that the classes are unbalanced for this dataset, resulting in the machine learning models not being able to classify them properly. For all models, the precision, recall, and F1-score values of class 0 and 4 are high, with the highest precision being 0.90, highest recall being 0.96, and highest F1-score being 0.91 for class 0, and highest precision of class 4 being 0.85, 0.87, and 0.83. This shows an interesting implication that with enough given data, consisting of more start-up firms in class 1, 2, and 3, all machine learning models used here will be able to predict the classes of all start-up firms with ease. Figs. 3 and 4 illustrate the distribution metrics of deep learning and logistic regression, respectively, to show

Table 10
Analysis of each class.

Metrics	Engagement class	Decision trees	Random forest	Deep learning	Nearest neighbour	Logistic regression	Gradient boosting
Precision	0	0.89	0.84	**0.90**	0.87	0.74	0.88
	1	0.33	0.37	0.45	0.41	0.33	0.39
	2	0.32	0.38	0.43	0.42	0.26	0.42
	3	0.38	0.44	0.44	0.46	0.36	0.44
	4	0.77	0.74	0.85	0.79	0.64	0.79
Recall	0	0.87	0.92	0.92	0.93	**0.96**	0.92
	1	0.43	0.36	0.49	0.39	0.19	0.49
	2	0.29	0.35	0.37	0.40	0.13	0.31
	3	0.30	0.19	0.45	0.26	0.01	0.28
	4	0.78	0.86	0.80	0.87	0.84	0.85
F1-score	0	0.88	0.88	**0.91**	0.90	0.84	0.90
	1	0.37	0.36	0.47	0.40	0.24	0.43
	2	0.31	0.36	0.40	0.41	0.17	0.36
	3	0.34	0.27	0.45	0.34	0.04	0.34
	4	0.78	0.79	0.83	0.82	0.73	0.82

Table 11
Distribution of classes.

	Engagement class	Test data distribution
	0	773
	1	211
	2	188
	3	170
	4	345
Total test data size		1687

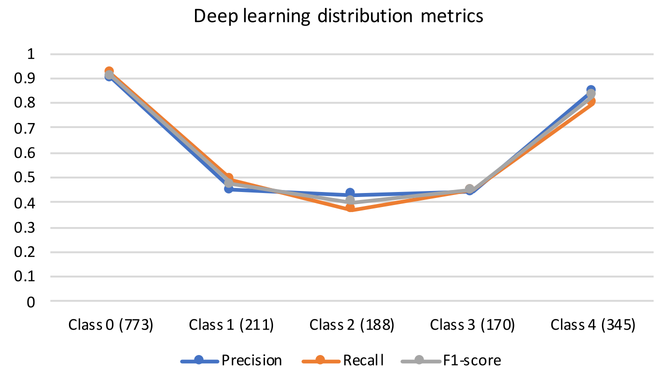


Fig. 3. Deep learning distribution metrics. Illustrates the stability of deep learning prediction.

the metric scores of each distribution according to the best and the worst overall prediction accuracy. It can be observed that deep learning is more stable in its distribution of metrics overall, with precision, recall, and F1-score all being of somewhat similar distribution for all classes. However, logistic regression shows more sporadic distribution of metrics, showing a greater fluctuation in regard to the data size of classes. From such analysis, deep learning shows more stable prediction distribution in comparison to logistic regression, resulting in better overall prediction accuracy. This also shows that machine learning models are sensitive to the data size. With more balanced data provided, machine learning will be able to predict all classes with more balance. With more sophisticated models such as deep learning, the imbalance in the dataset is somewhat mitigated as compared to more simplistic models such as logistic regression. This also implies that proper data preparation will result in better predictive models, making data inspection crucial for such analyses.

4.5. Most significant features

With machine learning methods, it is possible to identify the most significant features in predicting the target feature. As the result shows that deep learning is the best performer among all machine learning methods tried, the most significant features while using deep learning are identified through the use of permutation importance. Permutation importance is done with a model that has already been fitted onto a dataset. It will randomly shuffle one column of the validation dataset, calculating the effects of the permutation in the prediction accuracy. The less accurate the prediction is for permutation of a column, the more important that particular column is in prediction the target feature. The permutation importance result of deep learning is presented in Fig. 5, with the values towards the top showing most important features, while those towards bottom show least important features.

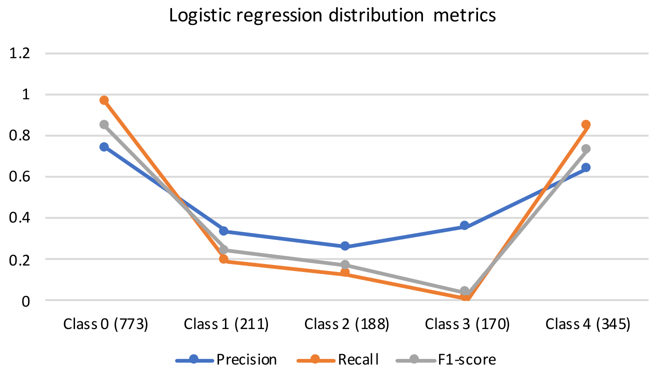
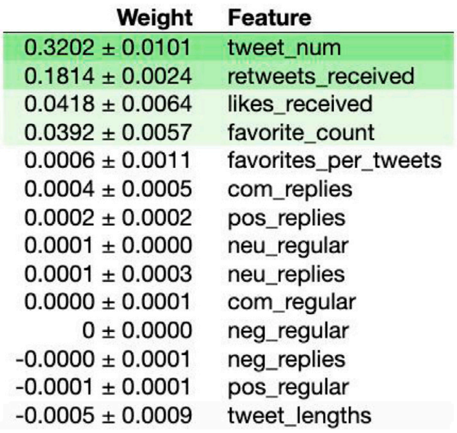


Fig. 4. Logistic regression distribution metrics. Illustrates the more sporadic distribution of prediction.



Weight	Feature
0.3202 ± 0.0101	tweet_num
0.1814 ± 0.0024	retweets_received
0.0418 ± 0.0064	likes_received
0.0392 ± 0.0057	favorite_count
0.0006 ± 0.0011	favorites_per_tweets
0.0004 ± 0.0005	com_replies
0.0002 ± 0.0002	pos_replies
0.0001 ± 0.0000	neu_regular
0.0001 ± 0.0003	neu_replies
0.0000 ± 0.0001	com_regular
0 ± 0.0000	neg_regular
-0.0000 ± 0.0001	neg_replies
-0.0001 ± 0.0001	pos_regular
-0.0005 ± 0.0009	tweet_lengths

Fig. 5. Permutation importance result of deep learning.

The results show that the number of tweets by the start-up firms matter the most in predicting the social media marketing effects, followed by the number of retweets received and the number of likes received on their posts. This implies that start-up firms must pay particular attention to the number of tweets that they post and must strive to maximise their number of retweets and likes received on their posts. This might seem straightforward but seeing that the sentiments related to these reactions are not that significant, this implies that start-up firms should try to maximise their exposure instead of trying to control the sentiments of their reactions. Furthermore, the lengths of tweets are not significant at all for this prediction task, implying that the messages that the start-up firms send out do not have to be particularly long or short. This further shows that start-up firms only need frequent exposure first in getting their desired effects.

4.6. Further discussion and limitations

Since deep learning is the best performer among the machine learning models used, researchers and analysts should move to different platforms and analysis methods to improve data analysis of business processes and strategies. Deep learning will also be able to have better accuracy in predicting social media engagement levels with larger amount of data being analysed. Deep learning performance greatly improves with more data involved, and big data in business is becoming more and more prominent. Social media platforms are also full of raw data, and through proper cleaning and usage of such data, start-up firms can improve their predictive analysis. For start-up firms, predicting aspects of their strategic business processes is necessary to stay vigilant and up to date. Additionally, big data coupled with data science analysis methods utilising deep learning can provide tremendous predictive insights that start-up firms should seek to utilise. Use of a more accurate prediction model can also allow start-up firms to better convince and entice investors that they need. With this in mind, analysts and researchers should move to a different analytics platform to perform various business intelligence analysis. More specifically, computer science techniques, such as Python and R language, should be combined with business intelligence analytics, more so in marketing. Start-up firms should endeavour to gain predictive business insights using big data, such as social media data, to gain better insights into their markets and themselves, as well as to gain competitive advantage and enhance strategic management.

Furthermore, such methodology of social media data analytics can be applied to business practices, such as in marketing or in other areas of business. Using similar methods, start-up firms can predict their sales and revenue using historical data of themselves or historical data of similar firms that have been existing in the industry before them. Other than that, social media marketing features can be directly related to the future sales and revenues of the firms, which future research can try to solve. Apart from sales and revenues, intangible assets of the start-up firms can be identified, established, and predicted using this methodology. Social media platforms allow several inherent relationships and benefits to be formed between consumers and firms. In this case, the feature to be used can be in the form of relationship links and other social media features that reflect the interpersonal relationships. This method can also be utilised in the industry level, where the contents and reactions of social media platforms can be used to predict the market condition of the industry. Apart from that, industry level revenues and sales can be predicted using insights from social media platforms. All these can be done in future studies, further elaborating the possibilities and benefits of using social media features as business intelligence features.

Some limitations of this study are lack of data used, simplistic dataset, and lack of direct impact of such prediction of social media engagement level on start-up firms' revenues. Data size is important in predictive analysis, especially when it involves machine learning. Specifically, for deep learning, big data is crucial in its predictive abilities, and the data size used in this study may not show the full potential of the predictive model. It would be beneficial for future studies to implement big data from social media platforms to perform similar analysis. Number of features generated can also be a limitation of such analysis. Among similar studies relating to social media engagement prediction, this study utilised only two of the four major categories of social media data features. Utilising all four categories of features can most likely improve the predictive accuracy of the machine learning models. These features can represent the various aspects of social media data and can present more information about the raw social media data. Future studies can expand on the feature types and include all possible features to further improve the prediction accuracy.

5. Conclusion

This study seeks to address the problem of lack of methodology among start-up firms in predicting their social media marketing efforts. Social media marketing should be utilised by start-up firms more extensively since social media brings about numerous benefits for start-up firms, such as direct communication with their consumers, increasing brand awareness, and even creating a brand community, that can provide better financial performance. By predicting the social media engagement level, the market insights can be better discovered, more strategic benefits can arise, and can improve competitive advantages of the start-up firms.

This study utilised several machine learning methods to perform predictive analysis. Utilising machine learning techniques is becoming more and more easier for start-up firms since dedicated APIs using platforms, such as Python and R, allows easy implementation of such methods. Several case studies conducted by Rabbi [74] proved that utilising machine learning in business analytical areas can improve the performance and insights gained than traditional methods. Therefore, this study created a methodology of predicting the engagement level of start-up firms in social media platforms using machine learning models to gain such abovementioned benefits.

The results of this study show that deep learning is the best predictor among the machine learning methods used, with an accuracy score of 73.42%. Even with simplistic data structure and limited features generated, deep learning has performed better than the traditional marketing predictions. It is probably with high certainty that an even better result can be achieved with more data gathered and more features generated. Furthermore, the most significant features can be identified for such tasks, revealing the most important features that the start-up firms should be focusing on. In this case, the most important features were the number of tweets, the number of retweets received, and the number of likes received.

This study showed that a new methodology involving machine learning, a fast developing, state-of-the-art technology that is becoming more and more essential to integrate with business processes, can be made using cost-effective and relatively easy processes. In the case of start-up firms, such cost-effective and efficient analysis of the social media market cannot only improve their relationships with their consumers, but also gain better venture financing and financial returns.

CRediT authorship contribution statement

Sang Hoon Jung: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing, Data curation. **Yong Jin Jeong:** Supervision, Writing - review & editing.

Acknowledgements

This work was supported by Kwangwoon University, South Korea and by the MISP Korea, under the National Program for Excellence in SW (2017-0-00096) supervised by IITP.

References

[1] A. Dziadkiewicz, et al., Social media and start-ups, Zarz. Mediami 3 (2015) 257–267.

[2] B. Rugova, B. Prenaj, Social media as marketing tool for SMEs: Opportunities and challenges, Acad. J. Bus. 2 (2016).

[3] V. Sharma, S.V. Bharathi, Social media for start-ups an effective marketing tool.

[4] M. Corvera, K. Johansson, Smes and Social Media: a Study About How Smes in the Fashion Industry Should Approach Social Media, University of Borås/Swedish School of Textiles, 2012.

[5] S. Cox, Social Media Marketing in a Small Business: A Case Study, 2012.

[6] D.J. Dalrymple, Sales forecasting methods and accuracy, Bus. Horiz. 18 (1975) 69–73.

[7] D. Jobber, G. Hooley, S. Sanderson, Marketing in a hostile environment: The British textile industry, Ind. Mark. Manag. 14 (1985) 35–41.

[8] J.S. Armstrong, R. Brodie, Forecasting for Marketing, 1999.

[9] I. Bose, R.K. Mahapatra, Business data mining—a machine learning perspective, Inf. Manage. 39 (2001) 211–225.

[10] S. Yang, S. Lin, J.R. Carlson, W.T. Ross Jr., Brand engagement on social media: will firms' social media efforts influence search engine advertising effectiveness? J. Mark. Manag. 32 (2016) 526–557.

[11] T. Boone, R. Ganeshan, R.L. Hicks, N.R. Sanders, Can Google trends improve your sales forecast? Prod. Oper. Manage. 27 (2018) 1770–1774.

[12] R.Y.K. Lau, W. Zhang, W. Xu, Parallel aspect-oriented sentiment analysis for sales forecasting with big data, Prod. Oper. Manage. 27 (2018) 1775–1794.

[13] M. Reyneke, L. Pitt, P.R. Berthon, Luxury wine brand visibility in social media: An exploratory study, Int. J. Wine Bus. Res. 23 (2011) 21–35.

[14] V. Schmid, Matthew, M. Brenner, J. Weberg, Why social media is important for business marketing, 2019, URL https://marketinginsidergroup.com/content-marketing/why-social-media-is-important-for-business-marketing/.

[15] W.S. Basri, M.R.A. Siam, Maximizing the social media potential for small businesses and startups: A conceptual study, Int. J. Econ. Perspect. 11 (2017) 241–245.

[16] J. Kuruzovich, Y. Lu, et al., Entrepreneurs' activities on social media and venture financing, in: Proceedings of the 50th Hawaii International Conference on System Sciences, 2017.

[17] R. Aggarwal, R. Gopal, A. Gupta, H. Singh, Putting money where the mouths are: The relation between venture financing and electronic word-of-mouth, Inf. Syst. Res. 23 (2012) 976–992.

[18] B.N. Greenwood, A. Gopal, Research note—Tigerblood: Newspapers, blogs, and the founding of information technology firms, Inf. Syst. Res. 26 (2015) 812–828.

[19] W.G. Sanders, S. Boivie, Sorting things out: Valuation of new firms in uncertain markets, Strateg. Manag. J. 25 (2004) 167–186.

[20] M.S. Granovetter, The strength of weak ties, in: Social Networks, Elsevier, 1977, pp. 347–367.

[21] E. Gilbert, K. Karahalios, Predicting tie strength with social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2009, pp. 211–220.

[22] J.F. Gaski, Interrelations among a channel entity's power sources: Impact of the exercise of reward and coercion on expert, referent, and legitimate power sources, J. Mark. Res. 23 (1986) 62–77.

[23] R. Putnam, The prosperous community: Social capital and public life, in: The American Prospect, vol. 13, 1993.

[24] V. Sambamurthy, A. Bharadwaj, V. Grover, Shaping agility through digital options: Reconceptualizing the role of information technology in contemporary firms, MIS Q. (2003) 237–263.

[25] Q. Zhang, T. Ye, M. Essaidi, S. Agarwal, V. Liu, B.T. Loo, Predicting startup crowdfunding success through longitudinal social engagement analysis, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, 2017, pp. 1937–1946.

[26] L. Dugan, 4 social media marketing strategies for startups: Simply measured, 2019, URL http://gundersonmarketing.com/4-social-media-marketing-strategies-for-startups-simply-measured/.

[27] N. Deepa, S. Deshmukh, Social media marketing: The next generation of business engagement, Int. J. Manag. Res. Rev. 3 (2013) 2461.

[28] D. Sprott, S. Czellar, E. Spangenberg, The importance of a general measure of brand engagement on market behavior: Development and validation of a scale, J. Mark. Res. 46 (2009) 92–104.

[29] E.-P. Lim, H. Chen, G. Chen, Business intelligence and analytics: Research directions, ACM Trans. Manag. Inf. Sys. 3 (2013) 17.

[30] M.J. Mortenson, N.F. Doherty, S. Robinson, Operational research from Taylorism to Terabytes: A research agenda for the analytics age, European J. Oper. Res. 241 (2015) 583–595.

[31] J.C. Ranyard, R. Fildes, T.-I. Hu, Reassessing the scope of OR practice: The influences of problem structuring methods and the analytics movement, European J. Oper. Res. 245 (2015) 1–13.

[32] Stratlytics, Machine learning in business, 2016.

[33] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[35] M. Kraus, S. Feuerriegel, A. Oztekin, Deep learning in business analytics and operations research: Models, applications and managerial implications, European J. Oper. Res. (2019).

[36] G. George, M.R. Haas, A. Pentland, Big Data and Management, Academy of Management Briarcliff Manor, NY, 2014.

[37] G. George, E.C. Osinga, D. Lavie, B.A. Scott, Big Data and Data Science Methods for Management Research, Academy of Management Briarcliff Manor, NY, 2016.

[38] F. Sadeque, S. Bethard, Predicting engagement in online social networks: Challenges and opportunities, 2019, arXiv preprint arXiv:1907.05442.

[39] J. Mahmud, J. Chen, J. Nichols, Why are you more engaged? predicting social engagement from word use, 2014, arXiv preprint arXiv:1402.6690.

[40] F. Sadeque, T. Solorio, T. Pedersen, P. Shrestha, S. Bethard, Predicting continued participation in online health forums, in: Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, 2015, pp. 12–20.

[41] J. Chen, P. Pirolli, Why you are more engaged: factors influencing twitter engagement in occupy Wall Street, in: Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[42] E. Joyce, R.E. Kraut, Predicting continued participation in newsgroups, J. Computer-Mediated Commun. 11 (2006) 723–747.

[43] W.L. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, J. Leskovec, Loyalty in online communities, in: Eleventh International AAAI Conference on Web and Social Media, 2017.

[44] J. Arguello, B.S. Butler, E. Joyce, R. Kraut, K.S. Ling, C. Rosé, X. Wang, Talk to me: foundations for successful individual-group interactions in online communities, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2006, pp. 959–968.

[45] M. Milošević, N. Živić, I. Andjelković, Early churn prediction with personalized targeting in mobile social games, Expert Syst. Appl. 83 (2017) 326–332.

[46] F. Sadeque, T. Pedersen, T. Solorio, P. Shrestha, N. Rey-Villamizar, S. Bethard, Why do they leave: Modeling participation in online depression forums, in: Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media, 2016, pp. 14–19.

[47] M.-C. Perreault, E. Mosconi, Social media engagement: Content strategy and metrics research opportunities, in: Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

[48] J.H. Moon, E. Kim, S.M. Choi, Y. Sung, Keep the social in social media: The role of social interaction in avatar-based virtual shopping, J. Interact. Adv. 13 (2013) 14–26.

[49] C. Oh, Y. Roumani, J.K. Nwankpa, H.-F. Hu, Beyond likes and tweets: Consumer engagement behavior and movie box office in social media, Inf. Manage. 54 (2017) 25–37.

[50] D. Agostino, M. Arnaboldi, A measurement framework for assessing the contribution of social media to public engagement: An empirical analysis on Facebook, Public Manag. Rev. 18 (2016) 1289–1307.

[51] C. Campbell, L.F. Pitt, M. Parent, P.R. Berthon, Understanding consumer conversations around ads in a web 2.0 world, J. Advert. 40 (2011) 87–102.

[52] S.-C. Chu, Y. Kim, Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites, Int. J. Advert. 30 (2011) 47–75.

[53] T. Daugherty, E. Hoffman, EWOM and the importance of capturing consumer attention within social media, J. Mark. Commun. 20 (2014) 82–102.

[54] X. Zheng, C.M. Cheung, M.K. Lee, L. Liang, Building brand loyalty through user engagement in online brand communities in social networking sites, Inf. Technol. People 28 (2015) 90–106.

[55] S. Boulianne, Social media use and participation: A meta-analysis of current research, Inf. Commun. Soc. 18 (2015) 524–538.

[56] J. Gummerus, V. Liljander, E. Weman, M. Pihlström, Customer engagement in a Facebook brand community, Manag. Res. Rev. 35 (2012) 857–877.

[57] V.A. Barger, L. Labrecque, An integrated marketing communications perspective on social media metrics, Int. J. Integr. Mark. Commun. Spring (2013).

[58] Z. Rahman, et al., The impact of social media engagement metrics on purchase intention: A study on brand fan page followers|, in: LUMEN Proceedings, vol. 1, Editura Lumen, 2017, pp. 665–681.

[59] S. Blaschke, A. Veh, Strategies for the use of social media in stakeholder conversations (Strategien für den Einsatz sozialer Medien in Stakeholderkonversationen). Die Betr. 75 (2015) 401.

[60] N. Misirlis, M. Vlachopoulou, Social media metrics and analytics in marketing–S3M: A mapping literature review, Int. J. Inf. Manage. 38 (2018) 270–276.

[61] T. Hennig-Thurau, C. Wiertz, F. Feldhaus, Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies, J. Acad. Mark. Sci. 43 (2015) 375–394.

[62] S. Edwards, A social media guide for startups and entrepreneurs, 2015, URL https://www.inc.com/samuel-edwards/a-social-media-guide-for-startups-and-entrepreneurs.html.

[63] K. Wagner, Linkedin: 81use social media, 2014, URL https://mashable.com/2014/02/13/linkedin-social-media-study/.

[64] P. Pandey, Simplifying sentiment analysis using vader in Python (on social media text), 2019, URL https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f.

[65] T. Antretter, I. Blohm, D. Grichnik, J. Wincent, Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy, J. Bus. Ventur. Insights 11 (2019) e00109.

[66] G. Mee, What is a good engagement rate on Twitter? URL https://www.scrunch.com/blog/what-is-a-good-engagement-rate-on-twitter.

[67] T.M. Mitchell, Machine Learning, McGraw hill, 1997.

[68] O. Harrison, Machine learning basics with the k-nearest neighbors algorithm, 2019, URL https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[69] H. Singh, Understanding gradient boosting machines, 2018, URL https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab.

[70] J. Brownlee, A gentle introduction to XGBoost for applied machine learning, 2019, URL https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.

[71] M.J. Baker, The IEBM Encyclopedia of Marketing, International Thomson business press, 2001, pp. 278–290.

[72] W.B. Gartner, R.J. Thomas, Factors affecting new product forecasting accuracy in new firms, J. Prod. Innov. Manage. 10 (1993) 35–52.

[73] D.S. Tull, The relationship of actual and predicted sales and profits in new-product introductions, J. Bus. 40 (1967) 233–250.

[74] F. Rabbi, A review of the recent trends in the use of machine learning in business.

Sang Hoon Jung: A graduate student of Department of Electronics and Communications at Kwangwoon University, South Korea. Received a B.S. degree in Business Management in Ateneo de Davao University. Main research interests are data science works related to business intelligence, big data analytics, and machine learning modelling of business strategic management.

Yong Jin Jeong received his B.S. degree in Control and Instrumentation Engineering from Seoul National University in 1983. He received his M.S. and Ph.D. degrees in Electrical and Computer Engineering from University of Massachusetts. He is currently a professor in the Dept. of Electronics and Communications Engineering, Kwangwoon University.