CHAPTER 5
RESULTS AND DISCUSSION

This chapter is dedicated to the results obtained from the experiments conducted based on the research methodology. For conducting experiments, the research methodology, we have taken four different Tweets datasets viz. Girl-Child, Intolerance, JNU, Price rise, which are relevant to the social impact stand point as is the title of this research.

5.1 PHASE 1: COLLECTING RAW DATA

For the analysis of social issues tweets, data is collected. For creating the corpus of tweets, the tweets are fetched from the Twitter database based on HashTags(#), using Twitter API for connecting and authenticating. The collected text is noisy and methods for cleaning and parsing of the data to form a corpus for further processing. All the datasets fetched from Twitter are cleaned and processed for obtaining results; one of the datasets is attached as ANNEXURE 1 at the end of this report for reference.

5.1.1 TOOL USED

The reenactment has been performed in Java Net Beans. NetBeans is an open-source venture devoted to providing rock strong programming enhancement objects (the NetBeans Platform and the NetBeans IDE) that address the requirements of designers, organizations and the clients who depend on NetBeans as a cause for their objects; specially, to empower them to create up these objects rapidly, effectively and effortlessly by using the qualities of the Java phase and other important industry gauges.

![Fig. 5.1 NetBeans IDE.](image-url)
The following programming libraries are used in the course of this application:

- **Weka**: Weka is a data mining tool. Weka.jar file is used with a purpose of utilizing the predefined methodologies in the research work.
- **AbsoluteLayout**: This is used to make the graphical user interface layout more attractive.
- **JFreeChart**: JFreeCart library is used for making the graphs or charts in java.
- **Mysqlconnector**: This library is used to interface the mysql database with the java code.

### 5.1.2 DATASET UPLOADING

This is the first window we come across, here we select the files that our dataset have that are preloaded when we created database at the backend. This interface helps us to choose the desired data set from any location and upload that data set.

![Dataset Upload Window](image)

**Fig. 5.2 Choosing and showing the content of the dataset.**

Figure 5.2 above presents the way of uploading various datasets for conducting experiments. After the loading of the dataset the next step is to perform cleaning steps on the text data, to convert string into words.

### 5.2 PHASE 2: REFINING RAW DATA

In order to make the data meaningful, the data needs to undergo through tokenization, removal of stop-words and finally stemming is done in this phase. Figure 5.3 is representing String to word vector filter which is applied to datasets for converting the String to words for further classification. This filter is
Word parsing and tokenization including **Removal of stop words and Stemming** Stop words are the words that contain little data so should have been evacuated. A rundown of around 300-350 words are made as a content document. Further stemming is done using Snowball stemmer to decrease the inferred word to their beginning.

![Fig. 5. 3 String to word vector filter applied to dataset for converting the String to words to further classification.](image)

**5.3 PHASE3-4: HYBRID APPROACH (MODIFIED K-MEANS WITH IMPROVED BAGGING)**

For experiments, we have considered various algorithms for classification apart from K-means with SVM and Modified K-means with Improved Bagging (research approach). The algorithms implemented are Naïve Bayes, Decision Trees, KNN and SVM. The results from these algorithms are presented for Girl-Child dataset. For results on other datasets, we have attached the snapshots in ANNEXURE 2 at the end of the thesis.

This research work is validated using two approaches viz. K-cross validation and split method. To achieve the last research objective, comparisons are made between K-means with SVM and Modified K-means with Improved Bagging (research approach) based on parameters viz. Accuracy rate, False Positive rate, True Positive rate, Efficiency, Recall, F-measure, Clustering Time and Precision.
5.3.1.1 GIRL CHILD TWEETS DATASET: RESULTS USING K-CROSS VALIDATION MODEL

NAÏVE BAYES CLASSIFICATION

The figure 5.4 below gives view of the classification results of Naïve Bayes algorithm.

Fig. 5. 4 Results of Naïve Bayes Classification using cross validation model. (69.53 % accuracy).

The results show the accuracy of 69.53% i.e. 89 instances are correctly classified out of 128 instances. It is one of the famous machine learning algorithm working on the principle of Bayes theorem. Bayes theorem calculate the posterior probability. The kappa statistics for Naïve Bayes algorithm is 0.4776. Class details parameters are also shown like precision which is 0.715, recall 0.695, F Measure 0.689 TP Rate 0.695 and FP rate 0.215.

DECISION TREE CLASSIFICATION
The figure 5.5 below shows the classification results of decision tree algorithm. The results show the accuracy of 57.81% i.e. 74 instances are correctly classified out of 128 instances. Decision tree comprise the root hub, branches and leaf hubs.

![Fig. 5.5 Results of Decision Tree Classification using cross validation model. (57.81 % accuracy).](image)

In this the tree is made in a best down, recursive and separate and vanquish way. It works like an insatiable procedure. The inward hub characterizes the condition on the property, each branch characterizes the yield of the condition and each leaf hub characterizes the class name. The kappa statistics for decision tree algorithm is 0.251 which denotes the enhancement in the algorithm its value should lie between 0 and 1 closer to 1 means algorithms performs better. Class details parameters are also shown like precision which is 0.591, recall 0.571, F Measure 0.556, TP Rate 0.578 and FP rate 0.339.
KNN CLASSIFICATION

The results in figure 5.6 below show the accuracy of 65.625% i.e. 84 instances are correctly classified out of 128 instances. The k-Nearest Neighbor (kNN for short) is a non-parametric example based learning technique or lazy learning.

![KNN Classification Results](image)

Fig. 5.6 Results of K-Nearest Neighbor Classification using cross validation model. (65.625% accuracy).

The kappa statistics for KNN algorithm is 0.422 which denotes the enhancement in the algorithm its value should lie between 0 and 1, the more close to 1 means algorithms performs better. Class details parameters are also shown like precision which is 0.745, recall 0.656, F Measure 0.649, TP Rate 0.656 and FP rate 0.237.

SVM CLASSIFICATION

The figure 5.7 below gives a view of the classification results of SVM algorithm.
The results show the accuracy of 46.875% i.e. 60 instances are correctly classified out of 128 instances. SVM is a managed machine learning procedure that is utilized for both order and relapse. In this we plot each datum object as a peak in n-dimensional area where n, is the amount of features. At that point it performs grouping by searching the hyper - plane that split the two distinct classes.

**K-MEANS WITH SVM CLASSIFICATION ALGORITHM**

The figure 5.8 below gives a view of the classification results of K-means with SVM algorithm.

**Fig. 5. 7 Results of Support Vector Machine Classification using cross validation model. (6.875 % accuracy).**
The results show the accuracy of \textbf{92.9688} \% i.e. 119 instances are correctly classified out of 128 instances. In this, firstly the data is clustered using K-means algorithm then the classification is performed using SVM. That’s why there is drastic change in the accuracy in SVM and K-means with SVM. K-means is used for clustering the data based on the similarity between the instances using Euclidean distance. SVM is a managed machine learning procedure that is used for both order and relapse. In this we plot each datum thing as a point in n-dimensional space where n= number of highlights. At that point it performs grouping by finding the hyper - plane that separate the two distinct classes. Class details parameters are also shown like precision which is 0.864, recall 0.93, F Measure 0.896, TP rate is 0.93 and FP rate is 0.93.

\textbf{MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM}

The figure below shows the classification results of Modified K-means with Improved Bagging.
algorithm. The results show the accuracy of **95.3125%** i.e. 122 instances are correctly classified out of 128 instances. In this, K-means is modified and the clustered data is further classified using Improved Bagging which comprises of Bagging with Boosting. Class details parameters are also shown like precision which is 0.94, recall 0.95, F Measure 0.94, TP rate is 0.953 and FP rate is 0.62.

![Fig. 5. 9 Results of Modified K-means with Improved Bagging using cross validation model. (95.3125% accuracy).](image)

Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 1057 milliseconds and proposed takes 250 milliseconds time to cluster the data into 2 clusters.

**5.3.1.2 GIRL CHILD TWEETS DATASET USING %AGE SPLIT VALIDATION MODEL**

**NAÏVE BAYES CLASSIFICATION**
Fig. 5.10 Results of Naïve Bayes Classification using percentage split validation model. (66.66 % accuracy).

The figure above shows the classification results of Naïve Bayes algorithm. In this the data is divided into training and testing based percentage split. Training percentage taken is 70% and Testing percentage taken is 30 % i.e. 89 instances are in training data and 39 instances are testing data out of 128 tweets datasets. The results show the testing accuracy of 66.66 % i.e. 26 instances are correctly classified out of 39 instances.

DECISION TREE CLASSIFICATION

The figure 5.11 below shows the classification results of decision tree algorithm. The results show the accuracy of 53.85% i.e. 21 instances are correctly classified out of 39 testing instances. In this the data is divided into training and testing based percentage split. Training percentage taken is 70% and Testing percentage taken is 30 % i.e. 89 instances are in training data and 39 instances are testing data out of 128 tweets datasets.
Fig. 5.11 Results of Decision Tree Classification using percentage split validation model. (53.85 % accuracy).

KNN CLASSIFICATION

The figure 5.12 below shows the classification results of KNN algorithm. The results show the accuracy of 61.5385% i.e. 24 instances are correctly classified out of 39 testing instances. In this the data is divided into training and testing based percentage split. Training percentage taken is 70% and Testing percentage taken is 30 % i.e. 89 instances are in training data and 39 instances are testing data out of 128 tweets datasets.
Fig. 5. 12 Results of K-Nearest Neighbor Classification using percentage split validation model. (61.5385 % accuracy).

SVM CLASSIFICATION

Fig. 5. 13 Results of Support Vector Machine Classification using percentage split validation model. (51.2821 % accuracy).
The figure 5.13 demonstrates the classification results of SVM algorithm. In this the data is separated into training and testing based percentage split. Training percentage taken is 70% and Testing percentage taken is 30 % i.e. 89 instances are in training data and 39 instances are testing data out of 128 tweets datasets. The results show the accuracy of 51.2821% i.e. 20 instances are correctly classified out of 39 testing instances.

### K-MEANS WITH SVM CLASSIFICATION ALGORITHM

The figure 5.14 above shows the classification results of K-means with SVM algorithm. The results show the accuracy of **92.3077 %** i.e. 36 instances are correctly classified out of 39 testing instances. The precision for this algorithm is 0.852 for girl-child dataset using split percentage validation.

![Fig. 5. 14 Results of K-means with SVM using percentage split validation model. (92.3077 % accuracy).](image)

**MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM**

The figure 5.15 shows the classification results of Modified K-means with Improved Bagging algorithm. The results show the accuracy of **92.3077%** i.e. 36 instances are correctly classified out of 39 instances. In this, K-means is modified and the clustered data is further classified using Improved
Bagging which comprises of Bagging with Boosting.

Although the accuracy in this case is same for K-means with SVM and Modified K-means with improved Bagging i.e. 92.3077. But the other parameters like error rates are less in the proposed algorithm i.e. mean absolute error is 0.0501in proposed and 0.0513 in existing, Root mean squared error is 0.2025 in proposed and 0.2265 in existing. Same is for Relative absolute error 0.4129 and Root relative squared error 102.1961%. Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 1116 milliseconds and proposed takes 250 milliseconds time to cluster the data.

5.3.2 JNU KANHIYA TWEETS DATASET: RESULTS USING CROSS VALIDATION MODEL

We will evaluate the dataset with both the validation techniques.

5.3.2.1 K-MEANS WITH SVM CLASSIFICATION ALGORITHM

The figure 5.16 shows the classification results of K-means with SVM algorithm. The results show the
accuracy of **51.96%** i.e. 53 instances are correctly classified out of 102 instances. Class details parameters are also shown like precision which is 0.27, recall 0.52, F Measure 0.355.

**Fig. 5.** Results of K-means with SVM using cross validation model. (51.96 % accuracy).

**MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM**
The figure 5.17 shows the classification results of K-means with SVM algorithm. The results show the accuracy of 93.13% i.e. 95 instances are correctly classified out of 102 instances. Class details parameters are also shown like precision which is 0.936, recall 0.931, F Measure 0.925.

Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 1121 milliseconds and proposed takes 218 milliseconds time to cluster the data into 2 clusters.

5.3.3.2 JNU KANHIYA TWEETS DATASET: RESULTS USING %AGE SPLIT VALIDATION

K-MEANS WITH SVM CLASSIFICATION ALGORITHM

The figure 5.18 demonstrates the classification results of K-means with SVM algorithm. In this the data is separated into training and testing based percentage split. Training percentage taken is 70% and Testing percentage taken is 30 % i.e. 71 instances are in training data and 31 instances are testing data out of 102 tweets datasets. The results show the accuracy of 74.19 % i.e. 23 instances are correctly classified out of 31 testing instances.
MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM

The figure 5.19 above shows the classification results of Modified K-means with Improved Bagging algorithm. The results show the accuracy of 87.096% i.e. 27 instances are correctly classified out of 31 instances. Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 1068 milliseconds and proposed takes 237 milliseconds time to cluster the data.
5.3.3 PRICE RISE TWEETS DATASET: RESULTS USING CROSS VALIDATION MODEL

We will evaluate the dataset with both the validation techniques

5.3.3.1 K-MEANS WITH SVM CLASSIFICATION ALGORITHM

The figure 5.20 shows the classification results of K-means with SVM algorithm. The results show the accuracy of 87.95 % i.e. 241 instances are correctly classified out of 274 instances. Class details parameters are also shown like precision which is 77.4%, recall 88.0 %, F Measure 82.3%.
MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM

The figure 5.21 shows the classification results of Modified K-means with Improved Bagging algorithm. The results show the accuracy of 87.096% i.e. 27 instances are correctly classified out of 31 instances. The precision for this algorithm is 0.982.

The results demonstrate that the proposed algorithm performs better in all the parameters than the existing algorithm. Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 1959 milliseconds and proposed takes 237 milliseconds time to cluster the data.
Fig. 5. 21 Results of Modified K-means with Improved Bagging using cross validation model. (98.17% accuracy).

5.3.3.2 PRICE RISE TWEETS DATASET: RESULTS USING P%AGE SPLIT VALIDATION

K-MEANS WITH SVM CLASSIFICATION ALGORITHM

The figure 5.22 demonstrate the classification results of K-means with SVM algorithm In this the data is separated into training and testing based percentage split. Training percentage taken is 70% and Testing percentage taken is 30 % i.e. 191 instances are in training data and 83 instances are testing data out of 274 tweets datasets. The results show the accuracy of 84.337 % i.e. 70 instances are correctly classified out of 83 testing instances.
MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM

The figure 5.23 shows the classification results of Modified K-means with Improved Bagging algorithm. The results show the accuracy of 97.59% i.e. 81 instances are correctly classified out of 83 testing instances. In this, K-means is modified and the clustered data is further classified using Improved Bagging which comprises of Bagging with Boosting. The results demonstrate that the proposed algorithm performs better in all the parameters than the existing algorithm. Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 1959 milliseconds and proposed takes 1187 milliseconds time to cluster the data.
5.3.4 INTOLERANCE TWEETS DATASET: RESULTS USING CROSS VALIDATION MODEL

We will evaluate the dataset with both the validation techniques.

5.3.4.1 K-MEANS WITH SVM CLASSIFICATION ALGORITHM

The figure 5.24 shows the classification results of K-means with SVM algorithm. The results show the accuracy of **98.07%** i.e. 357 instances are correctly classified out of 365 instances.

Class details parameters are also shown like precision which is 96.2%, recall 98.1% and F Measure is measured as 97.1%.
MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM

The figure 5.25 shows the classification results of K-means with SVM algorithm. The results show the accuracy of **99.72%** i.e. 363 instances are correctly classified out of 365 instances. Class details parameters are also shown like precision which is 99.7%, recall 99.7%, F Measure 99.7%. Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 1703 milliseconds and proposed takes 1126 milliseconds time to cluster the data into 3 clusters.
5.3.4.2 INTOLERANCE TWEETS DATASET: RESULTS USING %AGE SPLIT VALIDATION

K-MEANS WITH SVM CLASSIFICATION ALGORITHM

The figure 5.26 shows the classification results of K-means with SVM algorithm. In this the data is divided into training and testing based percentage split. Training percentage taken is 70% and Testing percentage taken is 30 % i.e. 254 instances are in training data and 110 instances are testing data out of 364 tweets datasets.

The results show the accuracy of 97.27 % as 107 instances are correctly classified out of 110 testing instances.
MODIFIED K-MEANS WITH IMPROVED BAGGING ALGORITHM

The figure 5.27 shows the classification results of Modified K-means with Improved Bagging algorithm. The results show the accuracy of 98.18% i.e. 108 instances are correctly classified out of 110 testing instances.

In this, K-means is modified and the clustered data is further classified using Improved Bagging which comprises of Bagging with Boosting.
The results show that the proposed algorithm performs better in all the parameters than the existing algorithm. Also, the clustering time of the proposed clustering algorithm is less as compared to the existing K-means algorithm. K-means takes 2665 milliseconds and proposed takes 1633 milliseconds time to cluster the data.

5.4 COMPARITIVE RESULTS

True positives (TP): These are cases in which we predicted Positive (tweet as Positive), and tweets were positive actually.

True negatives (TN): We predicted Negative, and tweets were Negative.

False positives (FP): We predicted Positive, but tweets were actually Negative.

False negatives (FN): We predicted Negative, but tweets were actually Positive.
ACCURACY

Accuracy is the percentage of correctly identifying the classes on the datasets. It can be calculated as the number of correctly classified tweets to the total number of tweets. It should be highest for the best technique.

\[
\text{Accuracy} = \frac{TP}{TP+FP+TN+FN}
\]  

(17)

The table below (Table 5.1) and Graph 5.1 represents the classification accuracy of the proposed technique with the existing techniques on various social issues tweets dataset.

Table 5.1 Accuracy of the proposed technique with the existing techniques on various dataset.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>69.53</td>
<td>71.5686</td>
<td>72.9927</td>
<td>57.1429</td>
</tr>
<tr>
<td>DT</td>
<td>57.8125</td>
<td>54.902</td>
<td>79.562</td>
<td>54.3956</td>
</tr>
<tr>
<td>KNN</td>
<td>65.625</td>
<td>59.8</td>
<td>75.1825</td>
<td>57.967</td>
</tr>
<tr>
<td>SVM</td>
<td>46.875</td>
<td>48.0392</td>
<td>70.073</td>
<td>45.8791</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>51.96</td>
<td>51.9608</td>
<td>87.9562</td>
<td>98.0769</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>93.1373</td>
<td>93.1373</td>
<td>98.1752</td>
<td>99.7253</td>
</tr>
</tbody>
</table>

Graph 5.1 Accuracy of the proposed technique with the existing techniques based on various dataset.
**PRECISION**

Precision measures the exactness of a classifier. A high precision indicates that false positives are less and a poor precision indicates that false positives are more. This is commonly inconsistent with recall; a simple approach to enhance precision is to reduce recall.

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (18)
\]

Table 5.2 is representing classification Precision of the proposed technique with the existing techniques on various social issues tweets dataset.

**Table 5.2 Precision of the proposed technique with the existing techniques on various dataset.**

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.715</td>
<td>0.732</td>
<td>0.714</td>
<td>0.566</td>
</tr>
<tr>
<td>DT</td>
<td>0.591</td>
<td>0.556</td>
<td>0.809</td>
<td>0.537</td>
</tr>
<tr>
<td>KNN</td>
<td>0.745</td>
<td>0.613</td>
<td>0.767</td>
<td>0.667</td>
</tr>
<tr>
<td>SVM</td>
<td>0.22</td>
<td>0.231</td>
<td>0.491</td>
<td>0.21</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>0.27</td>
<td>0.27</td>
<td>0.774</td>
<td>0.962</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>0.936</td>
<td>0.936</td>
<td>0.982</td>
<td>0.997</td>
</tr>
</tbody>
</table>

**Graph 5.2 Precision of the proposed technique with the existing techniques based on various dataset.**
Graph 5.2 is showing the graphical representation of classification Precision of the proposed technique with the existing techniques on various social issues tweets dataset.

**RECALL**

Recall measures the completeness, or sensitivity, of a classifier. Recall goes other way than precision; more recall implies few false negatives, whereas low recall implies high false negatives. Enhancing recall can usually reduce.

\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}
\]

Recall = True Positive / (True Positive + False Positive) \hspace{1cm} (19)

Table 5.3 and Graph 5.3 are representing the classification Recall of the proposed technique with the existing techniques on various social issues tweets dataset.

**Table 5.3 Recall of the proposed technique with the existing techniques on various dataset.**

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.695</td>
<td>0.716</td>
<td>0.73</td>
<td>0.571</td>
</tr>
<tr>
<td>DT</td>
<td>0.578</td>
<td>0.549</td>
<td>0.796</td>
<td>0.544</td>
</tr>
<tr>
<td>KNN</td>
<td>0.656</td>
<td>0.598</td>
<td>0.752</td>
<td>0.58</td>
</tr>
<tr>
<td>SVM</td>
<td>0.469</td>
<td>0.48</td>
<td>0.701</td>
<td>0.459</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>0.52</td>
<td>0.52</td>
<td>0.88</td>
<td>0.981</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>0.931</td>
<td>0.931</td>
<td>0.982</td>
<td>0.997</td>
</tr>
</tbody>
</table>

**Graph 5.3 Recall of the proposed technique with the existing techniques based on various social issues tweets dataset.**
**F-MEASURE**

F-measure is produced by combining precision and recall. “It is weighted harmonic mean of precision and recall”.

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} \times \text{Recall})} \tag{20}
\]

Table 5.4 is representing the classification F-measure of the proposed technique with the existing techniques on various social issues tweets dataset.

**Table 5.4 F-measure of the proposed technique with existing techniques on social issues tweets dataset**

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.689</td>
<td>0.714</td>
<td>0.717</td>
<td>0.568</td>
</tr>
<tr>
<td>DT</td>
<td>0.556</td>
<td>0.546</td>
<td>0.767</td>
<td>0.52</td>
</tr>
<tr>
<td>KNN</td>
<td>0.649</td>
<td>0.559</td>
<td>0.751</td>
<td>0.526</td>
</tr>
<tr>
<td>SVM</td>
<td>0.299</td>
<td>0.312</td>
<td>0.577</td>
<td>0.289</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>0.355</td>
<td>0.355</td>
<td>0.823</td>
<td>0.971</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>0.925</td>
<td>0.925</td>
<td>0.98</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Graph 5.4 is showing the graphical representing the classification F-measure of the proposed technique with the existing techniques on various social issues tweets dataset.
SENSITIVITY OR TP RATE

It checks how many tweets are correctly classified a sentiment (Positive, Negative and Neutral). It can be calculated by number of tweets that are correctly classified as sentiment to the total number of false tweets.

\[
TP\ Rate = \frac{TP}{TP + FN}
\]  

(21)

Table 5.5 is representing the classification TP rate of the proposed technique with the existing techniques on various social issues tweets dataset.

Table 5.5 TP rate of the proposed technique with the existing techniques on various social issues tweets dataset

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.695</td>
<td>0.716</td>
<td>0.73</td>
<td>0.571</td>
</tr>
<tr>
<td>DT</td>
<td>0.578</td>
<td>0.549</td>
<td>0.796</td>
<td>0.544</td>
</tr>
<tr>
<td>KNN</td>
<td>0.656</td>
<td>0.598</td>
<td>0.752</td>
<td>0.58</td>
</tr>
<tr>
<td>SVM</td>
<td>0.469</td>
<td>0.48</td>
<td>0.701</td>
<td>0.459</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>0.52</td>
<td>0.52</td>
<td>0.88</td>
<td>0.981</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>0.931</td>
<td>0.931</td>
<td>0.982</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Graph 5.5 is showing the graphical representing the classification TP rate of the proposed technique with the existing techniques on various social issues tweets dataset.

Graph 5.5 TP Rate of the proposed technique with the existing techniques based on various social issues tweets dataset.
FP RATE (FALL-OUT)

It can be calculated as number of tweets that are incorrectly classified to the total number of positive instances. It should be low.

\[
FP \text{ Rate} = \frac{FP}{FP+TN} \tag{22}
\]

Table 5.6 is showing the classification FP rate of the proposed technique with the existing techniques on various social issues tweets dataset. Graph 5.6 next to it is presenting the graphical representation of FP Rate of the proposed technique with the existing techniques based on various social issues tweets dataset.

*Table 5.6 FP rate of the proposed technique with the existing techniques on various social issues tweets dataset.*

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.215</td>
<td>0.211</td>
<td>0.425</td>
<td>0.295</td>
</tr>
<tr>
<td>DT</td>
<td>0.339</td>
<td>0.352</td>
<td>0.44</td>
<td>0.36</td>
</tr>
<tr>
<td>KNN</td>
<td>0.237</td>
<td>0.347</td>
<td>0.324</td>
<td>0.346</td>
</tr>
<tr>
<td>SVM</td>
<td>0.469</td>
<td>0.48</td>
<td>0.701</td>
<td>0.459</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>0.52</td>
<td>0.52</td>
<td>0.88</td>
<td>0.981</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>0.065</td>
<td>0.065</td>
<td>0.133</td>
<td>0.14</td>
</tr>
</tbody>
</table>

*Graph 5.6 FP Rate of the proposed technique with the existing techniques based on various dataset.*
CLASSIFIED INSTANCES

Table 5.7 below is representing the summary of the classification results based on the correctly classified and incorrectly classified instances of the proposed method with the current techniques on various social issues tweets dataset.

In Naïve Bayes, 89 tweets are correctly classified and 39 are incorrectly classified. Similarly, for KNN 84 tweets are correctly classified and 44 are incorrectly classified tweets, for K-means with SVM 53 tweets are correctly classified and 49 are incorrectly classified tweets, for proposed approach Modified K-means with Improved Bagging 95 tweets are correctly classified and 7 are incorrectly classified tweets for girl child tweets dataset. Similarly, for other social issues tweets dataset; the results of the proposed approach are better than the existing.

Table 5.7 Showing the classification results with Correctly classified and Incorrectly Classified Instances of the proposed method with the current techniques on various dataset.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correctly</td>
<td>Incorrectly</td>
<td>Correctly</td>
<td>Incorrectly</td>
</tr>
<tr>
<td></td>
<td>Classified</td>
<td>Classified</td>
<td>Classified</td>
<td>Classified</td>
</tr>
<tr>
<td>NB</td>
<td>89</td>
<td>39</td>
<td>73</td>
<td>49</td>
</tr>
<tr>
<td>DT</td>
<td>74</td>
<td>57</td>
<td>56</td>
<td>46</td>
</tr>
<tr>
<td>KNN</td>
<td>84</td>
<td>44</td>
<td>61</td>
<td>41</td>
</tr>
<tr>
<td>SVM</td>
<td>60</td>
<td>68</td>
<td>49</td>
<td>53</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>53</td>
<td>49</td>
<td>53</td>
<td>49</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>95</td>
<td>7</td>
<td>95</td>
<td>7</td>
</tr>
</tbody>
</table>

Graph 5.7 below is representing the summary of the classification results based on the correctly classified and incorrectly classified instances of the proposed method with the current techniques on various social issues tweets dataset in the graphical form.
**Graph 5.7 Showing the with Correctly classified and Incorrectly Classified Instances of the proposed technique with the existing techniques based on various dataset.**

**MEAN ABSOLUTE ERROR (MAE)**

The most frequently used metrics to measure accuracy for continuous variables are mean absolute error and root mean square error. Mean Absolute Error measures the average magnitude of the errors in a set of prediction, without considering their polarity.

**ROOT MEAN SQUARED ERROR (RMSE)**

The RMSE is a quadratic scoring principle which measures the normal extent of the blunder. The condition for the RMSE is given in both of the references. Communicating the equation in words, the distinction amongst gauge and relating watched values are each squared and after that found the middle value of over the example. At long last, the square foundation of the normal is taken. Since the mistakes are squared before they are found the middle value of, the RMSE gives a generally high weight to extensive blunders. This implies the RMSE is most valuable when substantial mistakes are especially unfortunate.

The MAE and the RMSE can be utilized together to analyze the variety in the mistakes in an arrangement of gauges. The RMSE will dependably be bigger or equivalent to the MAE; the more prominent contrast between them, the more prominent the change in the individual mistakes in the example. On the off chance that the RMSE=MAE, at that point every one of the mistakes are of a similar extent.
Table 5.8 Mean Absolute Error and Root Mean Squared Error of the proposed technique with the existing techniques on various dataset.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>NB</td>
<td>0.2245</td>
<td>0.4259</td>
<td>0.2245</td>
<td>0.401</td>
</tr>
<tr>
<td>DT</td>
<td>0.3107</td>
<td>0.4519</td>
<td>0.3123</td>
<td>0.4634</td>
</tr>
<tr>
<td>KNN</td>
<td>0.2403</td>
<td>0.4751</td>
<td>0.2761</td>
<td>0.4934</td>
</tr>
<tr>
<td>SVM</td>
<td>0.3542</td>
<td>0.5951</td>
<td>0.3464</td>
<td>0.5886</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>0.3203</td>
<td>0.5659</td>
<td>0.3203</td>
<td>0.5659</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>0.1937</td>
<td>0.2596</td>
<td>0.1937</td>
<td>0.2596</td>
</tr>
</tbody>
</table>

Graph 5.8 Mean Absolute Error and Root Mean Squared Error of the proposed technique with the existing techniques based on various dataset.

Table 5.9 below is representing the classification results with Relative Absolute Error and Root Relative Squared Error of the proposed technique with the existing techniques on various social issues tweets dataset.
Table 5. 9 Relative Absolute Error and Root Relative Squared Error of the proposed technique with the existing techniques on various dataset.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Girl Child Tweets</th>
<th>JNU Kanhiya Tweets</th>
<th>Price rise tweets</th>
<th>Intolerance Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAE</td>
<td>RRSE</td>
<td>RAE</td>
<td>RRSE</td>
</tr>
<tr>
<td>NB</td>
<td>55.97</td>
<td>95.21</td>
<td>56.4526</td>
<td>90.0747</td>
</tr>
<tr>
<td>DT</td>
<td>77.48</td>
<td>101.02</td>
<td>78.538</td>
<td>104.0822</td>
</tr>
<tr>
<td>KNN</td>
<td>59.9123</td>
<td>106.22</td>
<td>69.4327</td>
<td>110.8175</td>
</tr>
<tr>
<td>SVM</td>
<td>88.31</td>
<td>133.049</td>
<td>87.1181</td>
<td>132.194</td>
</tr>
<tr>
<td>K-means with SVM</td>
<td>82.43</td>
<td>128.64</td>
<td>82.4345</td>
<td>128.6432</td>
</tr>
<tr>
<td>Proposed Hybrid approach</td>
<td>49.86</td>
<td>59.02</td>
<td>49.8653</td>
<td>59.0203</td>
</tr>
</tbody>
</table>

Following is a graph presenting the plots for showing the Relative Absolute Error and Root Relative Squared Error of the proposed technique with the existing techniques based on various social issues tweets dataset.

Graph 5. 9 Relative Absolute Error and Root Relative Squared Error of the proposed technique with the existing techniques based on various dataset.

It is evident from the above table and graph that the proposed Modified K-means with Improved Bagging approach has very less error rate as compare to other classification algorithm.
The contributions provided in this thesis are focused on the development of hybrid method based on Modified k-means with improved bagging to address the problem of Tweet analysis on social causes in India. The results turned out to be satisfactory and confirmed that the research work reported in this thesis has made interesting contributions.