Methodology

Proposed Work

Machine Learning
It provides the facility to learn and improve from experience with no involvement, support or human intervention and adjust actions accordingly [5]. The development of learning starts with observations or data, such as understanding, training, examples in order to look for patterns in data and make enhanced decision in the future. Machine learning uses various methods to analyze data and broadly grouped into two types supervised and unsupervised.

In Supervised learning trained examples are used to make predictions. The training dataset includes the input data and their desired output values. From this, it can make predictions of the given values for a new examples. Supervised learning contains two types of algorithms i.e. Classification and Regression.

Unsupervised learning employ dataset without previous training. Cluster analysis is the most familiar unsupervised learning method. On the basis of similarity evaluated using metrics like Euclidean or else probabilistic distance the data set is grouped into various clusters.

Need of Machine Learning for analyzing Social Network data
Social security has provided researchers with countless new insights into the inner workings of the human nature. The statistical analysis of social networking data is very difficult. The data is enormous in volume comprising different types of data. These data are heterogeneous and noisy due to possible redundancy and lack of proper technology. The traditional statistical data analytic solutions are not efficient to draw effective conclusions from such high-resolution measurements. In machine learning approach the dimensionality of the data is effectively reduced by removing redundancy.

Learning Methods
We explore here two classifier training methods Logistic Regression and Support Vector Machine

Logistic Regression
Logistic Regression is the regression problem, with a small number of distinct values for prediction. Classification that can accept only two values i.e 0 and 1 will be discussed here.

We could move toward the classification without paying attention the detail that \( y \) is discrete-valued, and apply previous linear regression to attempted to calculate \( y \) given \( x \). we know that \( y \in \{0, 1\} \) and thus \( h_\theta(x) \) must take values between 0 and 1.
For this purpose, we will opt
\[
h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}
\]
where
\[ g(z) = \frac{1}{1 + e^{-z}} \]
is referred as the logistic or sigmoid function.

\( g(z) \) tends towards 1 as \( z \to \infty \), and \( g(z) \) tends towards 0 as \( z \to -\infty \). Moreover, \( g(z) \), and hence also \( h(x) \), is always bounded between 0 and 1. As before, we are keeping the convention of letting \( x_0 = 1 \), so that
\[
\theta^T x = \theta_0 + \sum_{i=1}^{n} \theta_i x_i
\]
The classification model is endowed with a number of probabilistic assumptions to fit \( \theta \) for it, and then the parameters are fitted using maximum likelihood.

Let us assume that
\[
P(y = 1 \mid x; \theta) = h_\theta(x) \\
P(y = 0 \mid x; \theta) = 1 - h_\theta(x)
\]
This can be presented more efficiently as
\[
p(y \mid x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}
\]
**Support Vector Machine**
SVM is considered as one of the most important methods for classification and linear kernel[4, 6], is used for training SVM. The support vector machine is a generalization of maximal margin classifier, it make use of a hyperplane for segregating two types of data. The maximal margin hyperplane outermost from the training examples is selected. The distance from each training example to a given segregating hyperplane, can be calculated. The margin is the minimal distance from the examples to the hyperplane. These two hyperplanes can be represented as given below:
\[
x.y - b = 1 \text{and} \\
x.y - b = -1
\]
Obviously, in order for \( \frac{2}{||w||} \) to get maximized the value of \( w \) should be minimum.

**Feature Selection**
The feature selection methods enumerated below are employed independently and as well as collectively in order to compare performance. All features present in the data are used to create sample data which is used for training and testing different learning models. A training example may have a number of features. As a result dimension of the feature vector can be very high.

This approach is very useful while data is assembled from diverse subjects. Features are chosen on the basis their capability to differentiate target method using F-ratiois employed to find out the involvement of every feature inside the feature set.

**Feature Standardization**
The common process of calculation is to found the distribution mean and standard deviation meant for each of the feature. Then the mean is deducted from every feature. after that the values of every feature is divided by its standard deviation.

\[ x' = \frac{x - \bar{x}}{\sigma} \]

Where \( x \) represent the original feature vector, \( \bar{x} \) represent the mean of that feature vector moreover \( \sigma \) represents its standard deviation.

We will use data in learning models using without standardization, means as we collected data from social networking data and with standardization, means scaling the data using standardization method. We present the results of both cases in different experiments.
Chapter 6
Work Plan

In Introduction, to introduce the scope of summarization, summarization methods and clustering technique, how these methods and techniques are used for better summarization.

In literature review section all the existing work is studied in depth. Much of the past research on identity deception among humans has been psychological in nature. We have found that the research work based on machine learning has more accuracy than the other related works.

In research objective to analyze all the existing research, compare each system with other to see the different result and basis on that result develop the new system.

The next is the methods are used in existing system and using now for developing or upgrading new system. Machine learning based approach or artificial analysis can provide better result based on content.

Finally the testing of system and result analysis will be done for use.