ISSN 2278-3091
Volume 10, No.3, May - June 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse441032021.pdf
https://doi.org/10.30534/ijatcse/2021/441032021

Gender Estimation on Social Media Using Recurrent Neural Network

Aamina Atta[1], Dr. Khalid Masood[2], Afrozah Nadeem[3], Sundus Munir[4]
[1]Lahore Garrison University, Pakistan, Aminaatta570@gmail.com
[2]Lahore Garrison University, Pakistan, Kmasoodk@gmail.com
[3]Lahore Garrison University, Pakistan, Afrozah@lgu.edu.pk
[3]Lahore Garrison University, Pakistan, sundusmunir@lgu.edu.pk

ABSTRACT

With the development of instant messaging innovation and social media, protection has turned into a significant issue. There is a danger of one's record being hacked and utilized by the unknown person unconsciously. While doing texting on social media many people use abbreviations, short messages, emojis, images. We tried with different methods to gain the best accuracy in this research. In this paper, we will attempt to check the personality of the individual based on his/her composing style. We will explore the possibility of predicting the gender of a writer utilizing semantic proof. For this reason, term and style-based grouping strategies are assessed over an enormous accumulation of text messages. This study depicts the development of a huge, multilingual dataset named with gender, and examines factual models for deciding the gender of unknown Twitter clients. Twitter gives a basic method to clients to express sentiments, thoughts and assessments, makes the client produced content and related metadata, accessible to the network, and gives simple to utilize web and application programming interfaces to get to the information. The fundamental focal point of this paper is to gather the gender orientation of the client from unstructured data, including the username, screen name, depiction and picture, or by the client produced content.

Key words: Abbreviation, Depiction, Enormous, Factual, Multilingual, Sentiments.

1. INTRODUCTION

Author classification is an issue while reading a research article [1]. The author classification could be determined as the issue of foreseeing the characteristics (e.g., natural qualities and social status) of the writer of a text report. The result of similar research is mainly used for economic forensic, implementation of the law, analysis of threats, and avoidance of terrorist actions. As a result of various study [2,3], efforts had been spent to enhance the forecast precision in author classification. In this study, we explore the issue of foreseeing the gender and age group of a message report creator. Specifically, we concentrate on the text-based conversation through the internet. This kind of conversations are noticed in online service like messaging, and press assisting composition like e-mail (electronic mail), group discussion, and social media platforms such as Facebook and Twitter. We should first develop the issue as a message categorization issue, wherein the wording in a report is utilized to characteristic a gender to the creator of the report. Secondly, we likewise examine the impact of stylistic-features (e.g., words lengths, the utilization of punctuation mark, and smileys) on foreseeing the gender. Finally, we make use of the contrary issue or discussed the impact of gender and age group on the composing style.

Social media websites like Facebook, Twitter have become many common means for an individual to interact and exchange their lives immediately along with short messages, pictures and music. The large variety of user's profile (see Figure 1 and Figure 2) and their network made the huge potential for advertising companies to produce a numerous value-added facility, like products publicity and valid inquiry. Although, the incomplete user profile is common because of user privacy-setting, that make obstacles for promotion and the compilation of online society views. Thus, user-profile recognition is important for online advertising approach and valid inquiry. Gender recognition is also a similar problem and can affect many applications, like user identification through social websites. In this study, we concentrate on the issue of gender recognition.

Existing works in the AI field focus on content examination arrangements. By removing highlights from companion systems, client profiles or post writings, analysts attempt to prepare models to make compelling distinguishing proof [4]–[5]. A significant number of these works utilize a similar philosophy: they believe every client to be an autonomous example from the populace, and dependent on the colossal information measures, the models are exact results of the unadulterated factual investigation. Guha and Wicker [6] noticed that in informal organizations, the all-inclusive community comprises of a practically equivalent number of males and females, albeit most interpersonal organizations that have been examined are not similarly organized. This discovering delineates those relations via web-based networking media framed by online practices might be one-sided and, along these lines, prescient on gender orientation. The worries of people constrained by associations among clients may prevent the precision of the techniques referenced previously.

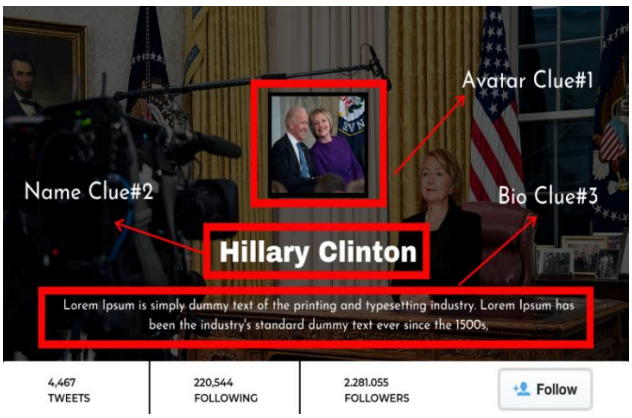

Figure 1: Female Twitter Profile

Figure 2: Male Twitter Profile

2. LITERATURE REVIEW

While the growth of machine learning, this is slowly been implemented on the review of revolutionary contents or opinions. Ferrara et al. [7] implemented machine learning capabilities on media platform messages to identify the cooperation of revolutionists. The suggested structure had tested over the group of 20,000 tweets originated through the revolutionary account, that is subsequently discontinued by Twitter. The major focus upon three functions: (i) identification of revolutionary utilizers, (ii) determining utilizers including revolutionary contents, (iii) forecasting consumers 'respond to revolutionary' posting. The tests are organized within two proportions, i.e., time dependence or real-time portent functions. A precision of nearly 93% is reached regarding the revolutionist's identification. By the same objective, the machine learning approach is suggested by [8] for categorizing of revolutionary connections. The Naïve Bayes approach is implemented using a standard group of elements. The structure is established over the categorization of a person's evaluation towards negative or positive class through little attention upon determining, that opinion classes (negative and positive) are related through revolutionary connection. In contrary with Ferrara et al. [7] effort, that have primarily highlights over categorization revolutionary connections on unbalanced information; the approach have applied NB computation over balance information shown better strong outcomes. Although, general dependency within the statement has not been examined. This matter could be managed through the implementation of deep learning model created over words integration characteristics. Searchers furthermore begin to examine different means of spontaneously testing revolutionary connections in a verbal language apart from English. In this respect, Hartung et al. [9] suggested a machine learning approach for identifying revolutionary post within German Twitter's account. Various characteristics are tested, like sentiments, language pattern, or literal signs. The structure waived enhanced outcomes through latest technology functions. Research on the arrangement of revolutionary connections within the environment of media platform contents is distinct in prohibited drugs using. For instance, in the works over marijuana relevant blogs, Nguyen et al. [10] gather over 30,000 tweets belonging to marijuana within 2016. The extraction of messages approach gives few beneficial perceptions towards obtained details like (i) person behavior may be classified as negative and positive, (ii) above 65% tweets are generated with smartphones, (iii) occurrence of tweets over the weekend is larger as compared to regular days.

Vocabulary supported unsupervised approach for opinions categorization primarily depend over few emotion terminologies or emotion marking module [11]. Alike different regions of opinion examine, revolutionary connections have

been inspected by Ryan et al. [12], by suggesting a new technology based on parts of speech labelling or opinion conducted identification of revolutionary writer through network platforms. The research has been established over 1 million posts upon over 25,000 several users browsing over four revolutionary platforms. The suggested approach has been established over a person's opinion marking, calculated by compilation of the scores of numbers of a negative post, period of negative post or intensity of negative post. The structure is compliant to identify internet-based doubtful activity over revolutionary consumers.

Unsupervised methods such as grouping should effectively be applicable within several regions like feature establishment review [13], inventory portent [14] or sensation categorization. Skillicorn [15], in their work over criminal inspection, suggested a structure in regard to contradictory examination of details. The structure is composed of three main sections containing information gathering, identification of defendants, or seeking of doubtful persons utilizing web conducted combination approach. Another approach is conducted over the grouped information or semantic analysis capabilities for inspection or execution of terrorist information. Therefore, writers have exploited Twitter to identify and categorize criminal facts using civil sensations.

Crossbreed method for establishing sensation supported requests have been accepted significant focus of searchers in various regions, like businesses, medical attention or politic [15]. This methodology, various characteristics of semi-supervised, supervised, or unsupervised proficiencies are accepted [13]. Within the framework of revolutionary categorization, Zeng et al. [16] performed at the Chinese texts partition problem in the region of terror utilizing a collective data. The basic modules utilize collective data for handling data in terror region. The technology has the capability for treating a wide range of Chinese texts information. Assessment of aggressive discussion from a chat over the internet. About 50 texts of communication on the internet over media platform have been assessed through implementing two methods that are quantitative or qualitative.

The above-mentioned study on identification or categorization of media-based platform revolutionary connection utilizes various methods, like supervised machine learning, unsupervised technology like 'lexicon or cluster-based or hybrid' model. Although, it is necessary to inspect the capability of the latest technology of deep learning model for categorizing revolutionary connections utilizing media platform contents.

3. METHODOLOGY

3.1 Data Collection

From this point on, Twitter permits easy access over its API to get tweets for the community. Twitter avoids the exchange of tweets to other regions against the intentional person. However, there is no openly accessible tweet dataset. For such work, we focused on English speaking people. For every Twitter user, Twitter provides geolocation details, it facilitates while collecting tweets based on the location of the person. We got 20,500 tweets with the name of the user and the content of the tweet. Even though every profile information is entirely inessential whenever any user is registered over the media platform, thus this provides some benefit from profile information which inform little bit regarding the user. The characteristics of the user profile have not been beneficial while trying to implement a supervised learning technique to discover gender features (see Figure 3).

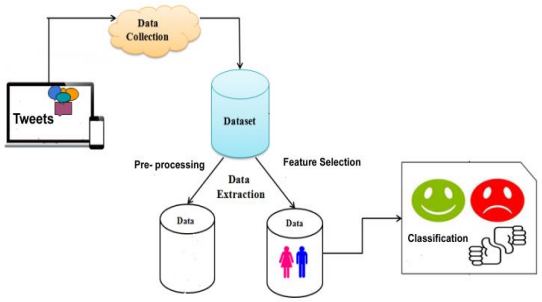

Figure 3: Methodology of Classification Process

3.2 Preprocessing

We have implemented various preprocessing approaches, like tokenization, removal of stop words, case conversion, and the removal of special symbols. Tokenization gives a unique set of tokens, that helps in making vocabulary through the training set, utilized for encrypting the text (see Figure 4).

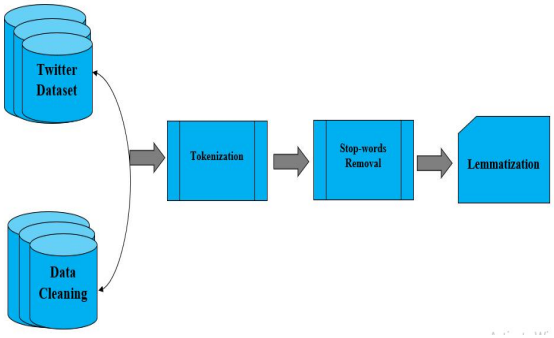

Figure 4: Classification Execution

3.3 Lemmatization

Lemmatization decreases the influence words appropriately make sure that the words about the language. The root words in lemmatization are called a lemma. For example, go, going, gone all are the forms of words go, thus go is the lemma of all these words. Because lemmatization proceeds an actual word of the language. It is applicable whereby it is essential to use. In this research, we use lemmatization to reduce the words that belong to the same phrase.

3.4 Training, Validating and Testing

Training data is utilized to train the model [17]. In this study, 80% of data is utilized training and it might differ by the conditions of experiments. Training data set contains input and the presumed output.

Data Validation is utilized to reduce misclassification and misspecification, that generally occurs due to the precision in the phase of training is frequently high and efficiency become decreased over test data. Thus, 10% of the validation set is

utilized to prevent efficiency error with the help of parameter tuning.

Test data 10% is utilized to verify either the trained model operates right over the unseen data. It is applicable to carry out the final analysis of the proposed model whenever the model is completely trained.

3.5 Proposed Network Model

Our general aim is to detect the gender from the tweets of a person so for this purpose we built our Long Short-Term Memory (LSTM) and Recurrent Neural Network.

3.5.1 Lstm Network

Long Short-Term Memory network is generally known as 'LSTMs'. It is a particular type of Recurrent Neural Network (RNN) which is able of understanding long term dependences. LSTM is specifically configured to escape from the long-term dependences. Memorizing the data for a prolonged period is their standard behavior, not a thing they conflict to understand. All the recurrent neural networks have the pattern of chain replicating the module of the neural network. We utilized a single LSTM layer network model, that comprises of 100 LSTMs cell/units.

All recurrent neural network has the group of chain replicate the modules of a neural network. In the normal RNNs, such replicating modules would have a quite simple configuration, similar to single tanh layer.

The first step in the LSTM is to choose which details we are going to drop through the cell state. Such resolution is made by the sigmoid layer which we called the "forget gate layer". In the next step, we have to decide which new details we are going to save in cell state. It has two parts. First, the sigmoid layer known as the "input gate layer" choose what kind of values we will update. Further, tanh layer builds a vector of new candidates' values, which can be added on to the state. After this, in the next step, we will merge these two for creating an update to the state cell. In the end, we want to choose what we are going to output. The output would be based over cell state, but it would be a filtered version. Firstly, we lead a sigmoid layer that chooses which part of the cell state we are going to output. Therefore, we place the cell state from tanh (to drive the values among -1 and 1) and multiply this by the output of the sigmoid gate, therefore we only output those parts which we want to. We implemented this algorithm in python language (see Figure 5).

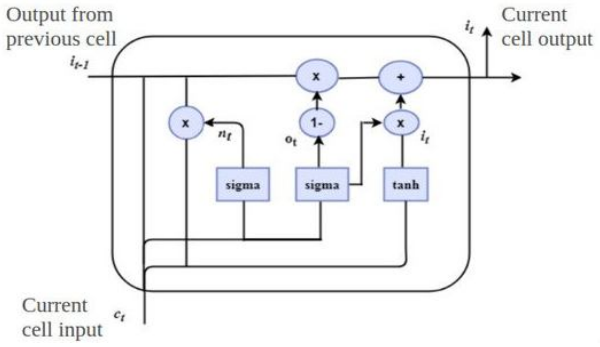

Figure 5: Working of Lstm

3.5.2 Recurrent Neural Network

Recurrent Neural Network is the induction of feedforward neural network which has internal memory. RNN is frequent when it carries out the same functions to each input of data whereas the output of the existing input relies upon the previous computations. Then generating the output, that is replicate and returned to the recurrent network. For producing the resolution, it takes the existing input and the output which has been learning by the past input.

In contrast to feedforward neural network, RNN may utilize their internal memory to proceed sequence of input. In the RNN, every single input is associated with the other inputs so they are connected.

Neural networks are the collection of artificial neurons or nodes which implement modifications to the data. Their essential performance gives input and produces the output. The fascinating feature of neural networks is that we did not tell them how the output will be generated. Instead, we establish how to 'learn' and output based on a huge amount of training dataset. Training dataset comprises inputs and output, generally marked by humans. The neural networks access the section of the training dataset, produce the output, match that output with the real outcome, or settle the weights on the nodes. Nodes are organized in layers within a recurrent neural network (see Figure 6).

3.5.3 Word Embedding (Input) Layer

The word embedding or the input layer which is the first layer of the LSTM+RNN model, that converts the words with real-value vector representations, like a lexicon of words are formed, that is changed to numerical format, which is known as word embedding. It is the illustration of text whereby words which have similar meanings have the same illustration. This means that words within an organized system whereby associated words, based upon the principle of relationship, are situated nearer simultaneously.

The input layer of the recurrent neural network consists of input neurons which fetch the data within the system for additional reprocessing through further layers of neurons. The input layer is the initial workflow for the neural network architecture. In our architecture, the Input layer has 100 input neurons.

3.5.4 Dropout Layer

The dropout layer relates to neglecting units (i.e., Neurons) within the training period of a particular group of neurons that is selected randomly. Dropout neurons are not reviewed within the certain forward and backwards pass. At the respective training phase, particular nodes are dropped out from network with the dropout rate 0.2, thus it ends up with a reduce network; incoming or outgoing sides to dropped-out nodes are also eliminated.

3.5.5 Dense Layer

We used an ordinary network of numerous completely connected (dense) layers to learn about the descriptive data. We design our neural network architecture thus congestion is constituted. In our architecture, the next layer of the neural network after the dropout layer is dense also called hidden layer. In our proposed model we have 100 neurons in the particular neural network layers. All the neurons are fully connected.

3.5.6 Output Layer

The output layer in the recurrent neural network is the last layer of neurons which gives output for the program. This layer is based upon the Sigmoid function for the binary classification and the SoftMax for binary and multi-classification output. Total learnable or trainable parameters in our neural network are: 10,080,501

Model Architecture

Maximum Sequence Length: 100

Embeddings Dimension: 100

Trainable params in network: 10,080,501

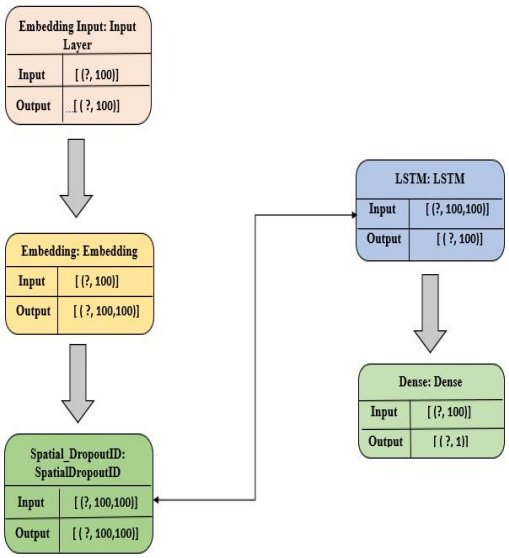

Figure 6: Model of Architecture

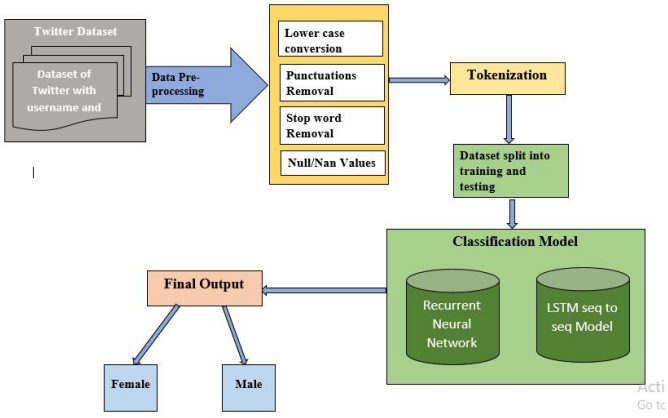

Figure 7: Block Diagram of Model

3.5.7 Sigmoid

The sigmoid activation function is very common in a neural network. This is the more extensively used function nowadays. It is a non-linear function; the input of the function is changed to the value between 0.0 and 1.0. Among (-2,2) over the x-axis, the function is extremely steep, which leads function to prioritize and classify the values either 1 or 0. It permits the nodes to accept any of the values between 1 or 0. At the end when there are multiple output classes, this would appear with various outcomes with the probabilities of 'activation' for every output class. And we will select the one which has high 'activation' (probability) value.

3.5.8 Rectified Linear Unit (ReLU)

In recurrent neural networks, the activation functions are reliable for modifying the aggregated weight input through the nodes within the activation of particular nodes or output of that input. Rectified Linear activation function or ReLU, in short, is a subsequent linear function which would output the input instantly if this is positive, contrary, this will give output zero. This activation model is easy to train and achieve better performance that's why it becomes the default activation model for neural networks.

3.5.9 SoftMax

It is a function which rotates a vector of real values into a vector of real values which sum to 1. The input values might be positive, negative, zero, or larger than one, however, the SoftMax function turns them into values with 0 and 1, thus they could be interpreting as probabilities. In case of any input is small and negative, the SoftMax function turn it into small probability, or if the input is large, then it will turn into large probability, but it remains lies within 0 and 1.

3.6 Logistic Regression

Logistic Regression is one of the more essential or commonly used machine learning approaches. It is a probabilistic classification model. Logistic regression is a predictive algorithm utilizing independent variable to forecast the dependent variables, similar to Logistic Regression, yet by the variance that the dependent variables must be absolute. Independent variable could be any numeric or absolute variable, but the dependent variable will be constantly categorical. The probability would be forever lying among 0 and 1. Logistic Regression could be utilized for the multi-class classification and binary classification. Logistic Regression utilizes logit function, as well stated to log-odds; this is the logarithm of odds.

$$ln\left(\frac{p}{1-p}\right) = \theta_1 + \theta_2 x + e$$

$$\frac{p}{1-p} = e^{\theta_1 + \theta_2 x + e}$$

$$P = \frac{1}{1+e^{-}}(\theta_1 + \theta_2 x)$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \; where \; z = \theta^T x$$

$$\theta^T x = \sum_{i=1}^{m} \theta_i x_i = \theta_1 x_1 + \theta_2 x_2 + \ldots\ldots + \theta_m x_m$$

- P is the probability that event Y occurs. P(Y=1)
- P/(1-P) is the odds ratio
- θ is a parameter of length m.

Logit function considers the probability among 0 and 1, and thus logistic regression is non-linear transition.

4. EXPERIMENTS

4.1 Dataset

We have used PAN Training and PAN Testing dataset provide by 2017 PAN Author Profiling. The classes labelled as Male and female. The dataset we used in this study is in the English language. We composed 240000 tweets, Total unique words in tweets are 260995 and the maximum number of words used: Top 100,000 words. We distributed PAN Training in two parts: 70% for the training, 30% for the validation (see Table 1).

Table 1: Tweets

Tweets	Gender
Omg now I remember that photo was from when I had my camera out and accidentally dropped the phone on my face and it took a burst	Female
I just randomly came across this photo on my camera roll and it's officially my favourite photo ever potatosutti https t co pG7bOaQaXf	Female
NUFC right let s see Perez and Gouffran off at HT and let s bring Townsend and Armstrong on oh we can t can we Think we will lose 1 3	Male
I hate how immigrant or refugee is considered a dirty word to some people My Nonna was a refugee I m an immigrant and PROUD	Female
realDonaldTrump go for the extra mayo You know it makes sense	Male

4.2 Logistic Regression

Logistic Regression is one of the more essential or commonly used machine learning approaches. It is a probabilistic classification model. Logistic regression is a predictive algorithm utilizing independent variable to forecast the dependent variables, similar to Logistic Regression, yet by the variance that the dependent variables must be absolute variables. Independent variable could be any numeric or absolute variable, but the dependent variable will be constantly categorical. For mapping predicted values t0 the probabilities, we used the sigmoid activation function. This function maps the real values to other values within the range of 0 and 1. For mapping the values we use the logit function, which maps the probabilities between (0,1). Cross entropy in logistic regression means that it is a measure of difference within the two probabilities. It calculates the difference between the two probability distributions. When we use the loss function for the classification model, both measures will calculate the similar and maybe utilized replaceable. First, we train the model over training dataset. For every pattern within test dataset, we implemented a logistic regression model to produce the probabilities. The training accuracy that we achieve after performing the logistic regression is 52.41%. we may test our logistic regression model through certain instances. The test accuracy that we achieve is 52.33%. So, the results we got in logistic regression are not good.

4.2.1 Confusion Matrix

A confusion matrix is a table which is utilized to measure the execution of the classification models. We might as well figure the efficiency of the algorithm (see Figure 8 and Figure 9). The essential of the confusion matrix is that the number of incorrect and correct predictions are summarized class wise. When we collect the dataset, after cleaning the dataset and preprocessing, the initial step we do is to take it to an extraordinary model and obviously, want the output in probabilities. But how we measure the efficiency of our model. When our model gives better performance automatically, we get good performance and that's the thing we want in our system. For this, we use a confusion matrix which is performance measurement in machine learning where output is maybe of two classes. Let's discuss the basic terminologies of confusion matrix:

True Positive: The male perception detected by the system as male.

True Negative: The female perception detected by the system as female.

False Positive: The female perception detected by the system as a male.

False Negative: The male perception detected by the system as a female.

Confusion Matrix (Training data)

0 = female

1 = male

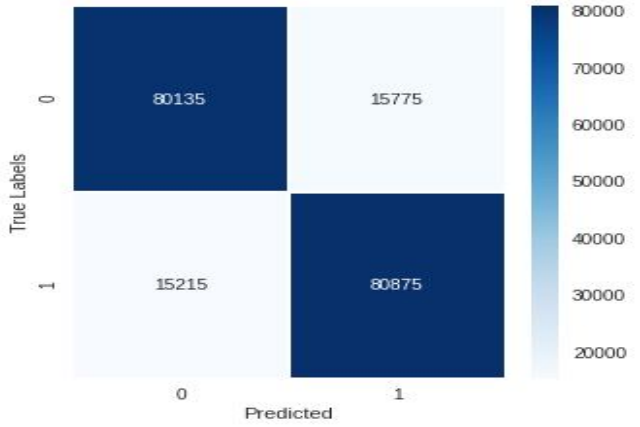

Figure 8: Confusion Matrix of Training Data

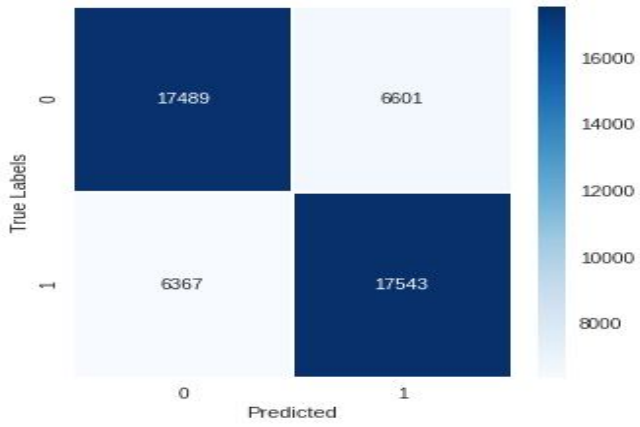

Figure 9: Confusion Matrix of Test Data

4.3 Precision Recall and F1-score

Precision means how accurate/definite is the model. Thoroughly measure the efficiency of the model, we must review both precision and recall. Thus, when the precision is improving it generally decreases the recall.

- *Precision*

Precision = (TP) / (TP+FP)

TP is the number of true positive, FP is the number of false positive. The insignificant procedure to acquire absolute

precision is to form one single positive prediction and make sure it is accurate (precision = 1/1 = 100%). Such a measure not to be very handy as long as the classifier will neglect all except one positive instance.

- *Recall*

Recall = (TP) / (TP+FP)

From all of the positive classes, how many times we predicted accurately, it must be as high as possible.

- *F1 Score*

This is generally appropriate to merge the precision and recall to singe metric known as the F1 score, in certain, if you require an easy way to contrast two classifiers. F1 score is the harmonic mean of precision and recall. The F1 score favour classifier which has the same precision and recall. It is not constantly what we want: in a certain context, we

generally, care about precision and in another context, we care about the recall.

F-measure = 2*Recall*Precision / Recall + Precision

4.4 Random Forest

Random forest is compliant, convenient to use an algorithm which generates better results sometimes. Random forest is a supervised machine learning algorithm. The "forest" builds up, is a composite of the decision tree, generally train through the "bagging" procedure. The comprehensive thought of bagging procedure. It builds various decision trees and combines them collectively to achieve more precise and consistent predictions. While training the dataset in the random forest we got 99.37% training accuracy and while testing the dataset in this algorithm we achieve 54.90% test accuracy. So, the results are average in the random forest algorithm not satisfactory.

Precision-Recall and F1-Score for Test data (see Table 2)

Table 2: Precision-Recall and F1 Score for Test Data

	Precession	**Recall**	**F1-Score**	**Support**
Female	0.73	0.73	0.73	24090
Male	0.73	0.73	0.73	23910
Accuracy	0.73	0.73	0.73	48000
Macro-avg	0.73	0.73	0.73	48000
Weighted-avg	0.73	0.73	0.73	48000

Precision-Recall and F1-Score for Training data (see Table 3)

Table 3: Precision-Recall and F1 Score for Training Data

	Precession	**Recall**	**F1-Score**	**Support**
Female	0.84	0.84	0.84	95910
Male	0.84	0.84	0.84	96090
Accuracy	0.84	0.84	0.84	192000
Macro-avg	0.84	0.84	0.84	192000
Weighted-avg	0.84	0.84	0.84	192000

4.5 Multinomial Naïve Bayes

Multinomial Naïve Bayes is a special version of Naïve Bayes which is configured better for the text document. While the simple naïve Bayes will make a document as some words are present and some are missing, multinomial naïve Bayes specifically model the word count and amend the essential calculations to handle within. we had performed our training and testing validation on multinomial naïve Bayes algorithm when we train the dataset over multinomial Bayes algorithm we got 51.80%

accuracy. And when we apply testing on the dataset, we got 51.57% accuracy. The results are not satisfactory, we didn't get the best results on the multinomial naïve Bayes algorithm.

4.6 Sequential Neural Network

The sequential neural network develops high-level characteristics with the successive layers. We proposed a neural network model whereby every layer is connected with the other layers. A sequential model is suitable for those layers which have a single input and the single output unit. The resultant model is organized by DAG like architecture, such that a pathway from the root node to the leaf node describes the sequence of conversions. Rather than examining general modifications, as in the classical multilayer network, such model permits us for learning a group of regional alterations. Thus, it is capable of processing of data through various features using particular sequences of comprehensive transitions, it increases the expression power of the model concerning the classical multilayer network. Let's have a look below about the layers their types, their output and shapes, and what are their parameters (see Table 4).

Table 4: Sequential Neural Network Based Model

Layer (Type)	Output Shape	Param #
Embedding (Embedding)	(None, 100, 100)	10000000
Spatial_dropoutId (SpatialDr)	(None, 100, 100)	0
lstm (Lstm)	(None, 100)	80400
dense (Dense)	(None, 1)	101

Total Params: 10,080,501
Trainable Params: 10,080,501
Non-trainable Params: 0

4.7 Long Short-Term Memory (Lstm)

Long Short-Term Memory network is generally known as 'LSTMs'. It is a particular type of RNN which is able of understanding long term dependences. LSTM is specifically configured to escape from the long-term dependences. Memorizing the data for a prolonged period is their standard behavior, not a thing they conflict to understand. LSTM has corresponding sequence controller as a recurrent neural network. This processes the data passing over information since it disperses further. The dissimilarities are the operations among the LSTM's cells. Such operations are utilized to permit the LSTM to kept or neglect information. The fundamental concept of LSTM is the state of cells, and it's several gates. The states of cells behave as a transport which conveys the related details completely down to the sequence chain. We may think about it as "memory" of the network. The state of cells might transfer the appropriate details across the sequence of processing. So, the information by the previous time scale may create its path for subsequent time steps, lowering the impact of short-term memory. Since the cell state comes off on its excursion, data become inserted and deleted from the cell state through gates. Gates are a various neural network which decides that data is permitted over the cell state. This is decided by the gates that which data is appropriate to kept or neglect at the time of training. The scores are noted at the end of every epoch.

4.8 Comparison with Other Models

Table 5: Accuracy Comparison

Models	Parameters	Training Accuracy	Test Accuracy
Logistic Regression		52.41%	52.33%
Random Forest	n_estimators = 18	99.37%	54.90%
Multinomial Naive Bayes		51.80%	51.57%
Sequential Neural Network based model	3 FC hidden layers. Dropout used	98.2%	60.35%
LSTM 1	Custom built with 1 hidden layer, dropout used, 100 embeddings, With 80000 top words. Epochs: 30	79.30%	71.65%
LSTM 2	Custom built with 1 hidden layer, dropout used, 100 embeddings, With 100000 top words, Epochs: 40	84%	72.95%

The previous results of researchers are satisfactory they implement the already made algorithms through which they achieve satisfactory accuracy which is not pretty good. Researchers implement the logistic regression, Naïve Bayes, Random Forest. In my study, I built my model that is LSTM+RNN model which gives efficient result and accuracy. We use this model because we want Long Term Dependency for our dataset (see Table 5).

4.9 Results and Metrics

Training Accuracy: 84.43

Test Accuracy: 72.95

Validation Accuracy: 72.83

Training and Validation accuracy and loss curves

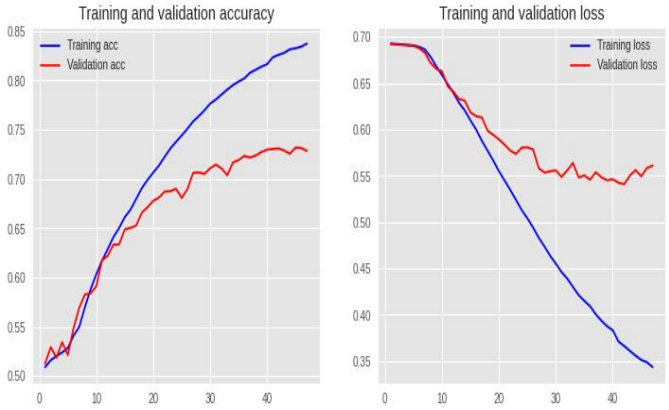

Figure 10: Accuracy Results

5. CONCLUSION

Both the corporation and humans are progressively reliant over social media platform for exchanging, contact and social relationships. Consequently, research workers have been offered over new sources of data and possibilities for extracting the data to enhance awareness and provision of services. Twitter is a well-known social media platform, has been perceived growing research concern recently as an opportunity to monitor, understand, and eventually forecast the real-life occurrence. Detecting gender over twitter is complex due to tweets as they are not available easily, tweets are short and, in several cases, they include informal texts, slang words.

In this research, we start through considering the previous methods to conquer the trouble of detecting gender on Twitter. In the earlier research, the most frequent characteristics are based upon text content, information about the profile and social contact. Al Zamal et al. [18] present the usage of characteristics correlate with the principle of homophily. It means, to conclude-person features based on nearest neighbors' features utilizing tweet contents and the information about the profile. The profile picture is generally neglected, although this may offer hints to users' gender. After examining earlier research, we have tested an approach to detect the gender particularly based on unorganized data accessible in a user profile. We began through mining data by Twitter and developed the dataset of English users. Once the dataset is created, we mined characteristics associate names found within user name and screen name through comparable gender. We assessed the execution of features utilizing various methods, comprising Multinomial Naïve Bayes, Logistic Regression, Random forest, Sequential Neural Network and LSTM. Results show the different accuracies in each of the approaches. Logistic Regression achieve 52.33% accuracy, Random Forest reached 54.90% accuracy, Multinomial Naïve Bayes achieve the 51.57% accuracy, Sequential Neural Network reached the 60.35% accuracy, but LSTM model proves suitable results by achieving 72.95% accuracy. W observed that the rising amount of dataset has a positive effect on the results.

Our further step was to produce prolonged marked dataset based on profile features and tweets. We made an English dataset, filter user through the tweet's language, comprised of 2400 users. From each user profile, we extracted 100 tweets. After the development of the data set, we classify the utilizers through the multinomial Naïve Bayes model. This is important to observe that when we executed our first experiment, we had only 3000 labelled datasets. When we increase the dataset further to 240000 tweets, we achieved better results.

We implemented an approach for gender detection with the usage of the combined classifier. Rather than implementing the same classifier for the entire characteristics, we classify the relevant features separately. We utilized the prolonged labelled dataset from our earlier experiments, divided into validation, train and test data set. The characteristics based on users content, profile details, split into various groups: name of the user, description, content of the tweets, profile picture and social contacts. The classes of characteristics to be assessed was screen name and the user name, description, content of tweets. we utilized 240 users name. LSTM achieves the best performance with the accuracy of 72.95% for English users. The other experiments we performed over the dataset was average, not satisfactory because the accuracy is very low in these experiments Logistic Regression, Random Forest, Multinomial Naïve Bayes algorithms.

5.1 Future Work

Although the classifier achieves the higher accuracy, our main aim is to enhance the classifier with the addition of new features. In future, we will incorporate the classification with detection of age and gender both. We will determine the age of the tweet sender that the particular person belongs to which age category. In future, we can proceed with this study based on a profile picture or based on facial features and expressions. With the help of the facial features from the pictures or profile, we will determine the gender.

REFERENCES

1. Love, H.: Attributing Authorship: An Introduction. Cambridge University Press (2018).
2. Corney, M.W.: Analyzing E-mail Text Authorship for Forensic Purposes. M.S. Thesis. Queensland University of Technology (2017).
3. Holmes, D.I.: Analysis of Literary Style - A Review. Journal of the Royal Statistical Society 148(4) (2019) 328–341.
4. J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, ‘‘Discriminating gender on Twitter,’’ presented at the Conf. Empirical Methods Natural Lang. Process., Jul. 2011. [Online]. Available: https://dl.acm. org/citation.cfm?id=2145568
5. X. Yan and L. Yan, ‘‘Gender classification of Weblog authors,’’ presented at the AAAI Spring Symp. Comput. Approaches Anal. Weblogs, Mar. 2006. [Online]. http://www.aaa~i.org/
6. Papers/Symposia/Spring/2006/SS-06-03/SS06-03-046. pdf.
7. Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni (2018). Gender, genre, and writing style in formal written texts. TEXT pp. 321--346...
8. Ferrara E, Wang WQ, Varol O, Flammini A, Galstyan A (2016) Predicting online extremism, content adopters, and interaction reciprocity. International conference on social informatics. Springer, New York, pp 22–39
9. Azizan SA, Aziz IA (2017) Terrorism detection based on sentiment analysis using machine learning. J Eng Appl Sci 12(3):691–698
10. Hartung M, Klinger R, Schmidtke F, Vogel L (2017) Identifying right-wing extremism in german Twitter profiles: a classification approach. International conference on applications of natural language to information systems. Springer, Cham, pp 320–325
11. Nguyen A, Hoang Q, Nguyen H, Nguyen D, Tran T (2017) Evaluating marijuana-related tweets on Twitter. IEEE 7th annual computing and communication workshop and conference (CCWC). IEEE, New Jersey, pp 1–7
12. Asghar MZ, Khan A, Ahmad S, Qasim M, Khan IA (2017) Lexicon-enhanced sentiment analysis framework using a rule-based classification scheme. PLoS ONE 12(2): e0171649
13. Ryan S, Garth D, Richard F (2018) Searching for signs of extremism on the web: an introduction to sentiment-based identification of radical authors. Behav Sci Terror Pol Aggres10:39–59. https://doi.org/10.1080/19434472.2016.1276612
14. Asghar MZ, Khan A, Zahra SR, Ahmad S, Kundi FM (2017) Aspect-based opinion mining framework using heuristic patterns. Cluster Comput pp 1–19 [32]
15. Asghar MZ, Rahman F, Kundi FM, Ahmad S (2019) Development of stock market trend prediction system using multiple regression. In: Computational and mathematical organization theory. pp 1–31
16. Skillicorn D (2017) Computational approaches to suspicion in adversarial settings. Inform Syst Front. https://doi. org/10.1007/s10796-010-9279-4
17. Zeng D, Wei D, Chau M, Wang F (2016) Domain-specific Chinese word segmentation using sufx tree and mutual information. Inform Syst Front. https://doi.org/10.1007/s10796-010-9278-5
18. Acharya A (2017) Comparative study of machine learning algorithms for heart disease prediction.
19. Al Zamal, F., Liu, W., and Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbours. ICWSM, 27