Natural Language Processing and Machine Learning Based Prediction for Traffic Accident

Noor-Ul-Saba

Department of computer science

University of Sialkot, Pakistan

noorrajpoot149@gmail.com ID

Abstract

Traffic management can be greatly helped by predicting the length of traffic incidents. In this study, we analyze this prediction task as a classification problem on order to generate a more precise real-time prediction of traffic accident duration and utilize the increasing amount of traffic texts in social networks. [2] Traffic accidents cannot be prevented, even with all these resources in the design and construction of automotive safety measures. Both urban and rural regions see a high rate of accidents. [3] By creating precise prediction models that can automatically separate distinct unintentional incidents, patterns related with different situations can be detected. These classifiers will help create safety precautions and Prevent incidents. [3] In this paper use some Machine learning models to analyses the results as much as possible while using limited resources.

Keywords: Natural language processing; urban traffic management; Traffic accident prediction; Machine Learning; NLP; ML

[4]

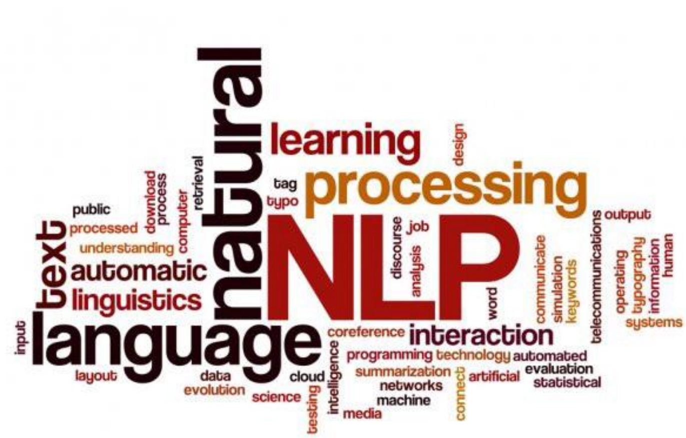

Figure 1

Introduction:

Traffic accidents routinely draw public attention as a result of the development of the road system and the increase in automobile ownership. The vehicles on the same road are often highly affected by the sudden occurrence of traffic accidents [2].

There are different efficient models and algorithms available today for predicting the duration of traffic accidents. Traditional ones, like those based on logistic regression and regression methods. After the maturity of each machine learning algorithm, models based on intelligent algorithms such as Naive Bayes model and simple Logistic model achieved a higher degree of accuracy for that task. At the same time, different ML algorithms performed this task even better.

Social network data may be easily browsed in real-time using technologies like crawlers. There is a probability that other road users who are on the scene at the first instant of a traffic accident would publish the news of the traffic accident on the scene [3].

The majority of social network data is text, and text analysis tools have shown great accuracy as natural language processing (NLP) has advanced in recent years. Furthermore, the end result of large-scale data training, such as Naive Bayes, Random forest and many algorithms is efficient in solving even the most challenging NLP tasks, including as classification and prediction.

Natural Language Processing

The word "natural language processing" (NLP) relates to how computers and human language interact. Even though it has been there for a while and is something that many people utilize on a daily basis, it is usually taken for granted. Similar to how would human can determine the proper word, phrase, or reaction by searching at context clues. [4] Basically it is a simple technique. The ability of a computer program to understand spoken and written is considered as NLP. [5]

Figure 2

Machine Learning

With the use of machine learning (ML), which is a type of artificial intelligence (AI), software tools can predict the outcome more accurately without having to be specifically instructed to do it at all. [6] In predicting new output values, machine learning algorithms use past data as input. The goal of machine learning is to create software programs that can access data and use it to acquire knowledge on their own. [7]

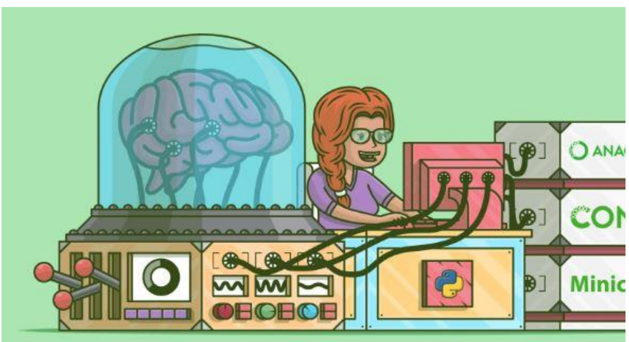

Figure 3

Method

The strategy that was specifically used to deal with the issue in this paper is described in this section. The classification problem was considered as the prediction problem. The text data is collected, and then it is categorized by accident duration. Firstly, the Bag-of-words model is used to extract text features, and that dataset is given to the WEKA software, then the labels of the data are converted into three categories, Fatal, serious non-fatal and serious. The dataset is applied on these five models [3].

Naïve Bayes
J48
Bayes Nets
Simple Logistic
Decision Stump

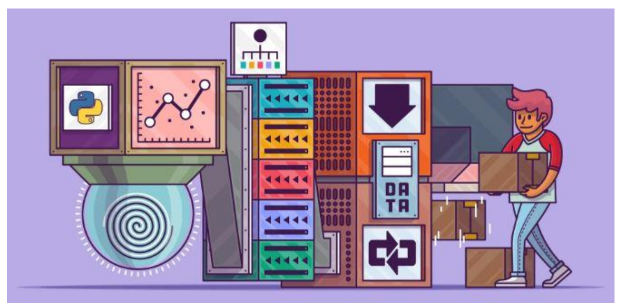

Figure 4

Gathered Dataset

In this paper, an already prepared datasets has been used from Stats NZ. The dataset is about serious injury outcome. Dataset contain 2748 instances. Dataset contain three features Label, Count, and weight.

Name: Series_reference　Type: Nominal
Missing: 0 (0%)　Distinct: 144　Unique: 0 (0%)

No.	Label	Count	Weight
1	W_A11	17	17.615
2	W_A12	17	17.615
3	W_F11B	17	17.615
4	W_F12B	17	17.615
5	W_W12	17	17.615
6	W_W14	17	17.615
7	M_S11	17	17.615
8	M_S12	17	17.615
9	M_F01C	19	20.284

...

No.	Label	Count	Weight
136	f21c	19	17.3
137	a22	19	17.3
138	a21	19	17.3
139	c22	19	17.3
140	c21	19	17.3
141	in22	19	17.3
142	in21	19	17.3
143	p22	19	17.3
144	p21	19	17.3

Figure 5

Confusion Matrix

Confusion matrix is widely used measurement when trying to solve classification issues. Both binary classification and multiclass classification issues can be solved with it.

```
=== Confusion Matrix ===

     a      b      c    <-- classified as
 828.96  56.99  30.05 |      a = Fatal
  23.49 874.36  18.15 |      b = Serious non-fatal
  20.03   3.64 892.33 |      c = Serious
```

Figure 6 Bayes Nets Confusion Matrix

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 916   0   0 |   a = Fatal
   0 916   0 |   b = Serious non-fatal
   0   0 916 |   c = Serious
```

Figure 7 J48 confusion Matrix

```
=== Confusion Matrix ===

     a      b      c    <-- classified as
 190.66 214.49 510.85 |      a = Fatal
   3.2  271.17 641.63 |      b = Serious non-fatal
   0.91 274.98 640.11 |      c = Serious
```

Figure 8 Decision Stump Confusion Matrix

```
=== Confusion Matrix ===

     a      b      c     <-- classified as
 896.31  19.69   0     |      a = Fatal
 519.92 396.08   0     |      b = Serious non-f
 401.55  12.75 501.71  |      c = Serious
```

Figure 9 Naive Bayes Confusion Matrix

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 916   0   0 |   a = Fatal
   0 916   0 |   b = Serious non-fatal
   0   0 916 |   c = Serious
```

Figure 10 Simple Logistic confusion matrix

Results and discussion

In this paper, the number of classifications is set into different categories by combining the data content and life reality. Multiple sequential classifications are obtained by the above method, and the duration range of each classification [3]

Dataset is applied on Weka 3.8.4 version.

With Text Preprocessing

The dataset csv file is loaded in weka. The following figures show the results with text preprocessing.

Total instance are 2748.

Algorithms	Correctly Classifier Prediction
Bayes nets	94%
Naïve Bayes	65%
J48	100%
Decision stump	40%
Simple logistic	100%

1. Bayes Nets

Using a variety of search techniques and quality indicators, Bayes Network learns. A Bayes Network classifier base class. Provides facilities and data structures that are common to Bayes Network learning algorithms like K2 and B, including network structure and conditional probability distributions, among others. [9]

It takes 0.21 seconds to build the model

```
Time taken to build model: 0.21 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       2595.649          94.4559 %
Incorrectly Classified Instances      152.351           5.5441 %
Kappa statistic                         0.9168
Mean absolute error                     0.064
Root mean squared error                 0.1678
Relative absolute error                14.399  %
Root relative squared error            35.5911 %
Total Number of Instances            2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.905    0.024    0.950      0.905    0.927      0.892    0.988     0.979     Fatal
                 0.955    0.033    0.935      0.955    0.945      0.917    0.992     0.990     Serious non-fatal
                 0.974    0.026    0.949      0.974    0.961      0.942    0.996     0.991     Serious
Weighted Avg.    0.945    0.028    0.945      0.945    0.944      0.917    0.992     0.987
```

Figure 11 BayesNet Results

2. Naïve Bayes

The Naive Bayes is the basis of the statistical machine learning method known as Binary Classification, which is utilized for a variety of classification problems. It has been used successfully for many things, but it best at solving natural language processing (NLP) issues. [10]

It takes 0.02 seconds to build model

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       1794.0972         65.2874 %
Incorrectly Classified Instances      953.9028         34.7126 %
Kappa statistic                         0.4793
Mean absolute error                     0.2229
Root mean squared error                 0.408
Relative absolute error                50.1518 %
Root relative squared error            86.5573 %
Total Number of Instances            2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979    0.503    0.493      0.979    0.656      0.474    0.980     0.985     Fatal
                 0.432    0.018    0.924      0.432    0.589      0.539    0.963     0.903     Serious non-fatal
                 0.548    0.000    1.000      0.548    0.708      0.668    0.964     0.939     Serious
Weighted Avg.    0.653    0.174    0.806      0.653    0.651      0.560    0.969     0.942
```

Figure 12 Naive Bayes Results

3. J48 Classifier

One of the greatest machine learning algorithms for categorical and continuous data analysis is the J48 algorithm. When used for such purpose, however, it takes up more memory and reduces classification performance and accuracy for medical data. [9]

It takes 0.06 seconds to build the model

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2748              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious non-fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000
```

Figure 13 J48 Classifier Results

4. Decision Stump

To create a decision tree with just one split, use the Decision Stump method. Unseen samples can be categorized using the given tree. When this operator is improved by other operators like the AdaBoost operator, it can be quite effective. Each example in the provided Example Set has a number of attributes and is a class (like yes or no). A decision node is any node other than the leaf nodes of a decision tree, which include the class name. Each branch (to another decision tree) is a potential value for the attribute tested at the decision node. [9]

It takes 0.04 seconds to build the model

```
Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1101.9382          40.0996 %
Incorrectly Classified Instances      1646.0618          59.9004 %
Kappa statistic                          0.1015
Mean absolute error                      0.4119
Root mean squared error                  0.4543
Relative absolute error                 92.6872 %
Root relative squared error             96.3634 %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.208    0.002    0.979      0.208    0.343      0.378    0.596     0.514     Fatal
                 0.296    0.267    0.357      0.296    0.323      0.030    0.545     0.354     Serious non-fatal
                 0.699    0.629    0.357      0.699    0.473      0.069    0.551     0.357     Serious
Weighted Avg.    0.401    0.300    0.564      0.401    0.380      0.159    0.564     0.408
```

Figure 14 Decision Stump Results

5. Simple Logistic

When a nominal variable and a measurement variable are present, we use simple logistic regression to determine whether variation in the measurement variable affects the nominal variable.

It takes 6.47 seconds to build the model

```
Time taken to build model: 6.47 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2748              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.1677
Root mean squared error                  0.2002
Relative absolute error                 37.7429 %
Root relative squared error             42.4675 %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious non-fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000
```

Figure 15 Simple logistic Results

Without Preprocessing Text

The dataset is loaded in weka then multiple models are applied on it without preprocessing text. The results are given in follows figures and table.

Algorithms	Correctly Classifier Prediction over 2748 instances
Bayes nets	94.5%
Naïve Bayes	65%
J48	100%
Decision stump	43%
Simple logistic	100%

1. Bayes Net

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2598               94.5415 %
Incorrectly Classified Instances       150                5.4585 %
Kappa statistic                          0.9179
Mean absolute error                      0.0633
Root mean squared error                  0.1658
Relative absolute error                 14.2873 %
Root relative squared error             35.2148 %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.903    0.023    0.950      0.903    0.926      0.892    0.989     0.978     Fatal
                 0.953    0.032    0.931      0.953    0.942      0.915    0.992     0.989     Serious non-fatal
                 0.976    0.027    0.954      0.976    0.965      0.945    0.995     0.992     Serious
Weighted Avg.    0.945    0.027    0.946      0.945    0.945      0.919    0.992     0.987

=== Confusion Matrix ===

   a   b   c   <-- classified as
 798  57  29 |   a = Fatal
  22 818  18 |   b = Serious non-fatal
  20   4 982 |   c = Serious
```

Figure 16

2. Naïve Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1789               65.1019 %
Incorrectly Classified Instances       959               34.8981 %
Kappa statistic                          0.4796
Mean absolute error                      0.2245
Root mean squared error                  0.4095
Relative absolute error                 50.6369 %
Root relative squared error             86.9655 %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979    0.497    0.483      0.979    0.647      0.472    0.980     0.985     Fatal
                 0.424    0.017    0.917      0.424    0.580      0.536    0.962     0.893     Serious non-fatal
                 0.557    0.000    1.000      0.557    0.715      0.666    0.965     0.948     Serious
Weighted Avg.    0.651    0.165    0.808      0.651    0.651      0.563    0.969     0.943

=== Confusion Matrix ===

   a   b   c   <-- classified as
 865  19   0 |   a = Fatal
 494 364   0 |   b = Serious non-fatal
 432  14 560 |   c = Serious
```

Figure 17

3. J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2748              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious non-fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b    c   <-- classified as
  884    0    0 |    a = Fatal
    0  858    0 |    b = Serious non-fatal
    0    0 1006 |    c = Serious
```

Figure 18

4. Decision Stump

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1191               43.3406 %
Incorrectly Classified Instances      1557               56.6594 %
Kappa statistic                          0.1105
Mean absolute error                      0.4105
Root mean squared error                  0.4534
Relative absolute error                 92.5903 %
Root relative squared error             96.3075 %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.210    0.002    0.979      0.210    0.346      0.383    0.596     0.504     Fatal
                 0.000    0.000    ?          0.000    ?          ?        0.542     0.330     Serious non-fatal
                 0.999    0.892    0.393      0.999    0.564      0.204    0.551     0.391     Serious
Weighted Avg.    0.433    0.327    ?          0.433    ?          ?        0.563     0.408

=== Confusion Matrix ===

    a    b    c   <-- classified as
  186    0  698 |    a = Fatal
    3    0  855 |    b = Serious non-fatal
    1    0 1005 |    c = Serious
```

Figure 19

5. Simple Logistic

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2748              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.1695
Root mean squared error                  0.2019
Relative absolute error                 38.2248 %
Root relative squared error             42.8809 %
Total Number of Instances             2748

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious non-fatal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Serious
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b    c   <-- classified as
  884    0    0 |    a = Fatal
    0  858    0 |    b = Serious non-fatal
    0    0 1006 |    c = Serious
```

Figure 20

Visualization Class Balance Results

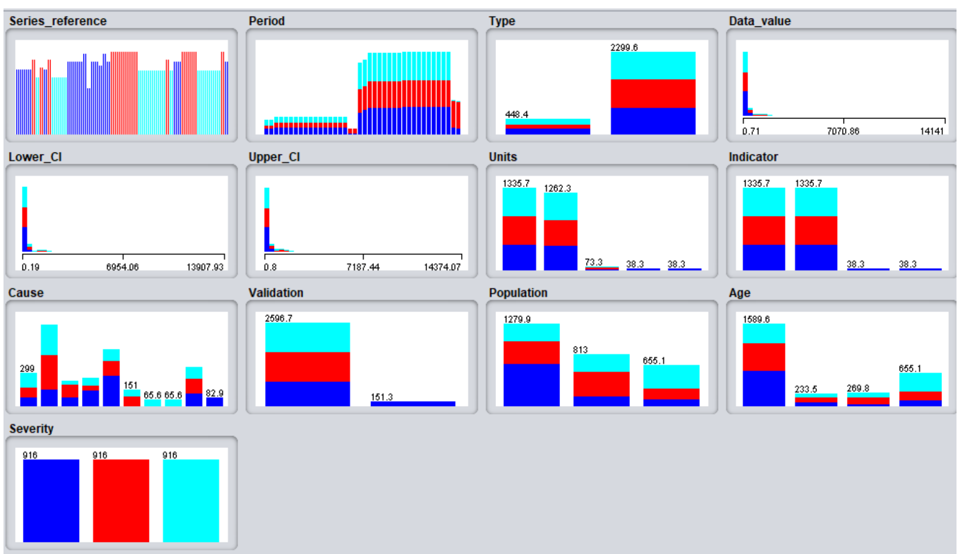

Figure 21 Visualization Class Balance Results

Conclusion

This paper has presented the analysis of traffic accident. Dataset is applied on five classifies: Bayes nets, Naïve Bayes, J48, Decision Stump, Simple Logistic. Applying dataset with preprocessing and without preprocessing and classification. Furthermore, execution time is also calculated with five classifier models.

REFERENCES

https://www.pantechsolutions.net/machine-learning-projects/road-accident-analysis-using-machine-learning

https://www.wonderflow.ai/blog/natural-language-processing-examples

https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

https://www.techtarget.com/searchenterpriseai/definition/machine-learning-

ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values.

https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11933/119330W/Traffic-accident-duration-prediction-based-on-natural-language-processing-and/10.1117/12.2614987.full?SSO=1

https://www.expert.ai/blog/machine-learning-definition/

https://www.google.com/search?q=natural%20language%20processing&tbm=isch&hl=en-GB&tbs=rimg:CY1fpuEIqV5_1YQGBbidmm2ES8AEAsgIMCgIIABAAOgQIABAA&sa=X&ved=0CBsQuIIBahcKEwi48cfspNX4AhUAAAAAHQAAAAAQIA&biw=1349&bih=610

https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/BayesNet.html

https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

https://www.geeksforgeeks.org/building-a-machine-learning-model-using-j48-classifier/

https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/decision_stump.html

https://www.statstest.com/simple-logistic-regression/

https://realpython.com/logistic-regression-python/

https://realpython.com/sentiment-analysis-python/

https://blog.kata.ai/basics-of-nlp

https://www.stats.govt.nz/information-releases/serious-injury-outcome-indicators-2000-2020