Predicting and Mitigating the Effect of Skewness on Credibility Assessment of Social Media Content Using Machine Learning: A Twitter Case Study

Shifaa Basharat, Saduf Afzal, Alwi M Bamhdi, Shozab Khurshid*, and Manzoor Chachoo

***Abstract*—Many strategies have been put forward to assess the credibility of online social media content, however, none of them focuses on the issue of accuracy paradox which mostly occurs in highly skewed datasets, a case that usually arises in real-life situations. The purpose of this paper is to explore the use of various machine learning models including Gaussian Naïve Bayes, Latent Dirichlet Allocation (LDA), Linear Regression, Logistic Regression, and Support Vector Machine (SVM) for identifying the credibility of tweets. This includes proposing a new algorithm where the generative properties of Gaussian naïve Bayes are integrated with the discriminative properties of logistic regression and the author evaluates its performance in terms of accuracy and prediction power of determining tweet credibility. The Machine Learning Models used in this study, implemented on the Twitter datasets extracted from various real-world events are compared based on their accuracy and predictive power, in determining the credibility of tweets, to identify various accuracy paradox cases. The proposed algorithm is then used for the credibility inference of tweets and the reduction in the number of accuracy paradox cases is monitored. An extensive experimental study is performed to evaluate the performance of the proposed model on Twitter datasets with varied degrees of skewness. Our proposed model achieved accuracy and predictive power of 97% and 94% for a balanced dataset and 99% and 93% for an imbalanced dataset with 99% skewness.**

***Index Terms*—Gaussian Gradient Descent (GGD), Gaussian Naïve Bayes (GNB), intra-skewness, inter-skewness, predictive index**

I. Introduction

Social media is quickly arising as a new and popular form of media. The last ten years have seen a huge rise in social media influence, not only in terms of communication but also in terms of business, travel, tourism, relationships, food, etc. According to studies conducted by the Pew research center, social media users have grown by more than 60% from 5% in 2005 to 69% in 2016 [1]. In another study, Pew Research Center has identified the Internet as the most important resource for the news for people under the age of 30 in the US and the second most important overall after television [2]. While social media is mostly used for everyday chatter, it is also used to share news and other important information [3, 4]. Now more than ever, people turn to social media as their source of news [5–7]; this is especially true for breaking-news situations, where people crave rapid updates on developing events in real-time. As Kwak and Lee *et al.* have shown, over 85% of all trending topics on Twitter are news [6]. Moreover, the ubiquity, accessibility, speed, and ease of use of social media have made them invaluable sources of first-hand information. However, a huge amount of misinformation including rumors, fake information, and fake identity engulfs online social media [8, 9].

Ever since Twitter and other forms of social media have replaced traditional sources of news, a lot of research has been carried out in terms of distinguishing fake and genuine information. The use of machine learning to determine the reliability of online news shared has received major attention from the research community as fake news and rumors are also propagated with genuine news [10] which leads to chaos among the population [11].

Although several machine learning techniques have been implemented for credibility evaluation, including Castillo and Mendoza *et al.* using the J48 decision tree for assessing the credibility of a given set of tweets [12], Gupta and Kumaraguru *et al.* [13] using Support Vector Machine (SVM) rank for credibility ranking of tweets during high impact events, Saikaew and Noyunsan [14] using SVM for determining the credibility of Facebook information among several others, but this does not suffice as the information explosion leads to a huge imbalance in datasets. The main motivation behind this research is to identify the reliability of tweets such that our model provides accurate results irrespective of the skewness prevalent in various social media datasets.

The work presented in this paper proposes to integrate the generative and discriminative properties as typified by Gaussian Naïve Bayes and Logistic regression into a single algorithm. Gaussian Naïve Bayes has been chosen because it supports continuous data which is the case for our Twitter datasets, whereby our feature vector gets reduced to continuous data after normalization. However, Multinomial Naïve Bayes and Bernoulli Naïve Bayes are well suited to discrete and Boolean data respectively. The integrated algorithm is then used to overcome the problem of accuracy paradox identified by determining the predictive index of the algorithm.

The paper is structured as follows. Section II presents the recent work done in evaluating the credibility of online social media content. Section III discusses the credibility inference framework used in this paper. The proposed integrated algorithm is described in Section IV. The experimental results followed by the conclusion are discussed in Section V and Section VI of this paper respectively.

Manuscript received June 27, 2022; revised August 12, 2022; accepted March 9, 2023.

Shifaa Basharat, Saduf Afzal, Shozab Khurshid, and Manzoor Chachoo are with Department of Computer Science, University of Kashmir, India.

Alwi Bamhdi is with the Computing College in AlQufudah, Umm Al-Qura University, Saudi Arabia.

*Correspondence: shozabkhurshid@gmail.com (S.K.)

II. Literature Review

A lot of research has been conducted for extracting and analyzing trustworthy content or analyzing its credibility from different social media platforms. Kang discussed the practical aspects of measuring blog credibility and validated a 14-item measure for calculating it [15]. Rubin and Liddy [16] combine credibility judgments of blog readers with NLP-based analysis of blogs. They used four profile factors which included the blogger's expertise and amount of offline identity disclosure, the blogger's trustworthiness, information quality, and appeals of personal nature for blog credibility assessment. Soiraya and Mingkhwan *et al.* [17] used the concept of text analysis to assess the trust of an E-commerce website. They used baseline and EC-word approaches for constructing feature sets for their classification algorithms and the best results were obtained on the application of the sequential Minimal Optimization algorithm together with the EC-word feature set. The algorithm achieved an accuracy of 83.5%.

Abbasi and Liu proposed a CredRank algorithm that uses the online behavior of social media users to measure their credibility [18]. Their method ranked Online Social Media users based on their credibility by assigning lower credibility scores to users involved in coordinated behavior. Barbier and Liu proposed a method for credibility evaluation by finding provenance paths leading to sources of the information that are being evaluated for their credibility [19]. Jamali and Ester in their work have proposed a random walk model based on the combination of trust-based and the collaborative filtering approach for a recommendation. They exploited the concept of trust to help online users collect reliable information in applications such as high-quality reviews detection and product recommendations [20]. Guha and Kumar *et al.* studied the problem of propagation of trust and distrust among Epinions users, who rate each other by assigning positive (trust) and negative (distrust) based on their experiences with different users [21]. They used the ratings of each user to rank them and thereby rank the content generated by them. Agichtein and Castillo *et al.* used community information to identify high-quality content in the question and answering portal (Yahoo! Answers 5) [22]. They used features generated from answers, questions, votes, and users' information and relationship to build a model to measure quality of the content.

Kuter and Golbeck showed the importance of user's trust in the source for the aggregation, filtering, and ordering of information [23]. Their work described a new trust inference algorithm based on the probabilistic sampling technique to estimate confidence in the trust information from some designated sources.

Castillo and Mendoza *et al.* determined the credibility of news propagated via Twitter using machine learning [12]. They extracted relevant discussion topics by studying bursts of activity using Twitter Monitor [24]. After using human accessors for classifying the topics as newsworthy/informal conversations, tweets belonging to the newsworthy class were labeled for credibility using Mechanical Turk evaluators. Later they trained several supervised classifiers to predict the credibility levels of different tweets using Mechanical Turk labeling out of which the J48 decision tree classifier produced the best results with an accuracy of 86%. They made some interesting observations, such as tweets that do not include URLs tend to be related to non-credible news; tweets that include negative sentiment terms are related to credible news.

Sahana and Pias *et al.* use a J48 classifier for classifying the tweets into rumor and non-rumor classes [25]. The authors used rumor tweets posted during the London riots and verified by the Guardian whereas the non-rumor tweets were collected using Twitter's Streaming API. The J48 model presented in this paper achieved a classification accuracy of ~87%. In addition to rumor-based tweet classification, the authors also conclude that tweet-based features play a relatively more important role than user-based features in rumor detection.

Another work introduced by Qazvinian and Rosengren *et al.* discussed the effectiveness of three different categories of Twitter features including content-based, network-based, and microblog-specific memes such as hashtags and URLs for correctly identifying rumors [26]. The experiments that were performed on more than 10K manually annotated tweets collected using Twitters Streaming API, achieved a mean average precision of more than 0.95. Hamidian and Diab presented a Rumor Detection and Classification (RDC) technique that not only detects a rumor but also classifies it [27]. The authors on comparing two different RDC techniques namely single-step RDC (SRDC) and two-step RDC (TRDC) conclude that TRDC outperforms SRDC by achieving an F-measure of 82.9% compared to only 74%.

Ajao and Bhowmik *et al*. used Hybrid Convolutional Neural Networks (CNN) and Long Short Term Recurrent Neural Network models to distinguish fake news from genuine one with an accuracy of 82% [28]. Naderam and Namjoo *et al.* used the concept of fuzzification of trust values into two, three, and five class categories and used three basic machine learning algorithms namely SVM, Decision Tree, and K-Nearest Neighbors (KNN) for trust classification. Their proposed approach achieved an accuracy of ~91% [29]. Xu and Yuan *et al.* used the concept of Hidden Bayesian Model to identify trustworthy people on Twitter and Facebook. Their proposed approach achieved an accuracy of 95% on the Twitter dataset and 87% on the Facebook dataset [30]. Verma and Agrawal suggested a combined approach for fake news detection. Their analysis showed that the combined approach Propagation, Pattern, Credibility, and Comprehensive approach (P2C2) gives improved results in the detection of fake news [31]. Kardas and Bayar *et al.* proposed the advantages of preprocessing before applying machine learning for the detection of spam tweets. The preprocessing approach was evaluated with four different Machine learning models namely Naïve Bayes Classifier, Neural Networks, Logistic regression, and SVM. The proposed approach together with SVM achieved an accuracy of ~93% [32].

III. Credibility Inference Framework

The credibility inference framework for Twitter consists of the data collection phase, data analysis phase, and machine learning phase as discussed in Fig. 1 [33]:

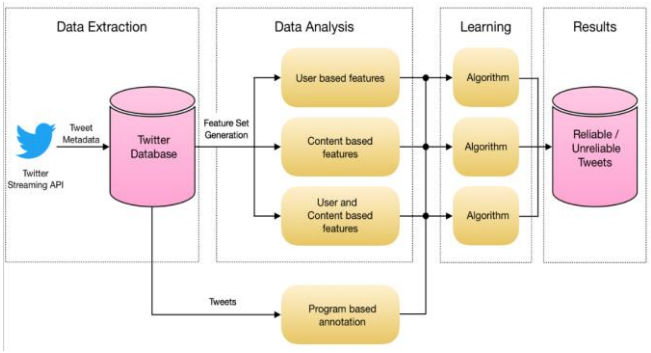

Fig. 1. Credibility inference framework for Twitter.

A. Extraction of Twitter Data

The process of data collection was based on collecting tweets from major events that took place between 2015–2017. Subject-specific tweets were collected by using hashtags about a particular event (Table I) as keywords for the search process in Twitter Streaming API and Chorus tools, which is software that allows us to retrieve tweets based on keyword matches and user timelines.

TABLE I: EVENTS USED FOR THE EXPERIMENT

Event Type	Trending Topics	Event Description
Paris Attacks	#ParisAttacks, #PrayForParis, #ISIS, #PrayForPeace, #Prayers4Paris	Series of terrorist attacks on November 2015 in Paris that caused 137 deaths.
Hamas Attacks	#Palestine, #Gaza, #Israel, #IslamicState	Attacks carried out by Hamas on Israel in the year 2017.
News	#News, #CelebrityNews, #tecnews, #digitalmarketing	All the trending topics in the field of sports, glamour, technology, and business.
India versus West Indies	#IndvsWI, #ViratKohli, #GoIndiaGo, #Dhoni, #Indians	News related to Microsoft from the period April 2017–June 2017.
Indian Elections 2014	#Modi, #IndianPM, #GST, #bjplies, #BJPEmpowers	News related to the 2014 Lok-Sabha elections.
Harrisburg Attacks	#terrorism, #centralIPA, #Harrisburg, #police	A terror attack on the Harrisburg police on December 2017.
Boko Haram Attacks	#Bokoharam, #Nigeria, #Attack	Attacks in Nigeria claimed 400 lives between April 2017- September 2017.
US Presidential Elections 2016	#PresidentObama, #Trump, #Trump, #US, #Obama	News related to the US
Terror Attacks	#IsraeliState, #terrorist, #hostage, #shooting, #Attack, #terrorattack	Terror attacks that took place in the year 2017 in any part of the world.
Microsoft	#startup, #bigdata, #cloudcomputing, #microsoft, #bitcoin	News related to Microsoft from the period April 2017–June 2017.

Twitters Streaming Application Programming Interface (API) delivers data to clients in web real-time. This data extraction step involved creating an application on Twitter which is followed by using our Open Authorization (OAuth) credentials for data extraction.

The output of the data extraction step is a JavaScript Object Notation (JSON) file that consists of the entire tweet metadata. The unofficial Java library, Twitter 4J was used to integrate the data extraction Java application with the Twitter service. The complete data extraction procedure has been depicted in Fig. 2.

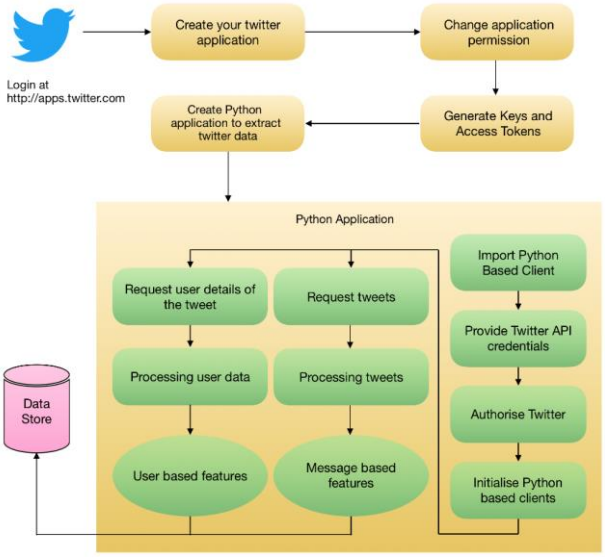

Fig. 2. Twitter data extraction.

We collected tweets related to different real-world events such as the US Presidential elections in 2016, Paris Attacks in 2015, Hamas Attacks, and so on. We collected data from 10 different events covering the fields of sports, technology, politics, and violence (Table I).

In addition to the details of the events, the total number of tweets extracted for each event and the number of tweets belonging to credible and non-credible classes have been depicted graphically in Fig. 3:

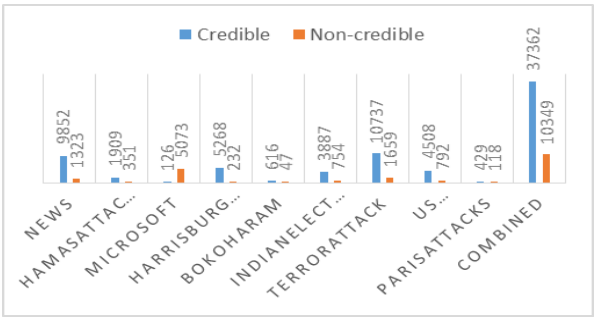

Fig. 3. Twitter data statistics.

B. Data Analysis and Interpretation

The JSON file output from the data extraction step goes through a series of Java and Python application programs (Fig. 4) thereby giving us the various user-based and tweet-based features along with the annotated Twitter dataset to be used for training various machine learning models.

- TPJava: It is a text-processing Java application that extracts the Twitter user's screen name who has tweeted only in the English language from the JSON file. It also extracts tweets from the JSON file which is used as input to the JTweet.
- PExtract: This is a Python-based application that extracts a particular Twitter user's profile details also known as user-based features such as followers_count, friends_count, status_count, listed_count, etc.
- JAnnotate: A Java based application program that uses various user-based features to annotate the Twitter dataset for use in training the machine learning models.
- JTweet: A Java-based application program that extracts various tweet-based features such as the number of

characters, number of words, number of URLs, and number of swear words in the tweet.

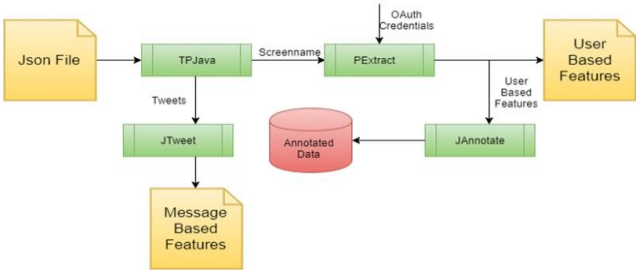

Fig. 4. Twitter data analysis and interpretation.

The feature selection process involved using a perceived credibility impact of each feature as presented by Morris et.al. 2012 in his paper. They had given a mean rating for tweet features' perceived credibility impact on a 5-point scale and attention typically allotted on a 3-point scale. We sorted tweet features based on their impact on reliability and came up with the top 11 features from the sorted list, which were used for the credibility assessment of tweets in our research.

C. Program Based Annotation

As our trust inference framework for Twitter is based on supervised machine learning models, the framework requires a trained set of tweets whose trust label is already known. To prepare a properly annotated Twitter dataset we use a Java based application program JAnnotate. JAnnotate uses the top 11 Twitter features which have a very high-reliability impact on the trust/reliability value of a tweet as presented in the literature [12, 34–37].

The following Algorithm 1 describes the steps involved in preparing a labeled Twitter dataset for training. We start with the Extract Feature Impact Value function which returns the impact value of the input feature vector that constitutes the Twitter dataset. SortAsc function sorts the Feature Impact vector in ascending order such that the top 11 elements of the sorted Feature Impact vector can be used for annotation. Extract Tweet Features returns a list of user-based features for the i^{th} tweet in the dataset. Finally, the Calculate Reliability function uses the feature vector F along with their impact value to calculate the reliability label for each tweet thereby providing us with the labeled training dataset for use in machine learning models.

Algorithm 1: Preparing a Labeled Twitter Dataset for Training

```
JAnnotate (Tweet [1..n])
  FeatureImpact < - ExtractFeatureImpactValue (Features)
  FeatureImpact' < - SortAsc (FeatureImpact)
  for i < - 1 to n do
      Fi < - ExtractTweetFeatures (Tweet[i])
  CredibilityScorei <- CalculateCredibility (FeatureImpact',
                                             F[i])
  if CredibilityScorei > 0
     TweetLabeli < - Credible
  else
      TweetLabeli < - Non-credible
  end if
  end for
  return TweetLabel[1..n]
```

D. Learning

In addition to the proposed integrated model, five popular machine learning models including Naïve Bayes, Linear Regression, Logistic Regression, Linear Discriminant Analysis, and Sequential Minimal Optimization, were used for the credibility inference of tweets. A total of 19 Twitter features (Table II) were used for training each model.

TABLE II: TWITTER FEATURES FOR RELIABILITY CHECK

Attribute Name	Attribute Type	Attribute Description
Friends_count	UserBasedFeature	No. of users this account has subscribed to
Listed_count	UserBasedFeature	No. of public lists this user is a member of
Followers count	UserBasedFeature	No. of users who have subscribed to this account
Status_Count	UserBasedFeature	No. of tweets the user has posted on his timeline
Registration_age	UserBasedFeature	Time since the user is on Twitter
Verified	UserBasedFeature	Whether the user has a verified account or not
Retweets_count	UserBasedFeature	No. of times the users status has been shared
Favorites_count	UserBasedFeature	Indicates the no. of times a particular tweet has been liked
Default_profile	UserBasedFeature	Whether the user has altered the theme and background of his Twitter profile or nor
Default_profile_ image	UserBasedFeature	Whether the user has uploaded the profile image or the default human silhouette has been used
Protected	UserBasedFeature	Whether the Twitter users profile is protected or not
No_of_ characters	MessageBased Feature	Total number of characters in the tweet
No_of_words	MessageBased Feature	Total number of words in a tweet
Pronoun_count	MessageBased Feature	Number of pronouns in a tweet
Retweet	MessageBased Feature	Whether the tweet is a retweet or not
URL_count	MessageBased Feature	Number of URLs in the tweet
Swearwords	MessageBased Feature	Number of swearwords used in the tweet
Special Symbols	MessageBased Feature	The number of special symbols a tweet has like @, #, $,?
Sentiment score	MessageBased Feature	Whether the tweet carries a negative, positive, or neutral sentiment

IV. PROPOSED MODEL

The proposed integrated model by Fazili and Ahmad [38], namely the Gaussian Gradient Descent Model (Fig. 5) is a web page-dependent approach that combines the generative and discriminative properties as typified by Gaussian Naïve Bayes and Logistic regression into a single algorithm. The Generative Module works by learning a model of joint probability, P (features, class), of Twitter features and the reliability label of each tweet, and the discriminative module works by modeling the conditional probability distribution, P (class | features).

The Generative Module accomplishes its task through a series of sub-modules which include:

A. Class-Based Segmentation Module

This module divides the entire Twitter dataset (only the training data) into two subsets of credible tweets and non-credible tweets respectively which it uses to obtain the prior probability of each class.

$$Prior_{credible} = \frac{Total\ number\ of\ credible\ tweets}{Total\ number\ of\ tweets} \quad (1)$$

where, $Prior_{credible}$ is the prior probability of credible tweets.

Similarly, we can obtain the prior probability of non-credible tweets.

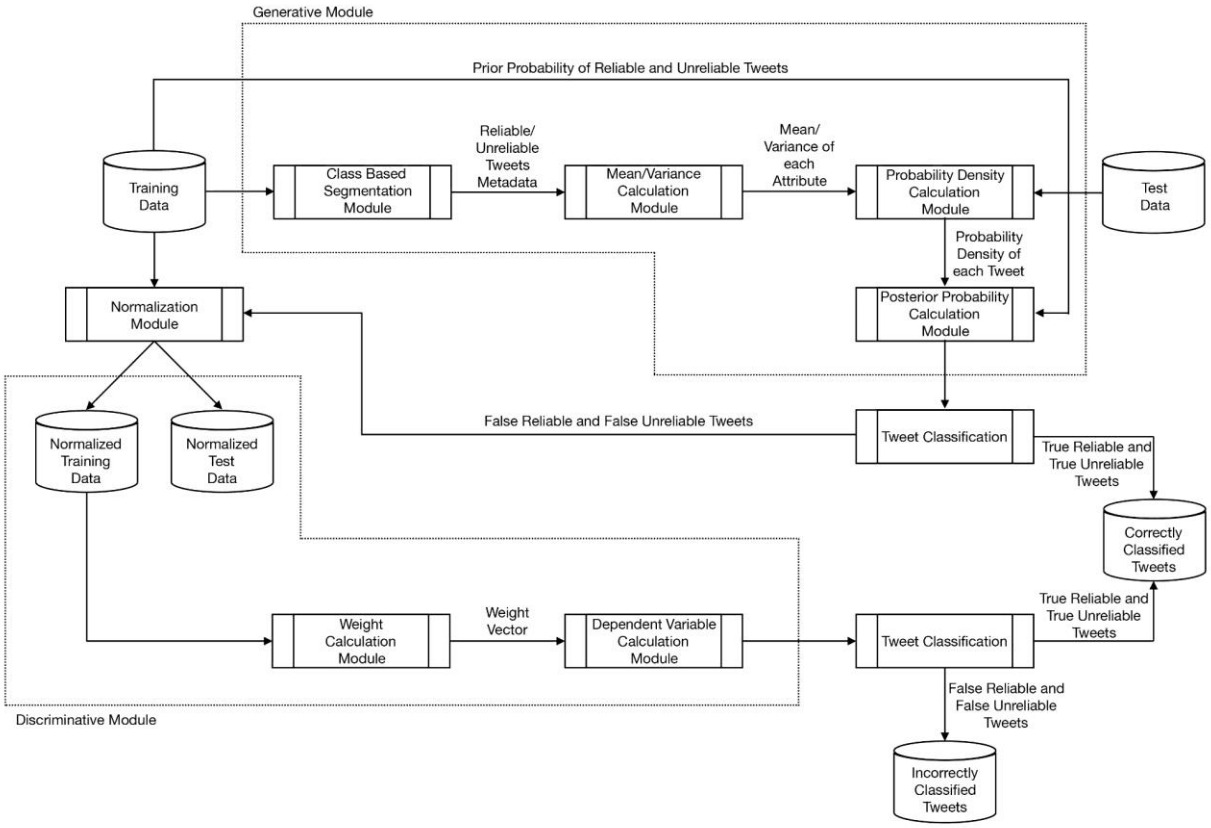

Fig. 5. Gaussian gradient descent model.

B. Mean and Variance Calculation Module

It takes the segmented datasets from the Class-Based Segmentation module as input to calculate the variance of both the credible and non-credible tweets for each feature of the tweet and the Twitter user.

$$cred\_variance_i = (cred\_tweet_i - 1/n \sum_{i=1}^{n} cred\_tweet_i) \quad (2)$$

$$noncred\_variance_i = (noncred\_tweet_i - 1/n \sum_{i=1}^{n} noncred\_tweet_i) \quad (3)$$

In Eqs. (2) and (3), $cred\_var_i$ and $noncred\_var_i$ represent the variance of the feature of the tweet or the Twitter user (such as the sentiment of a tweet or the number of friends a Twitter user has and so on) in the credible and non-credible tweet dataset respectively and $cred\_tweet_i$ represents the i^{th} feature of the tweet from the credible tweet dataset for which the data spread is to be calculated.

This variance of the tweets is used to calculate the z-score distance between the tweets and each class mean (which includes the credible and non-credible class) given by Raizada and Lee [39]:

$$z_{credible} = \frac{-(tweet\_test_i - 1/n \sum_{i=1}^{n} cred\_tweet_i)^2}{2 \cdot cred\_variance} \quad (4)$$

$$z_{noncredible} = \frac{-(tweet\_test_i - 1/n \sum_{i=1}^{n} noncred\_tweet_i)^2}{2 \cdot noncred\_variance} \quad (5)$$

C. Probability Density Calculation Module

After obtaining the measure of data spread for an individual tweet and the Twitter user features, the probability density module gives us the probability distribution of a particular point in the test dataset as follows [40]:

$$P(tweet\_test_i \mid C_{credible}) = \frac{1}{\sqrt{2 \cdot \pi \cdot cred\_variance_i^2}} \cdot e^{\frac{-(tweet\_test_i - 1/n \sum_{i=1}^{n} cred\_tweet_i)^2}{2 cred\_variance_i^2}} \quad (6)$$

D. Posterior Probability Calculation Module

Following the Probability Density module, the last step of the Generative Module includes the calculation of the posterior probability of each tweet which is obtained as the product of module 1 and module 3 as follows [40]:

$$posterior_{cred} = Prior_{credible} \cdot \prod_{i=1}^{k} P(tweet\_test_i | C_{credible}) \quad (7)$$

$$posterior_{noncred} = Prior_{noncred} \cdot \prod_{i=1}^{k} P(tweet\_test_i | C_{noncred}) \quad (8)$$

In Eqs. (7) an (8), $posterior_{rel}$ and $posterior_{unrel}$ gives us the probability of a particular tweet belonging to credible and non-credible classes respectively and then the tweet is placed in the class with higher probability. All the unclassified/wrongly classified tweets are forwarded to the Discriminative module where they go through a series of sub-modules as discussed below.

E. Normalization Module

This module is used to normalize the Twitter training dataset (same as that of the Generative Module) and the test dataset (which includes the unclassified/wrongly classified tweets from the Generative module). Each data instance can be normalized as:

$$tweet\_norm_i = \frac{tweet_i - mean(tweet_i)}{stddev(tweet_i)} \quad (9)$$

Here $tweet_i$, $mean(tweet_i)$ and $stddev(tweet_i)$ represent the i^{th} feature of a particular tweet, mean and standard deviation of the i^{th} feature of a particular tweet respectively and $tweet\_norm_i$ represents the normalized value of the i^{th} feature of a tweet.

F. Weight Calculation Module

This module is used to calculate the unknown weight vector (each of whose elements corresponds to a particular tweet feature) for use in the sigmoid function (discussed in the next section) which is applied to predict the reliability label of each tweet in the test dataset. After initializing the weight vector to random values between 0 and 1, each of its elements can be updated by the following Eq. (10) [41]:

$$wt_j = wt_j + step \times (train\_label_i - predicted\_label(tweet\_train_i)) \times tweet\_train_j^i \quad (10)$$

Here wt represents the weight vector, $step$ is a constant (=0.001), i represents the i^{th} instance of the training data, and j represents the j^{th} feature of the dataset. Also, the predicted label of each tweet can be obtained by the following Eq. (11) [41]:

$$predicted\_label(tweet\_train_i) = \frac{1}{1+e^{-wt^T \cdot tweet\_train_i}} \quad (11)$$

This weight update process continues until the predicted reliability label of the tweet is at a distance of epsilon (=0.001) from the actual reliability label.

G. Credibility Label Calculation Module

The weight vector from the previous module can be used to obtain the reliability label of the tweet by calculating the sigmoid value for each tweet using a sigmoid function as follows in Eq. (12) [41]:

$$tweet\_test\_value = \frac{1}{1+e^{-wt^T \cdot tweet\_test_i}} \quad (12)$$

The $tweet\_test\_value$ can be used for the classification of the tweet as shown in Eq. (13) [41]:

$$credibility\_label = \begin{cases} credible, & if\ tweet\_test\_value \geq 0.5 \\ non-credible, & if\ tweet\_test\_value > 0.5 \end{cases} \quad (13)$$

V. Experimental Results

The performance evaluation of the Gaussian Gradient Descent model described in Section IV is presented here. Ten different datasets from multiple Twitter events (Table I) with varying degrees of skewness (Table III) have been used to compare the performance of our proposed model with the five most popular machine learning models which include Naïve Bayes, Linear Regression, Logistic Regression, Linear Discriminant Analysis, and Sequential Minimal Optimization. The performance is evaluated using accuracy, precision, recall, F_1, and predictive index (P_i) [38] to identify pure accuracy paradox cases and complete paradox cases whereby each evaluation metric other than predictive index gives misleading results. The predictive Index is computed as follows in Eq. (14):

$$P_i = 1 - |S_t - S_p| \quad (14)$$

where $S_t = \frac{tc}{tc + fnc}$ and $S_p = \frac{tnc}{tnc + fc}$ refer to the sensitivity and specificity of a machine learning model [42]; tc refers to the true credible tweets; tnc refers to true non-credible tweets; fc refers to the number of tweets identified as credible but are non-credible; fnc refers to the number of tweets identified as non-credible, but are credible.

TABLE III: Skewness of Different Datasets

Event Type	Intra-skewness(1)		Inter-Skewness(2)
	Training Data	Test Data	
Paris Attacks	73%	73%	50%
Hamas Attacks	91%	76%	58%
News	89%	87%	50%
India versus West Indies	70%	70%	50%
Indian Elections 2014	83%	85%	65%
Harrisburg Attacks	98%	50%	95%
Boko Haram Attacks	92%	95%	70%
US Presidential Elections 2016	86%	84%	66%
Terror Attacks	89%	81%	73%
Microsoft	50%	99.7%	96%

(1) Intra-Skewness refers to skewness within the datasets, in terms of the number of observations belonging to each class either for the training dataset or test dataset.

(2) Inter-Skewness refers to skewness between datasets i.e. the number of observations belonging to the training dataset (both credible and non-credible) in comparison to those belonging to the test dataset (both credible and non-credible).

Predictive index is used to indicate the quality of a binary classification model, such that a P_i value of 1 indicates a perfect classification (i.e., all the instances of the test data have been correctly classified into their respective classes) whereas a P_i value of 0 together with ~80% accuracy value indicates the random prediction of classes due to skewness in the datasets.

TABLE IV: ALGORITHM PERFORMANCE ON BOKOHARAM DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.0476	0	0	0	0
Gaussian Naive Bayes	0.8381	0.9826	0.845	0.9086	0.855
Linear Regression	0.9286	0.9512	0.975	0.9629	0.025
Logistic Regression	0.9286	0.9947	0.93	0.9612	0.97
SMO	0.6381	1	0.62	0.7654	0.62
GGD	0.9809	0.9852	0.995	0.99	0.705

TABLE V: ALGORITHM PERFORMANCE ON HAMAS DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.7573	0.7589	0.9972	0.8619	0.0027
Gaussian Naive Bayes	0.6489	0.9712	0.5542	0.7057	0.6061
Linear Regression	0.7583	0.7591	0.9986	0.8625	0.0014
Logistic Regression	0.8406	0.86	0.9437	0.8999	0.5714
SMO	0.4167	0.9668	0.24	0.3846	0.266
GGD	0.9917	0.9891	1	0.9945	0.9654

TABLE VI: ALGORITHM PERFORMANCE ON NEWS DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.8745	0.8753	0.9985	0.9329	0.0199
Gaussian Naive Bayes	0.6133	0.9842	0.5663	0.7189	0.6289
Linear Regression	0.8751	0.8749	1	0.9332	0.128
Logistic Regression	0.8646	0.9437	0.8986	0.9206	0.7316
SMO	0.4083	0.9683	0.3335	0.4961	0.4089
GGD	0.9924	0.9914	1	0.9957	0.9402

TABLE VII: ALGORITHM PERFORMANCE ON HARRISBURG DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.5	0.5	1	0.6667	0
Gaussian Naive Bayes	0.9133	0.9079	0.92	0.9139	0.9867
Linear Regression	0.5	0.5	1	0.6667	0
Logistic Regression	0.5	0.5	1	0.6667	0
SMO	0.78	1	0.56	0.7179	0.56
GGD	0.97	0.9434	1	0.9709	0.94

TABLE VIII: ALGORITHM PERFORMANCE ON THE INDIAN ELECTION DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.8556	0.8635	0.9849	0.9202	0.1647
Gaussian Naive Bayes	0.7178	0.9915	0.6719	0.801	0.7034
Linear Regression	0.8543	0.8624	0.9848	0.9196	0.1569
Logistic Regression	0.7672	0.9754	0.7433	0.8437	0.8457
SMO	0.4452	1	0.4225	0.594	0.7136
GGD	0.9799	0.9949	0.9812	0.988	0.9912

From the results depicted in Tables IV–XIII, we observe that our proposed GGD model unlike other models performs consistently well irrespective of the skewness factor ranging between 50% and ~99% for 10 different datasets used in this paper. The accuracy and predictive power achieved for the proposed model came out to be 97%, 94% for balanced datasets (50% skewness), and 99%, 93% for highly imbalanced datasets (~99% skewness). However, the combination of accuracy, and predictive power for other machine learning models ranges between (91%, 98%) to (93%, 25%), (78%, 56%) to (82%, 73.8%), (50%, 0%) to (99%, 0%), (50%, 0%) to (54.5% to 60.6%) and (50%, 0%) to (98%, 7%) for Gaussian Naïve Bayes, SMO, LDA, Linear Regression and Logistic regression respectively.

TABLE IX: ALGORITHM PERFORMANCE ON TERROR ATTACK DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.8154	0.8166	0.9942	0.8967	0.0787
Gaussian Naive Bayes	0.6016	0.9832	0.5144	0.6755	0.5508
Linear Regression	0.8048	0.8057	0.9985	0.8918	0.0015
Logistic Regression	0.8268	0.8847	0.9028	0.8937	0.6086
SMO	0.2256	1	0.039	0.0752	0.039
GGD	0.9929	0.9913	1	0.9956	0.9636

TABLE X: ALGORITHM PERFORMANCE ON U.S. PRESIDENTIAL ELECTIONS DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.8433	0.8564	0.9768	0.9126	0.1733
Gaussian Naive Bayes	0.71	0.9833	0.6651	0.7935	0.7233
Linear Regression	0.8522	0.8552	0.9914	0.9183	0.1422
Logistic Regression	0.8333	0.9461	0.8495	0.8952	0.9005
SMO	0.5589	0.927	0.5139	0.6613	0.7228
GGD	0.985	0.9907	0.9914	0.9911	0.9607

TABLE XI: ALGORITHM PERFORMANCE ON PARIS ATTACK DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.7763	0.7799	0.9688	0.8641	0.2759
Gaussian Naive Bayes	0.608	1	0.4661	0.6359	0.4661
Linear Regression	0.7878	0.8138	0.9219	0.8645	0.4954
Logistic Regression	0.7878	0.8138	0.9219	0.8645	0.4954
SMO	0.8317	0.8246	0.9792	0.8952	0.4453
GGD	0.9732	1	0.9635	0.9814	0.9635

TABLE XII: ALGORITHM PERFORMANCE ON IND VS W.I DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.7593	0.7618	0.955	0.8476	0.3464
Gaussian Naive Bayes	0.6662	0.9183	0.5746	0.7069	0.6943
Linear Regression	0.7708	0.7932	0.91	0.8476	0.5349
Logistic Regression	0.7708	0.7932	0.91	0.8476	0.5349
SMO	0.53	0.9171	0.3619	0.5191	0.4385
GGD	0.9255	0.9524	0.9407	0.9465	0.9493

TABLE XIII: ALGORITHM PERFORMANCE ON MICROSOFT DATASET

Algorithm	Accuracy	Precision	Recall	F1	Predictive Index
LDA	0.9968	0	0	0	0
Gaussian Naive Bayes	0.9315	0.009	0.1875	0.0173	0.2536
Linear Regression	0.5449	0.0066	0.9375	0.013	0.6061
Logistic Regression	0.9825	0.0137	0.0625	0.0225	0.077
SMO	0.8228	0.0102	0.5625	0.02	0.7388
GGD	0.9998	1	0.9375	0.9677	0.9375

We also observe that our linear regression and linear discriminant analysis perform very well when evaluated across the commonly used evaluation metrics namely accuracy, precision, recall, and F1. However, the very low predictive index of these algorithms for different datasets shows us that fairly good numbers of the basic evaluation metrics are paradoxical that arise due to skewness in different datasets. The authenticity of predictive index in identifying the paradox cases can be shown by the number of True Credible (TC) and True Non-Credible (TNC) tweets identified (Table XIV) for different algorithms used in this paper.

TABLE XIV: NUMERICAL ANALYSIS OF CREDIBLE AND NON-CREDIBLE TWEETS IDENTIFIED BY VARIOUS ALGORITHMS

	Gaussian Naive Bayes		Linear Regression		Logistic Regression		LDA		SMO		GGD (Integrated Approach)	
Boko Haram	TC: 169	TNC: 7	TC: 195	TNC: 0	TC: 186	TNC: 9	TC: 0	TNC: 10	TC: 124	TNC: 10	TC: 199	TNC: 7
	FC: 3	FNC: 31	FC: 10	FNC: 5	FC: 1	FNC: 14	FC: 0	FNC: 200	FC: 0	FNC: 76	FC: 3	FNC: 1
Hamas	TC: 404	TNC: 219	TC: 728	TNC: 0	TC: 688	TNC: 119	TC: 727	TNC: 0	TC: 175	TNC: 225	TC: 729	TNC: 223
	FC: 12	FNC: 325	FC: 231	FNC: 1	FC: 112	FNC: 41	FC: 231	FNC: 2	FC: 6	FNC: 554	FC: 8	FNC: 0
News	TC: 2748	TNC: 659	TC: 4852	TNC: 9	TC: 4360	TNC: 443	TC: 4845	TNC: 13	TC: 1890	TNC: 701	TC: 4852	TNC: 661
	FC: 44	FNC: 2104	FC: 694	FNC: 0	FC: 260	FNC: 492	FC: 690	FNC: 7	FC: 2	FNC: 2962	FC: 42	FNC: 0
Harrisburg	TC: 138	TNC: 136	TC: 150	TNC: 0	TC: 150	TNC: 0	TC: 150	TNC: 0	TC: 84	TNC: 150	TC: 150	TNC: 141
	FC: 14	FNC: 12	FC: 150	FNC: 0	FC: 150	FNC: 0	FC: 150	FNC: 0	FC: 0	FNC: 66	FC: 9	FNC: 0
Indian Elections	TC: 932	TNC: 246	TC: 1366	TNC: 36	TC: 1031	TNC: 228	TC: 1366	TNC: 38	TC: 845	TNC: 82	TC: 1361	TNC: 247
	FC: 8	FNC: 455	FC: 218	FNC: 21	FC: 26	FNC: 356	FC: 216	FNC: 21	FC: 0	FNC: 1155	FC: 7	FNC: 26
Terror Attacks	TC: 1408	TNC: 635	TC: 2733	TNC: 0	TC: 2471	TNC: 337	TC: 2721	TNC: 48	TC: 107	TNC: 659	TC: 2737	TNC: 635
	FC: 24	FNC: 1329	FC: 659	FNC: 4	FC: 322	FNC: 266	FC: 611	FNC: 16	FC: 0	FNC: 2630	FC: 24	FNC: 0
US Presidential Elections	TC: 1003	TNC: 275	TC: 1495	TNC: 39	TC: 1281	TNC: 219	TC: 1473	TNC: 45	TC: 775	TNC: 231	TC: 1495	TNC: 278
	FC: 17	FNC: 505	FC: 253	FNC: 13	FC: 73	FNC: 227	FC: 247	FNC: 35	FC: 61	FNC: 733	FC: 14	FNC: 13
Parris Attacks	TC: 179	TNC: 139	TC: 354	TNC: 58	TC: 354	TNC: 58	TC: 372	TNC: 34	TC: 376	TNC: 59	TC: 370	TNC: 139
	FC: 0	FNC: 205	FC: 81	FNC: 30	FC: 81	FNC: 30	FC: 105	FNC: 12	FC: 80	FNC: 8	FC: 0	FNC: 14
India vs WI	TC: 281	TNC: 184	TC: 445	TNC: 93	TC: 445	TNC: 93	TC: 467	TNC: 63	TC: 177	TNC: 193	TC: 460	TNC: 186
	FC: 25	FNC: 208	FC: 116	FNC: 44	FC: 116	FNC: 44	FC: 146	FNC: 22	FC: 16	FNC: 312	FC: 23	FNC: 29
Microsoft	TC: 3	TNC: 4635	TC: 15	TNC: 2698	TC: 1	TNC: 4891	TC: 0	TNC: 4963	TC: 9	TNC: 4088	TC: 15	TNC: 4963
	FC: 328	FNC: 13	FC: 2265	FNC: 1	FC: 72	FNC: 15	FC: 0	FNC: 16	FC: 875	FNC: 7	FC: 0	FNC: 1

VI. CONCLUSION AND FUTURE SCOPE

This paper highlighted the effect of skewness in different datasets, which include paradox cases of basic evaluation metrics like accuracy, precision, and, recall, on some popular machine learning algorithms. The use of predictive index in identifying paradox cases was also shown. A model that integrates the generative and discriminative properties of GNB (Gaussian Naïve Bayes) and logistic regression was used in overcoming the effect of skewness on the classification of tweets. Experimental results demonstrated the success of identifying the effect of skewness on the performance of various machine learning models and how our proposed model performed better than all the other machine learning models used in this paper by not only achieving high accuracy but also high predictive power.

As part of future work, we tend to increase the number of features used in the credibility assessment of tweets and thereby draw a relation between the number of features and credibility assessment of the content. In addition, we can also identify users sharing non-credible tweets, which possibly can help us in identifying fake accounts on Twitter one of the major concerns that surfaced recently [43]. We would run our model on skewed datasets of other online social media platforms like Facebook.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Shifaa Basharat reviewed the existing research on machine learning approaches and worked on the significance of current study. She was also responsible for writing the paper. Saduf Afzal contributed in terms of data collection and annotation. Shozab khurshid identified the features used in the study. Manzoor Chachoo helped in identifying the machine learning models used for comparative study. Alwi Bamhdi was responsible for overall formulation of the paper. All the authors have approved the final version.

REFERENCES

[1] Social media fact sheet. (2021). Pew Research Center. [Online]. Available: http://www.pewinternet.org/fact-sheet/social-media/

[2] Internet overtakes newspapers as a news outlet. (2008). Pew Research Center. [Online]. Available: http://pewresearch.org/pubs/1066/internet-overtakes-newspapers-as-newssource

[3] K. Thirunarayan, P. Anantharam, C. Henson, and A. Sheth, "Comparative trust management with applications: Bayesian approaches emphasis," *Future Generation Computer Systems*, vol. 31, pp. 182–199, 2014.

[4] S. Pogatchnik, "Student hoaxes world's media on Wikipedia," *MSNBC Technology & Science*, vol. 12, 2009.

[5] S. Laird. (2012). How social media is taking over the news industry. *Mashable*. [Online]. Available: https://mashable.com/archive/social-media-and-the-news

[6] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. the 19th International Conference on World Wide Web*, 2010, pp. 591–600.

[7] W. Stassen, "Your news in 140 characters: exploring the role of social media in journalism," *Global Media Journal-African Edition*, vol. 4, no. 1, pp. 116–131, 2010.

[8] Anti-Phishing Working Group. Phishing activity trends report Q4/2012. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_Q4_2012.pdf

[9] Symantec Corporation. (2013). ISTR: Internet Security Threat Report 2013. [Online]. Available: http://www.symantec.com/threatreport/

[10] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we RT?" in *Proc. the First Workshop on Social Media Analytics*, 2010, pp. 71–79.

[11] P. Domm, "False rumor of explosion at white house causes stocks to briefly plunge; AP confirms its Twitter feed was hacked," *CNBC Newsletters*, 2013.

[12] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. the 20th International Conference on World Wide Web*, 2011, pp. 675–684.

[13] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Proc. the 1st Workshop on Privacy and Security in Online Social Media*, 2012, p. 2.

[14] K. R. Saikaew and C. Noyunsan, "Features for measuring credibility on Facebook information," *International Scholarly and Scientific Research & Innovation*, vol. 9, no. 1, pp. 174–177, 2015.

[15] M. Kang, "Measuring social media credibility: A study on a measure of blog credibility," Institute for Public Relations, pp. 59–68, 2010.

[16] V. L. Rubin and E. D. Liddy, "Assessing credibility of weblogs," in *Proc. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 187–190.

[17] B. Soiraya, A. Mingkhwan, and C. Haruechaiyasak, "E-commerce website trust assessment based on text analysis," *International Journal of Business and Information*, vol. 3, no. 1, 2008.

[18] M.-A. Abbasi and H. Liu, "Measuring user credibility in social media," in *Proc. International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 2013, pp. 441–448.

[19] G. Barbier and H. Liu, "Information provenance in social media," in *Proc. International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 2011, pp. 276–283.

[20] M. Jamali and M. Ester, "Trustwalker: a random walk model for combining trust-based and item-based recommendation," in *Proc. the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 397–406.

[21] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proc. the 13th International Conference on World Wide We*b, 2004, pp. 403–412.

[22] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 183–194.

[23] U. Kuter and J. Golbeck, "Sunny: A new algorithm for trust inference in social networks using probabilistic confidence models," in *Proc. AAAI*, vol. 7, 2007, pp. 1377–1382.

[24] M. Mathioudakis and N. Koudas, "Twittermonitor: Trend detection over the Twitter stream," in *Proc. the 2010 ACM SIGMOD International Conference on Management of Data*, 2010, pp. 1155–1158.

[25] V. P. Sahana, A. R. Pias, R. Shastri, and S. Mandloi, "Automatic detection of rumored tweets and finding its origin," in *Proc. 2015 International Conference on Computing and Network Communications (CoCoNet)*, 2015, pp. 607–612.

[26] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proc. the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1589–1599.

[27] S. Hamidian and M. T. Diab, "Rumor detection and classification for Twitter data," in *Proc. the Fifth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS)*, 2015, pp. 71-77.

[28] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on twitter with hybrid CNN and RNN models," in *Proc. the 9th International Conference on Social Media and Society (SMSociety'18)*, 2018, pp. 226–230, doi: 10.1145/3217804.3217917

[29] M. Naderan, E. Namjoo, and S. Mohammadi, "Trust classification in social networks using combined machine learning algorithms and fuzzy logic," *Iranian Journal of Electrical and Electronic Engineering*, vol. 15, 2019, doi: 10.22068/IJEEE.15.3.294

[30] C. Xu, Y. Yuan, and M. Orgun, "Using Bayesian networks with hidden variables for identifying trustworthy users in social networks," *Journal of Information Science*, vol. 46, 2019, doi: 10.1177/0165551519857590

[31] P. K. Verma and P. Agrawal, "Study and detection of fake news: P2C2-based machine learning approach," *Data Management, Analytics and Innovation*, pp 261–278, 2020, doi: 10.1007/978-981-15-5619-7_18

[32] B. Kardaş, İ. E. Bayar, T. Özyer, and R. Alhajj, "Detecting spam tweets using machine learning and effective preprocessing," in *Proc. the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'21)*, 2021, pp. 393–398, doi: 10.1145/3487351.3490968

[33] S. Basharat and M. Ahmad, "Inferring trust from message features using linear regression and support vector machines," in *Proc. International Conference on Next Generation Computing Technologies*, 2017, pp. 577–598.

[34] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing? Understanding microblog credibility perceptions," in *Proc. the ACM 2012 Conference on Computer Supported Cooperative Work*, 2012, pp. 441–450.

[35] S. Sikdar, B. Kang, J. O'Donovan, T. Höllerer, and S. Adah, "Understanding information credibility on Twitter," in *Proc. the 2013 International Conference on Social Computing*, 2013, pp. 19–24.

[36] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama, "Assessment of tweet credibility with LDA features," in *Proc. the 24th International Conference on World Wide Web*, 2015, pp. 953–958.

[37] K. Lorek, J. Suehiro-Wiciński, M. Jankowski-Lorek, and A. Gupta, "Automated credibility assessment on Twitter," *Computer Science*, vol. 16, no. 2, pp. 157–168, 2015.

[38] S. B. Fazili and M. Ahmad, "Guassian gradient descent model for trust inference in imbalanced data," in *Proc. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 929–934, doi: 10.1109/ICCONS.2018.8663243

[39] R. D. S. Raizada and Y.-S. Lee, "Smoothness without smoothing: Why gaussian naive bayes is not naive for multi-subject searchlight studies," *PLoS One*, vol. 8, no. 7, e69566, 2013, doi: 10.1371/journal.pone.0069566

[40] L. Ali, S. U. Khan, N. A. Golilarz, I. Yakubu, I. Qasim, A. Noor, and R. Nour, "A feature-driven decision support system for heart failure prediction based on a statistical model and Gaussian Naive Bayes," *Computational and Mathematical Methods in Medicine*, 2019.

[41] D. G. Kleinbaum and M. Klein, "Introduction to logistic regression," in *Logistic Regression. Statistics for Biology and Health*, New York, NY: Springer, 2010, doi: 10.1007/978-1-4419-1742-3_1

[42] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: A practical introduction," *BMC Medical Research Methodology*, vol. 19, pp. 1–18, 2019.
[43] S. Shead. (2022). Elon Musk says Twitter deal 'cannot move forward' until he has clarity on fake account numbers. *CNBC Newsletters*. [Online]. Available: https://www.cnbc.com/2022/05/17/elon-musk-says-twitter-deal-cannot-move-forward-until-he-has-clarity-on-bot-numbers.html