Received: October 23, 2018

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Behaviour Analysis Voting Model Using Social Media Data

Dalia Sameh[1]* **Ghada Khoriba[2]** **Mohamed Haggag[2]**

[1]*Computer Science Department,*
Modern Academy for Computer Science and Management Technology, Egypt
[2]*Computer Science Department, Faculty of Computers & Information,*
Helwan University, Egypt
* Corresponding author's Email: dalia_1991@live.com

Abstract: Nowadays, Social Media has become an important communication tool. Behavior could be measured using machine learning or deep learning techniques. This study represents a proposed model that shows how classification techniques can be used to recognize the personality according to the individuals' tweets using voting technique. The individual behavior is classified according to Eysenck's Three Factor personality model. A comparison was conducted between various machine learning and deep learning approaches as an input to a voting algorithm; which gave us more accurate results in the classifying task. The test results were 84.208% accurate. The study's main target is to make a new hybrid model for Twitter behavior analysis to enhance the accuracy of every approach individually on the dataset. Introducing the deep learning algorithm was to overcome the complexity and the time consuming obstacles. So this study can be used to predict future customers' behavior in order to increase satisfaction.

Keywords: Naïve Bayes, Social media analysis, Twitter analysis, SVM, SMO, CNN, RNN, Behavior analysis models.

1. Introduction

Twitter and Facebook have become an extremely popular communication tools through which users describe their opinions. The publicity and availability of data over the internet allow researchers, using data mining approaches, to analyze and to predict data from the tweets of the users. Researchers can measure the customers' satisfaction, get ratings, or make sentiment analysis [1]. Predicting future customer behavior is an important task in order to offer the customers the best possible experience and to improve their satisfaction. The large input dimensionality of the number of parameters to learn was the main problem of behavior analysis.

Text classification is how to distinguish to which category or class a new text belongs, based on a training dataset containing text or instances whose category or class is known or predefined [2].

Everyday millions of people share their opinions and thoughts through social media channels; this research is targeting identifying human behavior by analyzing their tweets. Previously, researchers used a system that was based on questionnaires to detect the personality of a person; this system had many limitations (i.e. maximum word limit is 140 characters). They also used Facebook data and got a high accuracy percentage of 91%, but the paper did not mention the percentage of accuracy when tested in real life.

Sentiment analysis is a way to categorize peoples' opinions and to know their directions; it also helps in decision making. The opinions could be assorted into: positive, negative and neutral. The demand on sentiment analysis is increasing due to an increase in need to analyze the hidden

information which come from different social media channels [3].

Techniques of sentiment analysis include: Lexicon-based or Corpus-based technique and machine learning-based techniques. To accurately predict users' behavior, some classification machine learning-based techniques are used to make a comparative study in this research. Those techniques are: Naïve Bayes, SVM, SMO, Bagging, Attributed Selected [4-8].

G.E Hinton [9] was the first to use deep learning expression in 2006. It involved many networks such as: (Convolutional Neural Networks) CNN, (Recurrent Neural Networks) RNN, Recursive Neural Networks, and (Deep Belief Networks) DBN [10].

Deep learning is a part of ML and a special type of artificial neural network (ANN) that is similar to a multilayered human cognition system. ANN was introduced in 1950, with many limitations in its application to solve real dilemmas; which has been now solved where big data is available, as it works on enhancing the computing power with graphics processing units (GPU), and training the deep neural network (DNN) with new algorithms [11].

Deep learning is a class of machine learning algorithms that uses a flow of multiple layers of nonlinear processing units for feature extraction and transformation. Each consecutive layer uses the output from the preceding layer as input, and learns in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners [12]. Thus, a voting technique is proposed by this study in order to choose the most efficient and effective performing algorithm that would increase the accuracy of behavior detection using a machine learning technique (SMO) and deep learning techniques (RNN and CNN). First, we detect behaviour from sentiment, and then predict, analyse and compare personality traits results using machine learning techniques (ex. SMO, SVM, Naive Bayes, Bagging, Attribute selected) and deep learning techniques (ex. CNN, RNN) for large twitter dataset (training 66731 records and a testing dataset containing 12500). Finally we analyze and compare prediction results from voting technique.

This paper is organized as follows: Section 2 presents some basic concepts and background for the approaches. Section 3 summarizes most of related work. Section 4 presents proposed model. Section 5 shows the experiments' results. Finally, this study's conclusion is found in Section 6.

2. Basics and backgrounds

Behaviour on social media is the way users interact through social media (as Facebook, Twitter, Instagram, etc.). Interaction may be a comment, post or just a share or a like. Analysing these actions allows us to predict the user's behaviour using for example: Big Five Model [13].

Personality traits studies with Facebook majorly use Five Factor Model (FFM) of personality as a guiding framework.

Five Factor Model is a widely researched model that uses the following personality traits: openness, conscientiousness, Extraversion, Agreeableness and Neuroticism [14-19].

Enrick Three Factor Model [20] is a personality theory that divides the person's behaviour into three major dimensions:

1- Extraversion
2- Neuroticism
3- Psychoticism

Each factor is described by a scale ranging from low to high; they are orthogonal and independent. Fig. 1 shows the measurements of personality.

Fabio Celli and Cristinna Zaga automatically annotated personality labels by an unsupervised system, then validated it on small set of Twitter users by collecting data from an online test and got results to determine the sentiment according to the individual's behaviour. This can be verified vice versa; determining the behaviour of the individual is based on his/her sentiment [21].

2.1 Classification techniques

a) Naive Bayes has been extensively studied since the 1950s. It is a popular Text categorization method for solving document problems as detecting spam or just specifying the category of the document if it belongs to politics, culture, sport, etc. [22].

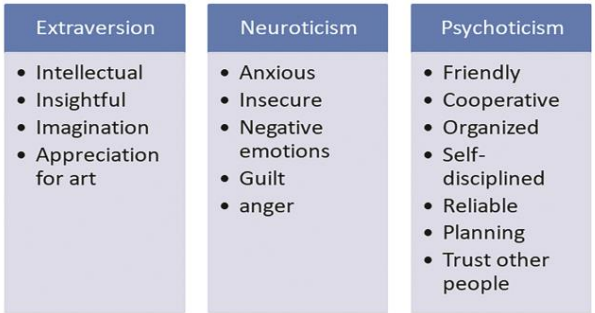

Figure. 1 Enrick three factor model traits

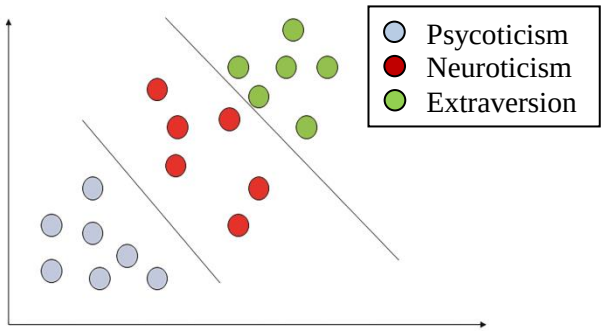

Figure. 2 SVM

The Naïve Bayes equation used to calculate the probability of occurrence of behaviour to the total number of occurrence to detect the highest one.

$$P(C_k|X)=\frac{P(C_k)P(X|C_k)}{P(X)} \quad (1)$$

Where $P(C_k|X)$ is the posterior probability of class (target), given predictor (attribute), $P(C_k)$ is the prior probability of class, $P(X|C_k)$ is the likelihood which is the probability of attribute given class, and $P(X)$ is the prior probability of predictor [23].

b) SVM (Support Vector Machine) is a supervised learning method that analyses data. Using this algorithm, it is easy to draw every data item as a point in n-dimensional space (where n is number of features you have) and the value of each feature. Then classification is preceded by finding the difference between the three classes very well (Fig. 2).

$$F(x) = \beta_0 + \beta^T x \quad (2)$$

where β^T is known as the weight vector and β_0 as the bias [24].

c) SMO (Sequential minimal optimization) is an algorithm to solve the quadratic problem that happens during the training of SVM, It is an iterative algorithm that shatters the problem into a series of smallest possible sub-problems, which are then, could be analytically solved.

$$0 \leq \propto_1, \alpha_2 \leq C \quad (3)$$

$$y_1\alpha_1 + y_2 \propto_2 = K \quad (4)$$

The algorithm proceeds as follows:

1- Find a Lagrange multiplier α_1 that breaks the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.

2- Pick a second multiplier α_2 and improve the pair (α_1, α_2).

3- Repeat steps 1 and 2 until convergence [25].

d) Bagging is a statistical method for considering a quantity from a data sample that is very small and has a mean error; so we have to improve that mean by:

1- Randomly creating (for example) 1000 sub-samples of the dataset with replacement

2- Calculating the mean of each sub-sample.

3- Calculating the average of all of our collected means and using that as our estimated mean for the data.

For example, if the sample result is 3; mean values are 2.3, 4.5 and 3.3 consecutively. The average of the mean values is 3.367.

When applying the same on Bagging using CART algorithm, the algorithm would work as follows:

1- Randomly creating (for example) 1000 sub-samples of the dataset with replacement

2- For each sample, applying CART model for training

3-Using a new dataset for testing

The average prediction is calculated from each model [26].

e) Attributed Selected is the process of picking out a subset of features to be used in the pattern construction. Attribute selection techniques are used for making the models much easier for users or researchers and minimizing the training time [27].

In this attributed selected technique, data contains many features that are redundant; so the technique removes the repeated features without much loss of information [28].

2.2 Artificial neural networks

Artificial Neural Networks consists of a series of linked neurons that load signals; they belong to Machine Learning Models that are based on the biological human brain. On the other hand, the computational power and the availability of data have been rising through the years, and the researchers have successfully trained the network to perform certain complex tasks as image recognition or speech recognition. This architecture is also called Feed-forward Neural Network.

2.3 Deep learning algorithms

Neural network worked only on one or two layers network, and then deep learning was discovered. This discovery was known as a new area of machine learning research [29]. Deep learning is also part of machine learning. When using the term,

Deep Learning, it may also means deep artificial neural networks [30].

a) Convolutional Neural Networks (CNN)

It is the farthest widespread used algorithm in NLP field. It is used in semantic parsing and sentiment analysis [31].

It performs particularly well on computer vision problems [32]. 1D convolution could be used in extracting local 1D patch (sub sequences) from sequences of image tensors in order to identify an identical transformation to every patch.

b) Recurrent Neural Networks (RNN)

It processes sequences by iterating through the sequence elements and maintaining a state containing information relative to what it has seen so far. So RNN is a kind of neural network that has an internal loop. The state of the RNN is reset between processing two different, independent sequences (such as two different IMDB reviews); still considering one sequence a single data point: a single input to the network. What changes is that this data point is no longer processed in a single step; rather, the network internally loops over sequence elements [32].

RNN is the way to share weights over time. It can be achieved using feed forward networks. It realizes the outputs as new inputs; so it is often known that recurrent networks have memory [30].

C) Long Short-Term Memory (LSTM)

This is an improvement of RNN; it converts existing information entirely by applying a function. It makes small adjustments to the inputs by multiples and additions. The information in a particular cell state contains three different dependencies [33]. These three dependencies are:

- The previous cell state.
- The previous hidden state.
- The input at the current time step.

Embedding

It is popular in NLP to represent words by word embeddings. It is like mapping words with Ids or one-hot encoding vectors. This has many advantages; one of them is being able to deal with the massive vocabulary sets when using NLP; word embedding reduces this dimensionality. Also word embedding would be useful in mapping words that are semantically similar to each other [34].

This is obvious in the following example of sentiment analysis. Given the words: awesome, fantastic, terrific, nice and good; the model can't recognize the difference between these words using one-hot representations to create a model that classifies good and bad reviews. But using word embedding mapping; the words "awesome" and "fantastic" are semantically similar, so they are mapped close to each other in the embedding space; so the good representation of data helps in correctly training the machine learning algorithm.

Word Embedding

Word embedding plays a substantial part in deep learning [35]. The usefulness of word embedding is representing a single word by dimensional vector and relating between two words. Not only does it relate the two words syntactically, but also it relates words with the same meaning (as ‘see’ and ‘watch’ are very different in syntactic, but their meaning is somewhat related). Another benefit is that the algorithm detects the words that appear mostly together (like ‘wear’ and ‘clothes’), and it shows their relationship; and hence this allows predicting the next word [36]. Word embedding is a technique which transforms words to vectors of continuous real numbers (e.g., word "hat" → (… , 0.15, … , 0.23, … , 0.41, …) [35]. The main idea of word embeddings is that words are given to vectors of real numbers, where words that have similar or related meanings are mapped or grouped together to nearby points. In 2013 Word2vec was introduced using neural networks in mapping words with vector representations in NLP tasks such as machine translation. Neural network has become essential to learn about word embeddings. Catching the semantic meaning of words is the main concept of Word2vec where words with likely meanings are given similar or nearby points [31].

Deep learning tools

Deep learning is now one of the most important trends in artificial intelligence and machine learning, here are a few popular tools used in it: [37]

- Theano
- Caffe
- Keras
- Pylearn2
- .Cuda-convnet
- Deeplearning4j

3. Literature review

Social media is a tool that is convenient to express ideas and opinions. J. Eliakin M. de Oliveira et al. [38] apply sentiment analysis and Naïve Bayes algorithm in order to study argumentative individuals on Twitter, Their objective was understanding patterns of changing opinions and the geographical allocation of sentiments whether it is

positive, negative, or neutral. They chose Republican Party candidate, Donald J. Trump, and built their study on this data to find the apportionment of users considering the resemblance of their sentiment, and what clusters they could get, but their problem was the lake of data.

Oberlander J., and Nowson used Naïve Bayes algorithm to classify blog authors into extraversion, agreeableness and conscientiousness using n-grams with accuracy 50% which is very low accuracy [39].

Golbech et al. have predicted the personality traits of 279 users of Facebook using LIWC (Linguistic inquiry and word count) and 279 Twitter users using Gaussian Process [40,41].

Quercia et Al. prophesied personality traits of 335 Twitter users using M5 rules algorithm [42].

Bai et Al. [43] applied decision tree to predict personality traits of 335 RenRen users, which is a popular Chinese social network. The small data set was the main withdraw for them all.

Ion Smeureanu et al., presented their work to explore the positive and negative comments on pre-classified movies' reviews using Naïve Bayes algorithm, which is applied on a collection of comments. Precision and recall methods are used for accuracy check of Naïve Bayes algorithm and its execution time [44, 45].

Tkalcicet et al., proposed a personalized intervention system (PIS) to predict whether a user is going to attend the concert or not through social media text mining, they didn't test the system on real users [46].

Ahmed Hassan et al., presented a method to detect participants' attitude in a reply to others. They used a combination of supervised Markov model of text, part-of-speech and dependency patterns. And they tested the results using Support Vector Machines (SVM) [47].

Pravesh and Mohd [48], used two different data sets: product review dataset and movie review dataset; where the former contains 8000 reviews and the later consists of 2000 reviews. They used clustering techniques to analyze and to compare results from Naïve Bayes, SVM, Clustering classifier and MLP (Multilayer Perceptron) to classify the datasets. They found that SVM is better that the other three techniques using N-gram feature.

Wenling Shang et al. provided a CNN and a comprehensive method in order to make CNN architecture demonstrated clearly. At the beginning, they examined existent CNN models, then they analyzed its rebuilding characteristic in CNNs, and finally they integrated CReLU into several modern CNN architectures. Besides, they demonstrated improvement in execution of lack of parameters in CIFAR-10/100 and ImageNet datasets [49].

Yoon Kim reached superb outcomes on numerous benchmarks by performing a series of experiments using convolutional neural networks, and he showed that a straightforward CNN with one convolutaional layer, little hyper-parameter regulations and constant vectors, perform excellent results on several benchmarks [50].

Ritesh Noothigattu et al., presented an approach that automates decisions. They provided a concrete algorithm that instantiated their approach. Finally, using predilection data collected from 1.3 million people through the Moral Machine website, they carried out and rated a system for moral decision making in the independent vehicle domain, they didn't answer the challenges: What is a good algorithm for generating the mixture of TM models? And, how should such a mixture be? [51].

Biagio Brattoli et al., applied their work on two datasets: Olympic Sports and Leeds dataset. They succeeded in learning the human posture and analyzing the motor kinematics. Testing their model with different number of hidden layers, they got that 512 nodes didn't enhance their final accuracy. But after applying CNN and LSTM together; they got an accuracy of 80.5% [52].

Arthur Toth et al. introduced a new approach that uses a mixture of recurrent Neural Networks to early predict shoppers' behaviors. Their goal was to classify incomplete sequences [53].

Mathieu Cliché made Twitter sentiment classifier using CNN and LSTM for SemEval-2017, they proved that GloVe unsupervised algorithm minimizes the score gotten by both FastText and word2vec [54].

Ye Yuan, You Zhou used the SemEval-2013 data set, which consisted of 6092 rows, divided into 4874 for training and 1218 for testing. They used GloVe word vectors to pre-train tweets, then used RNN for sentiment analysis with one-hidden-layer and two-hidden-layer, and got an accuracy of 84.17% and 80.68% respectively. Their model was a low-accurate model to predict negative label due to the lack of negative labels in the training data. [55].

Martinez-Cámara et al. used an unsupervised polarity classification system for sentiment analysis using twitter based on voting between three lexicon-based sentiment classifiers (negative, neutral, positive) according to the majority of class assigned to the tweet [56].

4. Proposed model

The proposed model aims to recommending an accurate user behavior analysis method according to users' tweets using a voting model of Naïve Bayes, SVM, SMO, Bagging, Attribute Selected and CNN (Fig. 3) and (Fig. 4).

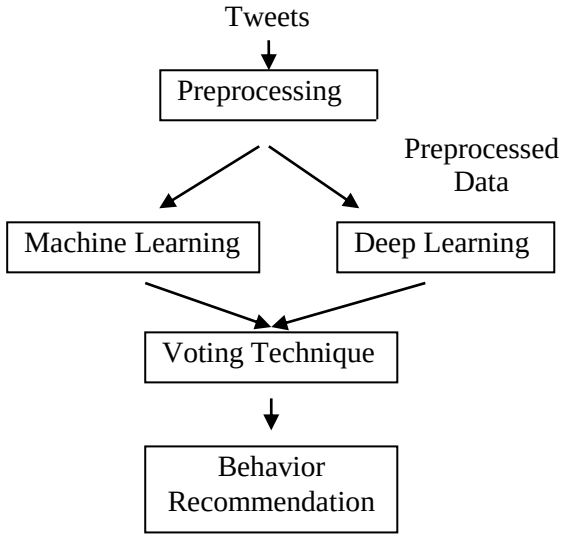

Figure. 3 A proposed model for the voting system

Algorithm 1 proposed model algorithm

Input: training dataset, testing dataset
Output: dataset with 3 cluster labels
1: procedure Naïve Bayes
2:initialisation i=1
LOOP Process
3: **while** i < N **do**
4: $P(C_K \mid X) = P(C_K)\, P(X \mid C_K)$ / $P(X)$
5: **for** each $P(C_K|X)$ **do**
6 **if** $P(C_1|X) > P(C_2|X) >$ $P(C_3|X)$ **then**
7: Assign tweet to cluster C1
8: **if** $P(C_2|X) > P(C_1|X) >$ $P(C_3|X)$ **then**
9: Assign tweet to cluster C2
10: **if** $P(C_3|X) > P(C_2|X) >$ $P(C_1|X)$ **then**
11: Assign tweet to cluster C3
12: procedure SVM
13:initialisation i=1
LOOP Process
14: **while** i < N **do**
15: $F(x) = \beta_0 + \beta^T x$,
find nearest cluster to the point in the dimensional space
16: procedure SMO
17:initialisation i=1
LOOP Process
18: **while** i < N **do**
19: Find a Lagrange multiplier α1 that violates the Karush– Kuhn–Tucker (KKT) conditions for the optimization problem,
Pick a second multiplier α2 and optimize the pair (α_1, α_2)
Repeat previous two steps until convergence
20: procedure Bagging
21: Create many random sub-samples of our dataset with replacement
22:initialisation i=1
LOOP Process
23: **while** i < M **do**
Calculate the mean of each sub-sample
Calculate the average of all of our collected means and use that as our estimated mean for the data
24: procedure Attribute Selected
25: select necessary attributes only
26:initialisation i=1
LOOP Process
27: **while** i < N **do**
classify the given tweets according to given attributes
28: initialisation i=1
LOOP Process
29: **while** i < N **do**
vote between output of the five algorithms
30: **for** each i **do**
31: **if** Max of 5 Algorithms output is $_{C1}$ **then**
32: Assign tweet to cluster C1
33: **if** Max of 5 Algorithms output is $_{C2}$**then**
34: Assign tweet to cluster C2
35: **if** Max of 5 Algorithms output is $_{C3}$ **then**
36: Assign tweet to cluster C3
37:**return** behavior

Figure. 4 Algorithm of proposed model for the voting system

4.1 Data set

The dataset used as an input for experiments consisted of tweets acquired from Twitter as strings and the classification of them, so there are two attributes:

1- Classify
2- String

In total, the training dataset contains 66731 records (tweets) [each user has 10 tweets with predicted behaviour] and a testing dataset contains 12500 records (tweets) [57].

Second step, after gathering the data, it had to be pre-processed in order to make sure that it didn't contain any dummy characters and strings; which might be difficult for an algorithm to work with. So

Table 1. Techniques comparison

Reference no.	Dataset size	Technique	Our system
21	Small (twitter)	Detect sentiment from behavior	Detect behaviour from sentiment
40,41	279 (twitter)	Predict personality traits using LIWC	Predict personality traits using SMO, SVM, Naive Bayes, Bagging, Attribute selected, CNN, RNN for large twitter dataset (training 66731 records and a testing dataset contains 12500)
39	(twitter)	used Naïve Bayes algorithm to classify blog authors into extraversion, agreeableness and conscientiousness using n-grams	
42	335 (twitter)	prophesied personality traits using M5 rules algorithm	
43	335 (RenRen)	applied decision tree to predict personality traits	
48	8000 (product review dataset) and 2000 (movie review dataset)	Analyze and to compare results from Naïve Bayes, SVM , Clustering classifier and MLP (Multilayer Perceptron) to classify the datasets. They found that SVM is better that the other three techniques using N-gram feature	Analyse and compare prediction results from machine learning (SMO, SVM, Naive Bayes, Bagging, Attribute selected)
53		Early predict shoppers' behaviors. In order to classify incomplete sequences using new approach which is a mixture of recurrent Neural Networks	Analyse and compare prediction results from deep learning techniques (CNN, RNN)
54	SemEval-2017 (Twitter)	Twitter sentiment classifier using CNN and LSTM	
52	Olympic Sports and Leeds dataset	learn the human posture and analyze the motor kinematics. After applying CNN and LSTM together; they got an accuracy of 80.5%	Analyse and compare prediction results from hybrid used approaches together
56	SemEval-2014 (Twitter)	voting strategy of three lexicon-based sentiment classifiers	Voting technique

working on this data was a must to make it suitable as an input for the model. Pre-processing the data gathered included the following:

1- Removing URLs, Hash tags, mentions, special characters and emotions.
2- Eliminating citations.
3- Tokenization.

Afterwards, the steps of applying algorithm and evaluating came into process. These steps could be summarized as follow:

1- Training data set that contained tweets that had been collected and grouped as class Extraversion, Neuroticism, and Psychoticism, then come the process of accuracy measurement for the output.
2- Tokenization: is the process of separating words (known as tokens) formulating the tweets.
3- Stopping word removing: is the process of removing words that are not affecting the meaning of the sentences, based on a list defined.
4- Applying a classification technique in order to classify to which class/ category the tweet belongs to.

4.2 Classification of a tweet using classification techniques

Table.2 is a summary for the used techniques and the accuracy of each. Fig. 5 is the graph representation of table.

So concluded from the previous table, SVM algorithm is better in detecting the Psycoticim behavior class, Naïve Bayes algorithm is better to detect the Neuroticism behavior class, and finally SMO algorithm is better for Extraversion.

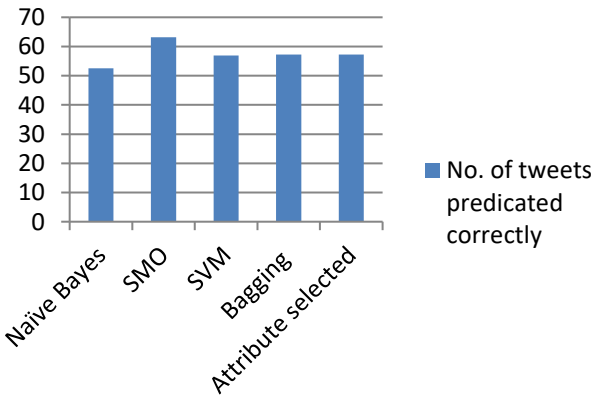

Figure. 5 Chart of accuracy

Table 2. Used machine learning techniques accuracy

Tweet no.	Actual	Naïve Bayes	SMO	SVM	Bagging	Attribute selected
1	P	P	P	P	P	P
3	E	E	E	E	E	E
5	E	P	E	E	N	N
6	P	P	P	P	E	E
16	N	E	E	E	E	E
109	N	E	P	E	E	E
…	…	…	…….	…….	…….	…….
Accuracy	**-**	**52.568 %**	**63.168 %**	**56.96 %**	**57.264 %**	**57.264 %**

Table 3. Capability of used machine learning techniques to predict behaviors

Classes	Test data set	SMO	Naïve Bayes	SVM	Bagging	Attribute Selected
Psychoticism	5387	2921	3193	3364	3039	3039
Neuroticism	1741	546	767	0	348	348
Extraversion	5371	4428	2610	3355	3770	3770

The results were tested and recorded for evaluation of accuracy and efficiency of that model. This research used to analyze the accuracy for the technique by comparing the result with a predefined dataset for test.

After applying the five previously mentioned machine learning algorithms, it was noticed that SMO algorithm scored the highest accuracy among all used algorithms. Then in Table 4, the algorithms were combined in the voting model to get the highest possible accuracy.

Table 4. Accuracy of experiments

Voting	Algorithm	Accuracy
3 algorithms (ML)	SMO Bagging Attribute Selected	57.264 %
	Naïve SVM Bagging	57.36 %
	SMO Naïve SVM	60.144 %
	SMO SVM Bagging	60.896 %
	SMO Naïve Bagging	61.568 %
5 algorithms (ML)	All	57.264 %
Deep Learning	CNN	73.68 %
	RNN	73.512 %
Hybrid	(CNN-SMO-SVM-Naïve-Bagging)	66.50 %
	(CNN-SMO-SVM)	65.848 %
	(RNN-CNN-SMO)	84.208 %

After applying the voting model on 3 different algorithms and 5 algorithms using machine learning; the accuracy was between 57.264 % and 61.568% which is less that SMO algorithm (accuracy 63.168 %), so start conducting CNN and RNN on the dataset, accuracy increased to 73.68% and 73.512% respectively. But after hybrid CNN with SMO-SVM-Naïve-Bagging and CNN with SMO, SVM accuracy was consecutively 66.50 % and 65.848 %.

It was obvious that the accuracy decreased after the hybrid, so using the voting technique between RNN, CNN and SMO to get a high accuracy was a must. The voting technique system mainly compares the results from CNN and SMO to predict the behavior. If both techniques predict the same behavior, then the result is verified. But if the prediction is different; the result is based on RNN technique. After applying this voting technique; the accuracy increased to 84.208%.

5. Conclusion

Before employing any employee in a company, it is imperative to record his/her behavior over a

certain time period and over various social media sites. This can be done through clustering and analyzing sentiments in order to predict his/her real motifs, values, beliefs, or behavior. This paper shows what behavior analysis means, what Enrick Three Factor Model is, its dimensions, and how researchers were concerned with personality and behavior analysis through social media recently. The study also presented the classification techniques using machine learning and deep learning. The researcher conducted these techniques and compared the results. At first, the accuracy was between 57.264% and 61.568% for machine learning techniques. Then after applying CNN and RNN algorithms, the highest accuracy reached was 73.68%, while hybrid CNN with SMO-SVM-Naïve-Bagging reached 66.50%. The accuracy, using CNN with SMO and SVM, was 65.848%. The voting between RNN, CNN and SMO resulted in 84.208% of accuracy. The challenge is to apply more sophisticated techniques and a bigger dataset on the model to show if the results obtained from this model will be more accurate or the previous models are better.

References

[1] M. Chetan and P. Mulay, "E3: effective emoticon extractor for behavior analysis from social media", *Procedia Computer Science*, Vol. 50, pp. 610-616, 2015.

[2] E. Alpaydin, "Introduction to Machine Learning. [Sl]", pp. 249-256, 2010.

[3] Q. T. Ain, M. Ali, A. Riazy, A. Noureenz, M. Kamranz, B. Hayat, and A. Rehman, "Sentiment Analysis Using Deep Learning Techniques: A Review", *Int. J. Adv. Comput. Sci. Appl.*, Vol. 8, No. 6, pp.424, 2017.

[4] C. Lee, "An information-theoretic filter approach for value weighted classification learning in naive Bayes", *Data & Knowledge Engineering*, pp.116-128, 2018.

[5] J. Platt, "12 fast training of support vector machines using sequential minimal optimization", *Advances in kernel methods*, pp. 185-208, 1999.

[6] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design", *Neural Computation*, Vol 13, No. 3, pp. 637-649, 2001.

[7] C. Sutton, "Classification and regression trees, bagging, and boosting", *Handbook of statistics*, Vol 24, pp.303-329, 2005.

[8] S. Dinakaran and P. Thangaiah, "Role of attribute selection in classification algorithms", *International Journal of Scientific & Engineering Research*, Vol 4, No. 6, pp. 67-71, 2013.

[9] M. Day and C. Lee, "Deep Learning for Financial Sentiment Analysis on Finance News Providers", *In Advances in Social Networks Analysis and Mining*, IEEE, No. 1, pp. 11271134, 2016.

[10] Y. Zhang, M. J. Er, R. Venkatesan, N. Wang, and M. Pratama, "Sentiment Classification using Comprehensive Attention Recurrent models", *International Joint Conference on Neural Networks*, pp. 1562–1569, 2016.

[11] J. Lee, S. Jun, Y.Cho, H. Lee, G.Kim, J. Seo, and N. Kim, "Deep learning in medical imaging: general overview", *Korean journal of radiology*, Vol 18, No. 4, pp.570-584, 2017.

[12] L. Deng and D. Yu, "Deep learning: methods and applications", *Foundations and Trends® in Signal Processing*, Vol 7, No. 3–4 pp.197-387, 2014.

[13] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, MF. Moens, and M. De Cock, "Computational personality recognition in social media", *User modeling and user-adapted interaction*, Vol 26, No.2-3, pp.109-142, 2016.

[14] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell. "Personality and patterns of Facebook usage", In: *Proc of the 4th Annual ACM Web Science Conference*, pp. 24-32, 2012.

[15] A. Poropat, "A meta-analysis of the five-factor model of personality and academic performance", *Psychological Bulletin*, Vol. 135, No. 2, pp.322.2009.

[16] R. Favaretto, L. Dihl, S. Musse, F. Vilanova, and A. Costa, "Using big five personality model to detect cultural aspects in crowds", In: *Proc. of the 30th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 223-229, 2017.

[17] T. Buchanan, J. Johnson, and L. Goldberg, "Implementing a five-factor personality inventory for use on the internet", *European Journal of Psychological Assessment*, Vol21, No. 2, pp115-127, 2005.

[18] J. Shropshire, M. Warkentin, A. Johnston, and M. Schmidt, "Personality and IT security: An application of the five-factor model", In: *Proc. of AMCIS 2006*, pp.415, 2006.

[19] A. Terracciano and P. Jr., "Smoking and the Five-Factor Model of personality", *Addiction*, Vol. 99, No. 4, pp. 472-481, 2004.

[20] D. Sewwandi, K. Perera, S. Sandaruwan, O. Lakchani, A. Nugaliyadde, and S. Thelijjagoda,

"Linguistic features based personality recognition using social media data", In: *Proc. of Technology and Management, National Conference*, pp. 63-68, 2017.

[21] F. Celli and C. Zaga, "Be conscientious, express your sentiment!", *Training*, Vol. 5747, No.495, pp.5252, 2013.

[22] S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach", *Prentice Hall*, pp.178, 2003.

[23] M. Murty and V. Devi, "Pattern recognition: An algorithmic approach", *Springer Science & Business Media*, 2011.

[24] C. Cortes and V. Vapnik, "Support-vector networks", *Machine Learning*, Vol 20, No. 3, pp.273-297, 1995.

[25] J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", 1998.

[26] L. Breiman, "Bagging predictors", *Machine Learning*, Vol 24, No. 2, pp.123-140, 1996.

[27] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An introduction to statistical learning", *Springer Texts in Statics*, p. 204, 2013.

[28] M. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. Wright, J. Wilson, F. Agakov, P. Navarro, and C. Haley, "Application of high-dimensional feature selection: evaluation for genomic prediction in man", *Scientific Reports*, Vol. 5, p.10312, 2015.

[29] T. Du and V. Shanker, "Deep Learning for Natural Language Processing", 2016.

[30] J. Schmidhuber, "Deep learning in neural networks: An overview", *Neural Networks*, Vol. 61, pp.85-117, 2015.

[31] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review", *EURASIP Journal on Wireless Communications and Networking*, Vol. 2017, No. 1, p.211, 2017.

[32] F. Chollet, "Deep learning with python", *Manning Publications Co*., 2017.

[33] B. Brattoli, U. Büchler, A. Wahl, M. Schwab, and B. Ommer, "Lstm self-supervision for detailed behavior analysis", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. 2017.

[34] D. Santolaya, "Using recurrent neural networks to predict customer behavior from interaction data", *University of Amsterdam*, 2017.

[35] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018

[36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", *In Advances in Neural Information Processing Systems*, pp. 3111-3119. 2013.

[37] www.kdnuggets.com

[38] J. de Oliveira, M. Cotacallapa, W. Seron, R. Santos, and M. Quiles, "Sentiment and Behavior Analysis of One Controversial American Individual on Twitter", In: *Proc. of International Conference on Neural Information Processing*, pp. 509-518, 2016.

[39] S. Nowson and J. Oberlander, "Identifying more bloggers", In: *Proc. of ICWSM*, 2007.

[40] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media", In: *Proc. of CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 253-262, 2011.

[41] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter", In: *Proc. of Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing*, pp. 149-156. 2011.

[42] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter", In: *Proc. of Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing*, pp. 180-185, 2011.

[43] S. Bai, T. Zhu, and L. Cheng, "Big-five personality prediction based on user behaviors at social network sites", *arXiv preprint arXiv:1204.4809*, 2012.

[44] I. Smeureanu and C. Bucur, "Applying supervised opinion mining techniques on online user reviews", *Informatica Economica*, Vol 16, No. 2, p.81, 2012.

[45] H. Wang and S. Wang, "A knowledge management approach to data mining process for business intelligence", *Industrial Management & Data Systems*, Vol. 108, No. 5, pp. 622-634, 2008.

[46] M. Tkalcic, B. Ferwerda, M. Schedl, C. Liem, M. Melenhorst, A. Odic, and A. Kosir, "Using social media mining for estimating theory of planned behaviour parameters", In: *Extended Proceedings of the Conference on User Modelling, Adaptation and Personalization*, 2014.

[47] A. Hassan, V. Qazvinian, and D. Radev, "What's with the attitude?: identifying sentences with attitude in online discussions",

In: *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1245-1255, 2010.

[48] P. Singh and M. Husain, "Methodological study of opinion mining and sentiment analysis techniques", *International Journal on Soft Computing*, Vol 5, No. 1, p. 11, 2014.

[49] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units", In: *Proc. of International Conference on Machine Learning*, pp. 2217-2225, 2016.

[50] Y. Kim, "Convolutional neural networks for sentence classification", *arXiv preprint arXiv:1408.5882*, 2014.

[51] R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. Procaccia, "A voting-based system for ethical decision making", *arXiv preprint arXiv:1709.06692*, 2018.

[52] B. Brattoli, U. Büchler, A. Wahl, M. Schwab, and B. Ommer, "Lstm self-supervision for detailed behavior analysis", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. 2017.

[53] A. Toth, L. Tan, G. Fabbrizio, and A. Datta, "Predicting Shopping Behavior with Mixture of RNNs", In: *Proc. of the SIGIR 2017 Workshop on eCommerce*, 2017.

[54] M. Cliche, "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs", *arXiv preprint arXiv:1704.06125*, 2017.

[55] Y. Yuan and Y. Zhou, "Twitter sentiment analysis with recursive neural networks", *CS224D Course Projects*, 2015.

[56] E. Martinez-Cámara, S. Jiménez-Zafra, M. Martin, and L. Lopez, "SINAI:Voting system for twitter sentiment analysis", In: *Proc. of the 8th International Workshop on Semantic Evaluation*, pp. 572-577, 2014.

[57] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter", In: *Proc. of the 11th International Workshop on Semantic Evaluation*, pp. 502-518. 2017.