Comparative Study of Personality Prediction From Social Media by using Machine Learning and Deep Learning Method

Thahira M
Dept. of Computer Science & Engineering
MGM College of Engineering and
Pharmaceutical Sciences, Valanchery,
Kerala, India

Mubeena A K
Dept. of Computer Science & Engineering
MGM College of Engineering and
Pharmaceutical Sciences, Valanchery,
Kerala, India

***Abstract*: The social media networks are an online forum that is used to improve social relationships with others by allowing people to share their thoughts, emotions, and experiences, among other things. The use of social media networks has skyrocketed in recent years. Predicting personality traits from social media networks has become a difficult challenge. This proposed approach uses social media networks to predict a person's personality. The big-five-factor model (OCEAN) for defining personality is used in this project, which includes openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). The classification is done using machine learning and deep learning neural networks as classifiers. Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT) are machine learning classifiers and LSTM is deep learning classifier to predict a person's personality. We compare machine learning and deep learning personality prediction in this project. Finally, it was discovered that deep learning is more accurate than machine learning.**

Keywords: Support vector machine (SVM), Naive Bayes(NB), Random Forest(RF), Decision Tree(DT), LSTM.

I. INTRODUCTION

The use of social media networks has skyrocketed in recent years. These social media networks are used to create social networks and relationships between people. Facebook, Twitter, Google+, and Instagram are some of the most well-known social media sites. These websites are referred to as entities, and each entity is linked to others as friends,, followers and so on. The popularity of these sites is growing by the day, owing to the ease with which they can be accessed from anywhere in the world and their user- friendly interfaces. We can start communicating with others in a short amount of time by using these social media channels. Many things, such as posting statuses, sharing other people's posts, supporting other people's posts, commenting on other people's posts, communicating directly with friends, and playing online games with friends, are available to users when using these services.

In this project, we predict user's personalities from these digital social media platform [2][6] and it is a challenging task.User's behaviour is different in social media and real life. In the social media, User generate content like status updates, post, comments etc. This content provides the reflection of user's personality like his/her current situation, culture, identity [5] political interest etc. We can identify the user's behaviour based on these personality traits. We are using five different personality traits for defining user's personality that is widely used big-five-factor model (BFFM) [7]. It include four positive personality traits, namely, openness-to-experience(O), conscientiousness (C), extraversion (E), agreeableness (A) and the only one negative personality trait is neuroticism (N). This BFFM model is also known as OCEAN model.

- **Open to Experience:** It encompasses a number of characteristics, including creativity, sensitivity, attentiveness, a preference for variety, and curiosity.
- **Conscientiousness:** This personality trait is used to characterise a person's attention to detail and diligence. It's a characteristic that defines how well-organized and productive someone is.
- **Extraversion:** It is the personality trait that explains how well a candidate can communicate with others, or how strong his or her social skills are.
- **Agreeableness:** It is a personality trait that assesses an individual's generosity, compassion, cooperativeness, and willingness to respond to others.
- **Neuroticism:** This personality trait defines someone who is moody and has a lot of expressive ability.

Feature extraction and feature selection methods are applied for extract the most relevant features. By using these relevant features training and testing are performed. Relevant feature extraction method is one of the challenging tasks, in order to get better accuracy over the prediction

The Main Contribution of this project is:

1. For defining personality we are using big-five-factor model(BFFM) such as openness-to-experience(O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N)
2. Five different machine learning classifiers are

used for training and testing such as Support Vector Machine(SVM),Naive Bayes(NB), Random Forest (RF), Simple Logistic Regression (SLR) and Decision Tree (DT)

3. Deep learning classifier such as Long Short Term Memory (LSTM) is used for the training and testing and it get better accuracy over the prediction.

II. RELATED WORKS

A. Predicting Facebook-Users' Personality based on Status and Linguistic Features via Flexible Regression Analysis Techniques

Here, machine learning methods are applied to the personality prediction namely LR, SVR with linear kernel, SVR with polynomial kernel and SVR with RBF kernel. SVR is a widely used regression technique which is based on Support Vector Machine. By use this method accuracy will be very less. Linguistic Inquiry and Word Count (LIWC) tool and Latent Dirichlet Allocation (LDA) are used to extract facebook user's statuses and posts.

B. Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction From Social Media

Here, Five different classifiers, namely, Naive Bayes(NB), decision tree (DT), random forest (RF), simple logistic regression (SLR), and support vector machine (SVM) for the training and testing of classification model. These five classifiers are machine learning classifiers. By using these classifiers the accuracy of prediction will be less compare to Deep learning method.

III. PROPOSED SYSTEM

In this project, Predict personality traits from social media sites by using two different classifiers namely machine learning and deep learning classifiers. Machine learning classifiers are support vector machine, naive bayes, random forest, simple logistic regression and decision tree. Deep learning classifiers are Long short term memory and multilayer perceptron. Here we compare accuracy of both classification method.

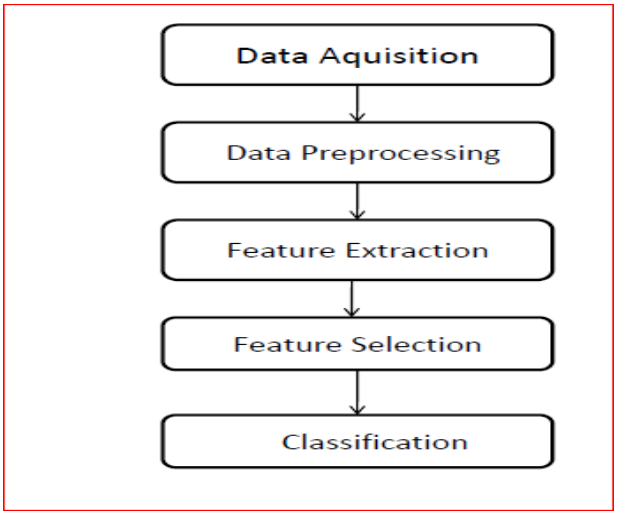

Fig.1 Step of the machine learning experimental method.

A. Data Acquisition

In this experiment, we have used the *myPersonality* dataset [3], [4] like https://www.16personalities.com/ sites for data collection that consist of status updates, social network features, comments, article etc. The traits of the data sets are formalized in big-five-factor model. The five personality traits are openness-to-experience (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N). Each BFFM, the personality score and class value (yes or no) are given in the dataset. This dataset contain 2500 data for personality prediction. 90 percentage of datas are taken as training and remaining 10 percentage is used for testing.

B. Data Preprocessing

In the dataset, all the statuses are in English and it follow every preprocessing step. The preprocessing step consist of removal of unnecessary details such as URLs, symbols, unnecessary spaces and stemming.

C. Feature Extraction

Feature extraction is the one of the important step for find out the most relevant features. The extracted features are two types that is linguistic features and social network features. Here we extracted both psycholinguistic features and traditional linguistic features. LIWC (Linguistic inquiry and word count) [1] is applied on preprocessed textual data for extracting linguistic features.

D. Feature Selection

Feature selection algorithms are used for find the relevant features. There are three different types of algorithms are used here for the feature selection namely pearson correlation coefficient (PCC), information gain (IG), chi-squared (CHI) method.

E. Classification Method

Here, we have applied two classic classification methods for evaluate the performance of this project. The different classification methods are machine learning techniques and deep learning techniques.

a) By using Machine learning techniques

Machine learning classification is the prediction of class from the given dataset. Here the classes are

personality traits, it include positive and negative traits. In this project, five different machine learning classifiers are used for classification such as Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) and Decision Tree (DT).

- Support Vector Machine

For two-group classification problems, a support vector machine (SVM) is a supervised machine learning model that uses classification algorithms. SVM models will categorised new text after being given sets of labelled training data for each group. SVM is a fast and dependable classification algorithm that does a great job with a small amount of data. The SVM algorithm is simple in concept, and applying it to natural language classification doesn't require much in the way of technical knowledge.

- Naive Bayes

The Bayes theorem inspired Naive Bayes, a probabilistic classifier that works under the simple assumption that the attributes are conditionally independent. Naive Bayes is a very easy algorithm to implement, and it has shown successful results in the majority of cases. Since it takes linear time rather than the costly iterative approximation used by many other types of algorithms, it can easily scale to larger datasets. The zero probability problem may be a problem with naive Bayes. When the conditional likelihood for a given attribute is zero, the prediction is invalid. Using a Laplacian estimator, this must be set specifically.

- Random Forest

Random forest is a versatile, easy-to-use machine learning algorithm that, in most cases, produces excellent results even without hyper-parameter tuning. Because of its simplicity and versatility, it is also one of the most widely used algorithms (it can be used for both classification and regression tasks). It is based on ensemble learning, which is a method of combining multiple classifiers to solve a complex problem and improve the model's accuracy.

- Decision Tree

The classification technique is a method for creating classification models from a collection of data. Different techniques to solve a classification problem include decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers. Each technique uses a learning algorithm to find the model that best matches the relationship between the input data's attribute set and class mark. As a result, one of the learning algorithm's primary goals is to create a predictive model that can reliably predict the class labels of previously unknown records.

A basic and commonly used classification technique is the Decision Tree Classifier. To solve the classification problem, it employs a straightforward approach. The Decision Tree Classifier presents a series of well-crafted questions about the test record's attributes. Each time it receives a response, it asks a follow-up question until a conclusion about the record's class label is reached.

b) By using deep learning technique

An artificial neural network with many layers between the input and output layers is known as a deep neural network. Each mathematical operation is referred to as a layer, and complex DNNs have several layers, hence the term "deep" networks. DNNs are capable of modeling non-linear relationships that are complex. Here Long short term memory (LSTM) is used as the deep learning classifier.

- Long Short Term Memory (LSTM)

Long Short Term Memory Networks (LSTMs) are a form of RNN that can learn long-term dependencies. Hochreiter & Schmidhuber (1997) introduced them, and many people improved and popularised them in subsequent work. They work incredibly well on a wide range of problems, and are now widely used.

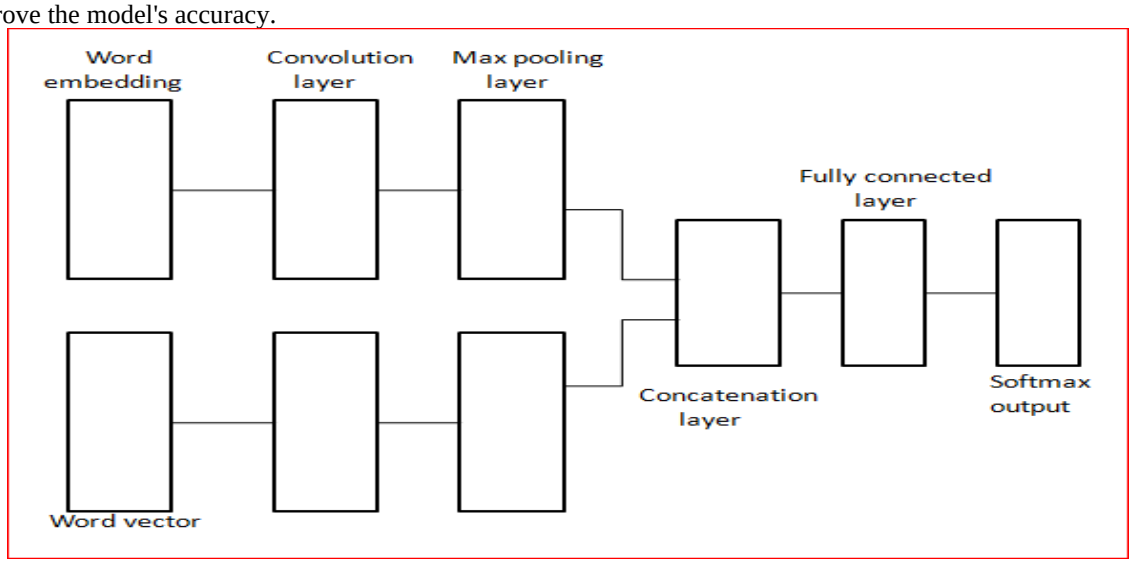

Fig.3. Network architecture of deep learning.

Here, the network include seven layers: input-word embedding(word vectorization), convolution(sentence vectorization), max pooling(sentence vectorization), concatenation(document vectorization), Fully connected layer (classification), and two neuron softmax output (classification).

- **Input**

Here, we have dataset like a set of documents that referred as input. Each document represented as d and it is a sequence of sentences. Each sentence is represented as *s* and it is a sequence of words, and each word *w* is a real- valued vector of fixed length known as word embedding. In our experiments, Google's pretrained word2vec embeddings is used. Our input layer is a four dimensional real-valued array. To ensure that all documents had the same number of sentences, we padded shorter documents with dummy sentences during implementation. Similarly, we used dummy words to fill in the gaps in shorter sentences. To extract unigram, bigram, and trigram features from each sentence, we use three convolutional filters after max pooling, the sentence vector is a concatenation of the feature vectors obtained from these three convolutional filters.

- **Convolution layer**

We concatenate the vectors for the three forms of n-grams to get the vector s, which represents the sentence. Each sentence in the document is subjected to convolution and max pooling. All of the document's

sentences share the same network parameters. In particular, although we pad all sentences to a common size with dummy words, we do not need to pad all documents to a common size with dummy sentences.

- **Classification**

For classification, we use a two layer perceptron consisting of a fully connected layer of size 200 and the final softmax layer of size two, representing the *yes* and *no* classes.

IV. RESULT AND ANALYSIS

Here, we have five personality traits for predicting personality namely, openness-to-experience (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N). In this experiment we have focused on deep learning and machine learning method for classification. By using machine learning method SVM classifier shows the highest accuracy nearly 58% for the openness-to-experience. The remaining personality traits value reported in Table 1.

By using the deep learning method, we have better accuracy than machine learning classifier nearly 80%.

Table I

Traits	NB	DT	RF	SVM	LSTM
OPN	0.5203	0.4725	0.5532	0.6108	0.8523
CON	0.4472	0.5321	0.4871	0.5407	0.8264
EXT	0.4675	0.4821	0.4965	0.5732	0.8144
AGR	0.5447	0.5132	0.5384	0.4756	0.8312
NEU	0.5285	0.5732	0.4658	0.5041	0.8265

V. CONCLUSION

Here, we have presented a comparative study of user's personality prediction from social media. For predicting personality we have used positive and negative traits. There are two types of classification techniques are used here for identifying the personality such as machine learning and deep learning neural network techniques. By using these deep learning techniques such as LSTM, we can get better accuracy over the machine learning techniques. Here we can predict more accurate user's personality from their own social media activities such as their statuses, comments, post, tweets etc.

VI. REFERENCES

[1] J. W. Pennebaker, M. E. Francis, and R. J. Booth. (2001). Linguistic Inquiry and Word Count: LIWC2001. Erlbaum, Mahwah, NJ, USA. [Online]. Available: https://www.erlbaum.com

[2] M. M. Hasan, N. H. Shaon, A. A. Marouf, M. K. Hasan, H. Mahmud, and M. M. Khan, "Friend recommendation framework for social networking sites using user's online behavior," in *Proc. 18th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2015, pp. 539–543.

[3] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines," *Amer. Psycholog.*, vol. 70, no. 6, pp. 543–556, Sep. 2015.

[4] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 5802–5805, Apr. 2013.

[5] C. P. Williams. (Feb. 23, 2013). *Language, Identity, Culture, and Diversity*.[Online].Available:https://www.newamerica.org/educationpolicy/ edcentral/ multilingualismmatters/

[6] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *Proc. Extended Abstracts Hum. Factors Computing Syst.*, Vancouver, BC, Canada, May 2011, pp. 253–262.

[7] L. R. Goldberg, "The development of markers for the big-five factor structure," *Psychol. Assessment*, vol. 4, no. 1, pp. 26–42, 1992, doi: 10.1037/1040-3590.4.1.26.