A comprehensive analysis of consumer decisions on Twitter dataset using machine learning algorithms

Vigneshwaran Pandi[1], Prasath Nithiyanandam[1], Sindhuja Manickavasagam[2], Islabudeen Mohamed Meerasha[3], Ragaventhiran Jaganathan[4], Muthu Kumar Balasubramanian[4]

[1]Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India
[2]Department of Information Technology, Rajalakshmi Engineering College, Thandalam, Chennai, India
[3]Department of Computer Science and Engineering, School of Engineering, Presidency University, Bengaluru, India
[4]School of Computing and Information Technology, REVA University, Bengaluru, India

Article Info

Article history:

Received Jul 15, 2021
Revised Jun 2, 2022
Accepted Jun 20, 2022

Keywords:

Artificial intelligence
E-commerce
Machine learning
Sentimental analysis
Twitter dataset

ABSTRACT

An exponential growth posting on the web about the product reviews on social media, there has been a great deal of examination being done on sorting out the purchasing behaviors of the client. This paper depends on utilizing twitter for sentiment analysis to comprehend the customer purchasing behavior. There has been a significant increase in e-commerce, particularly in persons purchasing products on the internet. As a result, it becomes a fertile hotspot for opinion analysis and belief mining. In this investigation, we look at the problem of recognizing and anticipating a client's purchase goal for an item. The sentiment analysis helps to arrive at a more indisputable outcome. In this study, the support vector machine, naive Bayes, and logistic regression methods are investigated for understanding the customer's sentiment or opinion on a specific product. These strategies have been demonstrated to be genuinely for making predictions using the analysis models which examine the client's conclusion/sentiment the most precisely. The exactness for each machine learning algorithm will be analyzed and the calculation which is the most precise would be viewed as ideal.

Corresponding Author:

Vigneshwaran Pandi
Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology
SRM Nagar, Kattankulathur - 603 203, Chengalpattu District, Chennai, India
Email: vigenesp@srmist.edu.in

1. INTRODUCTION

Social media has become one of the most important channels for communication and content generation. It fills in as a bound together stage for clients to communicate their contemplations on subjects going from their day by day lives to their sentiment on organizations and items. This, thus, has made it a significant asset for digging client feelings for errands going from anticipating the exhibition of films to aftereffects of stock market exchanges and races. Even though the vast majority is reluctant to answer reviews about items or administrations, they express their considerations unreservedly via online media and employ a huge impact in molding the assessments of different buyers. These customer voices can impact brand recognition, brand dedication and brand support. Therefore, it is basic that big companies give more consideration to mining client assessment identified with their brands and items from web-based media. With web-based media checking, they will have the option to take advantage of shopper bits of knowledge to improve their item quality, offer better assistance, drive deals, and even recognize new business openings. What is more, they can lessen client care costs by reacting to their clients through these web-based media channels, as half of clients incline toward arriving at specialist organizations via online media as opposed to a call place.

It is a phenomenal device for undertakings to dissect clients' communicated conclusions via online media without expressly posing any inquiries as this methodology frequently mirrors their actual sentiments. In spite of the fact that it has disadvantages with respect to the populace examined, it very well may be utilized to surmise general assessment. The objective of this exploration is to manufacture a framework that can give exact outcomes, helping brands to see how the clients are responding to the specific item. Nowadays interpersonal organizations, web journals, and other media produce an enormous measure of information on the Internet. This tremendous measure of information contains pivotal sentiment related data that can be utilized to profit organizations and different parts of business and logical ventures. Manual following and separating this valuable data from this monstrous measure of information is practically inconceivable. Sentiment analysis of user posts is required to help take business decisions. It is a cycle which extricates notions or suppositions from audits which are given by clients over a specific subject, zone, or item on the web. Estimation may be divided into two types: i) good or ii) negative that determines an individual's overall attitude toward a given subject. Predicting the sentiment of a tweet is our main priority. Purchasing objectives are often assessed and used by advertising executives as a contribution to decisions regarding new and current goods and administrations. Until date, many businesses have used client overview frameworks in which they offer questions such, "How likely are you to buy an item in a certain time span?" and then use that data to calculate the buy goal. We need to see whether we can use Twitter tweets to train a model that can differentiate tweets that indicate a purchase intention for a product.

2. RELATED WORKS

Tan *et al.* [1] proposed interpreting public sentiment variation to be able to further understand the reason behind the shift of public opinion on product or even people. In this case, they proposed using two models: one foreground and background latent dirichlet allocation (LDA) to filter out background topics that have no significance in the most recent public sentiment variation, and the other reason candidate and background LDA to rank the various reasons based on their "popularity" in the given period. It also employed Gibb's sampling since it was simple to expand and shown to be a successful approach. A sentiment analysis tool for slang word translation was also used, which could translate slangs into legitimate terms, which may be beneficial for more accuracy. They used data from the Stanford Network Analysis Platform. The suggested approach outperformed previous models in terms of accuracy and might be used for product evaluations, scientific publications, and many other applications; it is also the first effort to assess public sentiment changes. Xia *et al.* [2] developed dual sentiment analysis to solve the polarity shift problem in sentiment analysis, which affects the entire order but is otherwise treated the same in a typical model. So, in order to address the polarity shift, they offer dual training and dual prediction algorithms to assess both original and reversed data in order to comprehend not only how positive or negative the original data is, but also how positive or negative the reversed data is. They also expanded their polarity paradigm to a three-class structure that includes neutral data. They created language-independent pseudo-antonym dictionaries to lessen their reliance on external antonym dictionaries. Support vector machine (SVM), naive Bayes, and logistic regression classifiers were used, and it was discovered that they exceed the baseline by 3.0 and 1.7% on average, respectively. Hamroun *et al.* [3] advocated using latent semantics instead of current models that employ polarity terms and matching phrases and may fail when views are stated using latent semantics, which is known as customer intents analysis. They combined OpenNLP, W3C Web Ontology Language (OWL) ontologies, and WordNet natural language processing processes with additional meanings. Their strategy was to automatically extract patterns from Twitter for consumer intention research. The idea is to use domain ontology for two key purposes: creating ontology representations and using ontology representations in pattern learning. They utilized five distinct datasets, with the continuous integration (CI) pattern outperforming the baseline by 3-6% on average.

Li *et al.* [4] proposed combining two models: Sentiment-specific word embeddings and Weighted text feature modal. Because the majority of conventional models are either lexicon-based or machine learning-based. Instead of immediately using the word embeddings approach, it will be done by first constructing vectors in order to avoid missing out on semantic hints and to enhance semantic categorization. weighted text feature model that generates two sort of features: the first is a negation feature based on negation terms, and the second is generated by computing the similarity of tweets and their polarity. The suggested strategy outperformed the previous model and separated sentiment specific word embeddings (SSWE) and (weighted text feature model (WTFM); moreover, when SSWE + word2vec was used, the performance was extremely near to SSWE. Tweepy, a Twitter Application Program Interface (API), was utilized to generate the dataset. Ren and Wu [5] created a lexicon-based learning method that is also language dependent to anticipate unknown user subject opinions. They attempted to include topical and social information into the current prediction model mathematically. They understood the association between social and topical context after applying an appropriate hypothesis and also utilized topic content similarity (TCS) to quantify the same. The findings revealed that the suggested ScTcMF framework was really superior to the existing one. The scope of the project was just for twitter and the dataset was also from twitter API. Chen *et al.* [6] evaluated a very hard constraint project which only focused on engineering students' difficulties faced during their program. Naive Bayes and multi-label classification algorithms were employed in the technique. The method used was a combination of qualitative analysis and large-scale data mining approaches. It is a machine learning method that is also language dependent. It was founded on the notion that informal social media data might give additional information about students' experiences. Purdue University provided the tweets, which included subjects ranging from sleep deprivation to food. The dataset was taken from twitter API Tweepy. Bollegala *et al.* [7] looked to address the mismatch problem arising in trained dataset and target dataset that is when the trained dataset has been for selected words and the test data does not contain those words, it creates a mismatch. In order to overcome this mismatch problem, they came up with a cross-domain sentiment classifier where they used already extracted sentiment sensitive words and were able to determine that the existing models such as SentiWordNet, which is a lexical resource were outperformed by cross-domain classifier. It also uses a lexical based approach and is a language dependent model aimed mainly at product reviews and the dataset was taken from amazon.com.

Lin *et al.* [8] presented a joint sentiment analysis model as well as a reparametrized version of supervised joint sentiment-topic because it was frequently observed that the weakly supervised joint sentiment topic, which is a component of LDA, failed to produce acceptable performance when shifting to new domains. As a result, our model can now recognize both sentiment and the subject of a certain data set. It is a machine learning method that is also language dependent. The dataset came from Amazon.com and IMDB.com and was based on product or movie reviews. Wang *et al.* [9] proposed that for complete sentiment analysis of a tweet, we should also consider hashtags as complete words, and that three types of information are required to generate the complete sentiment polarity for hashtag, which differs from sentence and document level sentiment analysis. They also suggested using improved boosting classification, which would allow us to use the literal meaning of hashtags as a semi-supervised training set. To construct the hashtag sentiment, they utilized an SVM classifier; it was a language dependent model for the Twitter dataset. Mudinas *et al.* [10] assessed both lexicon-only and learning-only approaches and presented a hybrid strategy that takes the best of both worlds from lexicon and learning-only algorithms. When they ran the experiment, they discovered that the sentiment polarity classification and sentiment strength detection values in their pSenti system were higher, which is very near to the pure learning model and higher than the pure lexicon model. It was language-specific and used both machine learning and lexical models. This model was created for software and movie reviews, including data from computer network (CNET) and internet movie database (IMDB). Yu *et al.* [11] built their whole research around a movie domain case study and assessed the difficulty of forecasting sales using sentiment analysis. They investigated several hidden sentiment components in order to use sentiment Probabilistic Latent Sentiment Analysis (PLSA) to evaluate complicated forms of sentiment. They then suggested an updated version of the auto-regressive sentiment aware model to boost accuracy. It was a language-dependent, machine-learning-based model that focused on sales prediction in a movie-based case study. The dataset was derived from the Twitter API, Tweepy, and was created exclusively for Twitter.

Jose and Chooralil [12] evaluated and tried to address the problem with selecting just one algorithm for sentiment analysis, so they came up with the solution of combining machine learning algorithms along with lexicon-based algorithms which would choose the appropriate algorithm for its use so as to remove the risk of selecting inappropriate classifiers. They chose SentiWordNet classifier, naive Bayes classifier, and Hidden Markov model classifier, which showed to be more accurate. So, after analyzing sentiment classification on numerous tweets, they concluded that their ensemble technique produced an accuracy of roughly 71.48%, which was higher than all three classifiers combined. Kouloumpis *et al.* [13] recommended using Twitter hashtags to achieve even more accurate sentiment analysis since hashtags and emoticons may occasionally add significantly to model accuracy. In contrast to basic sentiment or non-sentiment analysis, they would employ a three-way classifier. To work on the datasets, they concentrated on n-gram features, lexicon features, and part of speech features. They employed three datasets for development and training: hashtagged dataset from Edinburgh Twitter Corpus, emoticon dataset from twittersentiment.appspot.com, and iSieve company for assessment. After doing their investigation, they discovered that combining the n-gram, lexicon, and microblogging features resulted in an accuracy of 74-75%. Park and Seo [14] used sentiment analysis to rank the three AI assistants, Siri from Apple, Cortana from Microsoft, and Google Assistant from Google, based on user feedback. They evaluated tweets using valence aware dictionary and sentiment reasoner (VADER), the Kruskal Wallis test, and the Mann-Whitney test to determine statistical significance between groups. They employed null hypotheses and the t-test to determine how the similarity of various aides varied over time.

Prakruthi *et al.* [15] assess people's feelings towards a person, trend, product, or brand. The Twitter API is used to directly retrieve tweets from Twitter and construct sentiment classifications for the tweets. The data are categorized and represented using a Histogram and a Pie Chart. The pie chart depicts the%age of positive, negative, and neutral attitude, which is believed to be roughly 65% positive, 20% negative, and 15% neutral. The histograms below depict positive, negative, and neutral emotion. Go *et al.* [16] tested many models and performed trials to identify the best classifier for organizations who wish to analyze the sentiment of their products. Twitter tweets with emoticons serve as training data. Three classifiers were used: naive Bayes, maximum entropy, and SVM; all methods had an accuracy of more than 80% when trained using emoticon data. However, the SVM was the most accurate, with an accuracy of 85%.

Trupthi *et al.* [17] want to do real-time sentiment analysis on tweets retrieved from Twitter and present the results to the user. The tools and processes used here are natural language processing. Naïve Bayes and Twitter API. Natrual Language Processing (NLP) is used to remove the words with tags which is not helpful for the building of the classifier. The tweets removed by the Streaming API are then arranged into positive, negative, or unbiased tweets. The analytics for word nepotism from twitter is evident that Twitter verse feels mostly negative about nepotism. The results for the word education were mostly positive. Karthika *et al.* [18] evaluated different models and the experiments were conducted to find the best classifier to analyze the reviews from shopping site amazon. Based on those reviews the product is classified as positive, negative, or neutral. Algorithms used here are random forest and SVMs. Random forest gave the best accuracy showing 84% while SVM showed 81% accuracy. Dataset contains reviews from 7 different products. Ramalingam *et al.* [19] tested numerous models and performed trials to discover the best classifier for identifying similar qualities among depressed persons and identifying them using various machine learning methods. The algorithms are intended to examine tweets for emotion detection as well as the identification of suicide ideation among social media users. logistic regression, SVM, and Random Forest are the algorithms employed here. The goal of these strategies is to leverage data accessible on Twitter and other social media to forecast people's mindsets by studying their numerous social media posts. When compared to logistic regression and random forest, SVM has the highest accuracy of 82.5%. Singh and Kumar [20] analyzed numerous models and conducted trials to determine the best method for predicting cardiac disease using various machine learning techniques. K-nearest neighbor, decision tree, linear regression, and SVM are the approaches. Jupyter notebook is employed as the simulation tool in this case. The dataset contains 14 variables such as sex, age, blood sugar, and so on. We discovered that the accuracy of each algorithm was 87%, 79%, 78%, and 83%, respectively. As a result, k-nearest neighbor (KNN) is the most precise. Sujath *et al.* [21] tested many models and performed tests to determine the optimal method for analyzing the impact of COVID-19 on the stock market. Using several algorithms, we attempt to determine which method provides the best accurate prediction of the impact of COVID-19 on the stock market. The algorithms are random forest, linear regression, and SVM. The dataset was discovered on Kaggle. We discovered that SVM had the highest accuracy of 82%.

Mujumdar and Vaidehi [22] analyzed different models and experiments were conducted to find the best algorithm to predict diabetes among patients. The dataset contains 800 records and 10 attributes. Algorithms used here are decision tree, logistic regression and KNN. Logistic regression shows the most accuracy with 96% compared to the other two which shows only 90% and 86% accuracy. Huq *et al.* [23] examined many models and performed tests to determine the best algorithm to predict the sentiment of a tweet on social media, i.e., whether it is good, negative, or neutral. It generally focuses on the tweet's wording and sentiment. KNN and SVM are the algorithms applied in this case. The dataset was obtained from the website Kaggle. According to the research, KNN is the most accurate, with an accuracy rate of 84%. Lassen *et al.* [24] examined many models and performed trials to determine the best algorithm to forecast iPhone sales based on tweets. The tweets are categorized as good, negative, or neutral. The dataset utilized here contains 400 million tweets from 2007-2010. Predictions are performed using linear regression and multiple regression models. Multiple regression has the smallest gap between anticipated and actual sales (5-10%), making it the most accurate. Dhir and Raj [25] examined many models and performed trials to determine the best algorithm for predicting movie performance. In this section, we analyze the internet movie database (IMDB) and estimate the IMDB score, as well as how it influences the movie collection. Logistic regression decision tree and random forest are the methods employed in this case. With 61% accuracy, random forest is the best. It demonstrates that social media likes, the number of voted users, and the length all have a significant impact on the IMDB score. Labib *et al.* [26] used machine learning methodologies to examine multiple models and perform tests to discover the optimal algorithm to analyze traffic incidents to predict the intensity of accidents. The algorithms employed in this case include naive Bayes, decision trees, KNN, and AdaBoost. It classifies the severity of incidents as deadly, serious, or minor harm. AdaBoost has the highest accuracy rate of 80%. It also revealed that accidents are more common at no-joint exits and T intersections. Wongkar and Angdresey [27] created this model for the 2019 presidential election using Python and the naive Bayes, SVM, and K-NN classifiers. Crawlers were employed to get tweets from Twitter, which were then tokenized to discover significant terms. They discovered that naive Bayes was more accurate, with an accuracy of 75-76%, after extensive study.

Gamon [28] proposed to perform sentiment analysis on even noisy data by the use of large feature vectors with feature reduction. As customer feedback are received at a very large volume, to be able to react to it quickly there has to be an efficient model to class the tweets into positive, negative, and neutral. They used NLPW in natural language processing for linguistic analysis. The accuracy at the end was 85.47%. Kusrini and Mashuri [29] proposed two classifiers SVM and naive Bayes and compared both classifiers to understand which classifier gives the best result. It first takes the dataset, uses tokenization to segregate the words, removes various slangs and then uses stemming using python to reduce the volume of data. The accuracy at the end was around 82-83%. Mandloi and Patel [30] proposed using three different classifiers namely SVM, naive Bayes and maximum entropy classification to understand the user's sentiment towards the following product, movie, and the people's alignment towards the political parties. To extract the data, they used three features namely unigram, bi-gram and n-gram features and the accuracy came out to be 85% for naive Bayes.

3. COMPARISON ANALYSIS

Table 1 (as seen in Appendix) shows the comparison of existing systems. To summarize all the existing works on sentiment analysis, we've gone through, we can divide it categorically into four types, which are document-level, sentence-level, phrase-level, and aspect-level. These existing papers tried to either tackle any one of the four types or even clubbed them, some tried to incorporate hashtags, some even tried to incorporate emoticons, some had language dependency, and some even had language independency. Some had greater accuracy but could tackle only one of the types, where some even had lesser accuracy but could incorporate a lot, some even tried building a complete corpus-based antonym dictionary.

Overall, we have a lot to dig in to use opinion mining to its fullest potential. What we will be doing in our model is, we will be taking the three best performing algorithms which are SVM, naive Bayes, and logistic regression to build a model which would allow enterprises to actually understand how well their products are performing, what shortcomings did customers feel, what could be better and many more. The proposed system will be much more efficient.

4. CONCLUSION AND FUTURE WORK

This article addresses a number of machines learning methods, including naive Bayes, SVM, logistic regression, and random forest. After extensive research, we discovered that SVM, naive Bayes, and logistic regression may be utilized to develop a model for our project that will provide a more accurate model than the present one, as demonstrated in the publications above. As we all know how analysis of twitter is being done to mine the opinions of users or customers in order to bring in potential customers or to enhance their products or services. Hence, it has become very important to constantly evolve and bring out even more accurate models. This work will help enterprises to draw out a basic idea on how the customers are reacting to the products which will then help them to make the product even better. This may help enterprises to leave behind the traditional methods of feedback forms which anyways is not very accurate.

People now have the option to organize the unrelenting rise of knowledge from interpersonal organizations. Because virtually all actual complicated concerns ranging from natural to mechanical in nature may be addressed via social media, its challenges should be heard. Rumor detection, evaluation repetition, patterns of online conversations resulting in riotous circumstances, and online shaming, all shift assumptions, allowing us to understand social pervasiveness in the form of preferences, shares, and retweets. Finding the right content and the right time to publish are two of the most important difficulties that need be addressed in interpersonal organizations before fully integrating into people's life. Indeed, even the detection of fraudulent remarks should be attended to at the tiniest level of social places like Twitter to avoid unnecessary badgering from spammers. Medical issues of genuine concern should be addressed in additional study so that they have a strong impact via web-based media clients. It would be appropriate at this point to prepare a tied up unified model that comprehends the assessments of the clientele when she/he is making remarks on social media.

APPENDIX

Table 1. Performance analysis comparison of existing systems (continue)

Name of the authors	Year of publication	Methodology	Algorithms used	Accuracy
Tan *et al.* [1]	2014	Foreground and Background LDA, Reason Candidate and Background LDA	Gibbs sampling, Parameter estimation, Average word entropy	69.70%
Xia *et al.* [2]	2015	Dual Training and Dual Prediction along with corpus-based antonym dictionary	Naive Bayes, SVM, logistic regression	85-87%
Hamroun *et al.* [3]	2015	OPEN NLP, WordNet, OWL ontology	CI patterns	72%
Li *et al.* [4]	2016	LibLinear Model and RNDN	N-gram, SSWE, WTFM	66.8
Ren and Wu [5]	2013	The social and topical contexts Factorization of Matrixes (ScTcMF)	Breadth-first search, user topic opinion labelling	60.35
Chen *et al.* [6]	2014	Use informal social medial data to provide insights	Naive Bayes, multilevel classification	61%
Bollrgala *et al.* [7]	2013	SentiWordNet lexica classifier, corpus based	Cross-domain sentiment classification	80%
Lin *et al.* [8]	2012	To identify sentiment and topic from text at the same time	Joint sentiment-topic (JST) model with weak supervision based on latent Dirichlet allocation (LDA, Reverse-JST).	71.20%
Wang *et al.* [9]	2011	To automatically create the overall sentiment polarity for a specific hashtag during a specified time period, which differs significantly from the typical sentence-level and document-level sentiment polarities.	SVM classifier	76%
Mudinas *et al.* [10]	2012	To classify polarity and detect sentiment strength	A hybrid strategy (lexicon-based + M/c learning) was used.	77%
Yu *et al.* [11]	2012	To Predict Sales Performance	Sentiment S-PLSA (an Autoregressive Sentiment and Quality Aware model)	73%
Jose and Chooralil [12]	2016	Three-way classifier unlike simple sentiment or non-sentiment analysis	n-gram feature, lexicon feature and part of speech feature	75%
Kouloumpis *et al.* [13]	2011	Three-way classifier unlike simple sentiment or non-sentiment analysis.	n-gram feature, lexicon feature and part of speech feature	75%
Park and Seo [14]	2018	Three AI assistants namely Siri by Apple, Cortana by Microsoft and Google Assistant by Google using sentiment analysis	VADER, Kruskal Wallis test and Mann-Whitney test	71%
Prakruthi *et al.* [15]	2018	Sentiment classification for the tweets using Histogram and Pie Chart.	Bag of Words algorithm	68%
Go *et al.* [16]	2009	Unigrams, Bi-grams, and parts of speech to use emoticons	Naive Bayes, SVM	81%
Trupthi *et al.* [17]	2017	Natural Language Processing – NLTK	Naive Bayes classification	74%

Table 1. Performance analysis comparison of existing systems

Name of the authors	Year of publication	Methodology	Algorithms used	Accuracy
Karthika *et al.* [18]	2019	Receiver Operating Characteristic (ROC) curve to evaluate classifier output	Random forest algorithm, SVM	84%
Ramalingam *et al.* [19]	2019	Machine learning and lexicon-based techniques to opinion mining, as well as assessment metrics	Logistic regression, SVM, and random forest	82.50%
Singh and Kumar [20]	2020	Machine learning algorithms' accuracy in predicting heart disease	k-nearest neighbor, decision tree, linear regression, and support vector machine	87%
Sujath *et al.* [21]	2020	forecasting model for COVID-19 pandemic	decision tree, logistic regression and KNN.	96%
Mujumdar and Vaidehi [22]	2019	best algorithm to predict diabetes among patients.	decision tree, logistic regression and KNN	96%
Huq *et al.* [23]	2017	To predict the sentiment of a tweet on social media	KNN and SVM classifiers	84%
Lassen *et al.* [24]	2014	predict iPhone sales using tweets based on iPhone	linear regression and multiple regression models	70%
Dhir and Raj [25]	2018	movie success prediction	logistic regression decision tree and random forest	61%
Labib *et al.* [26]	2019	determine the intensity of accidents	naïve bayes, decision trees, KNN and AdaBoost	80%
Wongkar and Angdresey [27]	2019	Data collection utilizing Python libraries, text processing, testing training data, and text categorization	Naive Bayes classifier, SVM classifier and K-NN classifier	76%
Gamon [28]	2004	Train linear SVMs to obtain high classification accuracy on difficult-to-classify data.	NLPW in natural language processing for linguistic analysis.	85%
Kusrini and Mashuri [29]	2019	Lexicon Based and Polarity Multiplication	SVM, naive Bayes	83%
Mandloi and Patel [30]	2020	Three features namely unigram, bi-gram and n-gram features	SVM, naive Bayes and Maximum Entropy	85%

REFERENCES

[1] S. Tan *et al.*, “Interpreting the public sentiment variations on Twitter,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1158–1170, May 2014, doi: 10.1109/TKDE.2013.116.

[2] R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi, and T. Li, “Dual sentiment analysis: considering two sides of one review,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2120–2133, Aug. 2015, doi: 10.1109/TKDE.2015.2407371.

[3] M. Hamroun, M. S. Gouider, and L. Ben Said, “Lexico semantic patterns for customer intentions analysis of microblogging,” in *2015 11th International Conference on Semantics, Knowledge and Grids (SKG)*, Aug. 2015, pp. 222–226., doi: 10.1109/SKG.2015.40.

[4] Q. Li, S. Shah, R. Fang, A. Nourbakhsh, and X. Liu, “Tweet sentiment analysis by incorporating sentiment-specific word embedding and weighted text features,” in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Oct. 2016, pp. 568–571., doi: 10.1109/WI.2016.0097.

[5] F. Ren and Y. Wu, “Predicting user-topic opinions in Twitter with social and topical context,” *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 412–424, Oct. 2013, doi: 10.1109/T-AFFC.2013.22.

[6] X. Chen, M. Vorvoreanu, and K. P. C. Madhavan, “Mining social media data for understanding students’ learning experiences,” *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, pp. 246–259, Jul. 2014, doi: 10.1109/TLT.2013.2296520.

[7] D. Bollegala, D. Weir, and J. Carroll, “Cross-domain sentiment classification using a sentiment sensitive thesaurus,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1719–1731, Aug. 2013, doi: 10.1109/TKDE.2012.103.

[8] C. Lin, Y. He, R. Everson, and S. Ruger, “Weakly supervised joint sentiment-topic detection from text,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012, doi: 10.1109/TKDE.2011.48.

[9] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, “Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach,” in *Proceedings International Conference on Information and Knowledge Management*, 2011, pp. 1031–1040., doi: 10.1145/2063576.2063726.

[10] A. Mudinas, D. Zhang, and M. Levene, “Combining lexicon and learning based approaches for concept-level sentiment analysis,” in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2012, pp. 1–8., doi: 10.1145/2346676.2346681.

[11] X. Yu, Y. Liu, X. Huang, and A. An, “Mining online reviews for predicting sales performance: a case study in the movie domain,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 720–734, Apr. 2012, doi: 10.1109/TKDE.2010.269.

[12] R. Jose and V. S. Chooralil, “Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach,” in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, Mar. 2016, pp. 64–67., doi: 10.1109/SAPIENCE.2016.7684133.

[13] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter sentiment analysis: the good the bad and the OMG!,” *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media*, vol. 5, no. 1, pp. 538–541, 2011

[14] C. W. Park and D. R. Seo, “Sentiment analysis of Twitter corpus related to artificial intelligence assistants,” in *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)*, Apr. 2018, pp. 495–498., doi: 10.1109/IEA.2018.8387151.

[15] V. Prakruthi, D. Sindhu, and D. S. Anupama Kumar, “Real time sentiment analysis of Twitter posts,” in *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, Dec. 2018, pp. 29–34., doi: 10.1109/CSITSS.2018.8768774.

[16] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” 2009.

[17] M. Trupthi, S. Pabboju, and G. Narasimha, “Sentiment analysis on Twitter using streaming API,” in *2017 IEEE 7th International Advance Computing Conference (IACC)*, Jan. 2017, pp. 915–919., doi: 10.1109/IACC.2017.0186.

[18] P. Karthika, R. Murugeswari, and R. Manoranjithem, “Sentiment analysis of social media network using random forest algorithm,” in *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Apr. 2019, pp. 1–5., doi: 10.1109/INCOS45849.2019.8951367.

[19] D. Ramalingam, V. Sharma, and P. Zar, “Standard multiple regression analysis model for cell survival/ death decision of JNK protein using HT-29 carcinoma cells,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 10, pp. 187–197, Aug. 2019, doi: 10.35940/ijitee.H7163.0881019.

[20] A. Singh and R. Kumar, “Heart disease prediction using machine learning algorithms,” in *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, Feb. 2020, pp. 452–457., doi: 10.1109/ICE348803.2020.9122958.

[21] R. Sujath, J. M. Chatterjee, and A. E. Hassanien, “A machine learning forecasting model for COVID-19 pandemic in India,” *Stoch. Environ. Res. Risk Assess.*, vol. 34, no. 7, pp. 959–972, Jul. 2020, doi: 10.1007/s00477-020-01827-8.

[22] A. Mujumdar and V. Vaidehi, “Diabetes prediction using machine learning algorithms,” *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[23] M. R. Huq, A. Ali, and A. Rahman, “Sentiment analysis on Twitter data using KNN and SVM,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, 2017, doi: 10.14569/ijacsa.2017.080603.

[24] N. B. Lassen, R. Madsen, and R. Vatrapu, “Predicting iPhone sales from iPhone tweets,” in *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, Sep. 2014, pp. 81–90., doi: 10.1109/EDOC.2014.20.

[25] R. Dhir and A. Raj, “Movie success prediction using machine learning algorithms and their comparison,” in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Dec. 2018, pp. 385–390., doi: 10.1109/ICSCCC.2018.8703320.

[26] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das, and F. Nawrine, “Road accident analysis and prediction of accident severity by using machine learning in Bangladesh,” in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, Jun. 2019, pp. 1–5., doi: 10.1109/ICSCC.2019.8843640.

[27] M. Wongkar and A. Angdresey, “Sentiment analysis using naive bayes algorithm of the data crawler: Twitter,” in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–5., doi: 10.1109/ICIC47613.2019.8985884.

[28] M. Gamon, “Sentiment classification on customer feedback data,” 2004., doi: 10.3115/1220355.1220476.

[29] Kusrini and M. Mashuri, “Sentiment analysis in Twitter using lexicon based and polarity multiplication,” in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Mar. 2019, pp. 365–368., doi: 10.1109/ICAIIT.2019.8834477.

[30] L. Mandloi and R. Patel, “Twitter sentiments analysis using machine learninig methods,” in *2020 International Conference for Emerging Technology (INCET)*, Jun. 2020, pp. 1–5., doi: 10.1109/INCET49848.2020.9154183.

BIOGRAPHIES OF AUTHORS

Vigneshwaran Pandi has obtained his Doctoral Degree in Anna University Chennai in 2016 and Master of Engineering under Anna University Chennai in June 2005. He is having 20 years of experience and specialization in Cybersecurity. Presently, He is working as Associate Professor at the SRM Institute of Science and Technology, Chennai. He has published more than 30 papers in various international journals and 10 in International Conferences. His area of interest includes Security, Routing, and Intelligent Data Analysis. He can be contacted at email: vigenesp@srmist.edu.in.

Prasath Nithiyanandam has obtained his Doctoral Degree from Anna University, Chennai in 2017 and Master of Technology from SASTRA University in 2009 and Undergraduate degree under Anna University Chennai in 2006. He is having 13+ years of experience in teaching and Industry. He has published 4 patents and published more than 15 research papers in refereed conferences and in journals. He is a member many professional societies. His research interests include MANET, Sensor Networks, IoT and Cloud, Cyber Physical System and Machine Learning. He can be contacted at email: prasathn@srmist.edu.in, prasath283@gmail.com.

Mrs. Sindhuja Manickavasagam M.Tech., Ph.D. Assistant Professor (Senior Grade) at Department of Information Technology, Rajalakshmi Engineering College from August 2006 onwards. She is having more than 15 years of experience in teaching. Presently she is acting as Assistant Professor (Senior Grade). She is pursuing Doctoral Program under Anna University in the field of Data Analytics. Her research interest includes artificial intelligence, machine learning, deep learning, bigdata analytics, and bioinformatics. She has published 14 international Journals including 3 from SCI and Scopus Indexed Journals. She has published one patent titled "Arm Band for Blood Testing" under the category of Design patent. She has visited Japan and Malaysia for presenting her research work in reputed International Conferences. Presently, she is working in the project of Cancer Analysis. She is a Certified Talend DI developer by Virtusa, IBM certified DB2 and Tivoli developer. She acted as resource person for various workshops held in other Engineering Colleges and University on her research topics. She can be contacted at email: sindhuja.m@rajalakshmi.edu.in.

Dr. Islabudeen Mohamed Meerasha received his B.E. Degree in Computer Science and Engineering from Madurai Kamaraj University, Madurai, India in 2001 and M.E. Degree in Computer Science and Engineering and Ph.D. Degree in Information and Communication Engineering from Anna University, Chennai, India in 2008 and 2021, respectively. He is currently working as an Associate Professor in Department of Computer Science Engineering, School of Engineering, Presidency University, Bengaluru, India. He is having more than 20 years of academic experience. His current research interests include Cryptography and Network Security, Blockchain, Wireless Networks and Data Mining. He is a life member of Computer Society of India (CSI) and Indian Society for Technical Education (ISTE). He has served as an active reviewer and chair for many reputed international conferences and journals including IEEE and Springer. He has more than 20 publications in reputed international conferences and refereed journals. He can be contacted at email: islabudeen@gmail.com.

Dr. Ragaventhiran Jaganathan received his B.E. and M.E. degree in Computer Science and Engineering and Ph.D. in Information and Communication Engineering from Madurai Kamaraj University and Anna University, Tamil Nadu, India respectively. He is currently working as an Associate Professor in School of Computing and Information Technology in REVA University, Bengaluru India. His research interest includes data mining, big data, data structure and cloud computing. He is a life member in computer society of India (CSI) and Institution of Engineers (IEI) India. He has reviewed and chaired various national and international conferences including IEEE conferences. He is also a reviewer for refereed journals. He can be contacted at email: jragaventhiran@gmail.com.

Dr. Muthu Kumar Balasubramania Professor, School of Computing and Information Technology, REVA University, Bengaluru received his B.E., (CSE) degree from Anna University, Chennai in the year 2005, M.Tech., (CSE) (Gold Medalist) received from Dr. MGR. University, Chennai in the year 2007 and Doctoral degree from St. Peter's University, Chennai in the year 2013. He is having more than 16 years of teaching experience in reputed engineering colleges. He has published more than 40 peer reviewed International Journals, 50 International/National Conference and attended more than 150 Workshops/FDPs/Seminars etc., He organized many events like Conference/FDPs/Workshops/Seminars/Guest Lecture. He has published more than 10 patents in various fields like Wireless Sensor Networking, Image Processing, Optimization Techniques and IoT. He received nearly 5.67 Lakhs funding from various agencies like AICTE, ATAL and IEI. He has written 2 books from reputed publishers. He received Best Researcher Award in the year 2021 and Innovative Research and Dedicated Professor Award in Computer Science and Engineering in the year 2018. He has professional membership on ISTE, CSI, IEI, IACSIT, IAENG, CSTA, and SIAM. He has invited as guest lecture, chairperson, examiner, and reviewer/editorial board member in various institutions, journals, and conferences. He is a recognized supervisor in Anna University, Chennai and currently guiding 4 research scholars. His areas of interest are image processing, wireless networks, IOT and computing techniques. He can be contacted at email: muthu122@gmail.com.