

Knowledge based Techniques for Pragmatic Feature Engineering and Opinion Mining on Divergent Data sets

Annie Syrien 1^a, M. Hanumanthappa 2^{a,*} and Ravi Kumar 3^b

^a Department of Computer Science and Applications, Bangalore University, Bengaluru 560056, India

^b Kalinga Institute of Industrial Technology(KIIT), Bhubaneswar 751024, India

ABSTRACT

Sentiment Analysis and Opinion Mining is a computational text processing technique used for extraction of emotion from the given text, for e.g. joy, sad, disgust, fear, anger which can be polarized as positive, negative and neutral. The advancement in technology and data augmentation has been an easy access for data collection for the research work. In this research paper, different machine learning algorithms such as Extreme Gradient Boost classifier, Extra tree classifier, K-Nearest Neighbour classifier and Long Short Term Memory deep learning model was computed to classify the divergent datasets such as Twitter Bengaluru traffic data, movie review, fake news classification and financial sentiments. Performance of these classifier was tabulated with respect to independent datasets in terms of accuracy, f1-score, precision, sensitivity, mean squared error and, log loss and graphically represented with the help of PyCharm. The contribution in this paper is a methodology that automates the sentiment analysis of the divergent dataset with context of apprehension of tweets.

Keywords: Sentiment Analysis, Opinion Mining, Polarity detection, Deep Learning, Natural Language Processing, Machine Learning

1. Introduction

The evolution of technology has aggrandized the transparency in the digital world, every information is easily available online through social networking sites and people have also strayed the concept of privacy. Individual, professional, product or circumambient information is shared on daily basis in social media. Due to vast consignment of information, which is available online, is utilized for various research and product reviews. This improvement has led to an advancement in data science and evolution of big data [1].

Twitter is extensively used social networking site, every user can share the information of 280 characters length. Since the Bengaluru traffic congestions are become matter of concern, the data on Bengaluru traffic is extracted from Twitter and context based model is build and evaluated for better traffic management. Twitter enables to tweet anyone on the microblog by mentioning the username proceeded by @ character. The tweet will be visible to all the followers of the mentioned user. Any tweet which is prefixed by # character is known as hash tagged and are used for commenting. Tweets also provide URL's and pictures. The average number of tweets per day remain 500 million supporting 35 different languages all over the world.

Sentiment classification is the field of Natural language processing used for extracting opinions and emotions from the text, the opinions are mined with respect to subjectivity, polarity, spam detection, summarization, and argument detection, whereas the emotions are mined to with respect to classification, polarity detection, cause detection. The same is applied, with respect to different level of analysis, such as document level, sentence level, aspect or entity level in the given document or text. The classification also takes place in methods, that is lexicon based and machine learning based. In machine learning model, the classification takes place based on supervised, unsupervised and semi-supervised learning models. Supervised learning model has corpus, with pre-training model for feature extraction and sentiment detection. In the unsupervised model, the group of unlabeled data is used of learning purpose, the clustering techniques are employed to group the unseen patterns of data. The semi-supervised makes use of both unlabeled and labeled data for classification.

The furtherance in Sentiment analysis has flexed in divergent applications and different aspects in applications. The main key is the person's mindset, to elevate the information of social networking sites. As human psychology is to seek approval from others, is highly gratified through the social networking sites such as Twitter, Instagram, YouTube and so on. Few examples of benefits of advancements involve brand management, political analysis, traffic sentiment analysis, product analysis, natural language processing [2], medical analysis [3], monitoring social media, unusual events tracking [4], market analysis and prediction [5], employee sentiment analysis, recommender systems, and so on.

Objective of the research paper is to demonstrate accuracy knowledge based technique model of Bengaluru traffic. The Bengaluru traffic data was built based on scrupulous noise removal and context based data refinement and storage. It is also remarkable the dataset is well preprocessed and dataset is trained to produce good accuracy for the existing machine learning and deep learning models. The Research paper is organized as introduction, which gives brief information on the research along with objective and motivation, followed by related work describing some recent works done in social media analysis, followed by methodology, describing the approach adopted in the classification model and results and comparison shares the brief information on various model performance of divergent dataset followed by conclusion.

2. Motivation

Evaluation of Technology and advancement in digital world also caused spiraled impact on daily life and the life style. The deep learning and machine learning has brought lot of alleviations in the field of NLP, Vision, Weather Robotics and there is also future scope for automatic traffic management. The motivation of the research work is to explore different machine learning algorithms with deep learning model LSTM with respect to different datasets.

3. Related Work

Easy access to enormous information has benefited the seeker to make important decision policies and Adebayo Abayomi-Alli et al. [6], coaxed on yahoo-yahoo hash tag tweets from

Nigeria, which is elucidated as cybercrime, for sentiment analysis and opinion mining, the 5500 tweets were obtained from twitter and preprocessed and pre-trained, four different methods such as dictionary to allocation methods terms as Latent Dirichlet Allocation, Latent Semantic Indexing, Valence Aware Dictionary and VADER were used for Sentiment analysis, Multidimensional Scaling graphs for the same generated. The VADER technique proved to be more efficient as it uses the emoji's for sentiment analysis. The sentiment of the statements depend upon the context as well. Siti Khotijah et al. [7], had influenced on context based approach for twitter data, the work focuses on retrieval of twitter data and preprocessing the data followed by paraphrasing, for finding the context of the data, then the classification technique Long short term memory networks with recurrent neural network was applied on balanced and unbalanced data of English and Indonesian languages. The results of Indonesian balanced data set was yielding to 88% and imbalanced data to 76%, whereas the English balanced data accuracy was recorded as 79% and imbalanced as 54%. It is very evident the balanced data set capitulated the highest accuracy, the future scope of the work was to relate the context and lexicon based methodology. The Twitter sentiment can also be performed on Political based data, Rajesh Bose et al. [8], swayed on political sentiment classification for 14th Gujarat Legislative Assembly Election, 2017 to predict the public opinion. NRC emotion lexicon was used for multiclass classification models such as ParallelDots. The data was collected for the period of 3 months, 1000 tweets were extracted from the Twitter. Data was visualized based on the sources, such as android, iPhone and web users. It was concluded that 12% of Indians have communicated via English Language and 55% of them are positive about the ruling party. Identifying the correlation between the events was the proposed future plans. The Covid-19 outbreak impacted the lives of many, many researchers explored the impact of virus. Rubul Kumar [9] worked on heterogeneous ensemble framework for covid 19 tweets classification, pandemic was a real stupor for the modern world, the tweet were collected during this period of time, an ensembling learning model was proposed to train and test the preprocessed tweets. The number of tweets collected were around 142,334 in the first dataset and second the dataset consists of 65,688 in consequent cycle. After the preprocessing of tweets, vectorization using TF-IDF was performed explicitly for bigrams and unigrams, different machine learning algorithms such as NB, DT, KNN, LR, and RF were used to build the ensembling learning model. The model capitulated the accuracy to 94%, Mohammed Bouhabous et al. [10], contributed something contemporary to the field of sentiment analysis, by forecasting the hybrid sentiment via BERT in crime predictions. The research work involved in foreseeing the crime and security threats via real time through Twitter. Two approach is carried out, the first one is bidirectional encoder representations from transformers and the second one is lexicon based model, using both the hybrid model was created. The hybrid model which proposed, produced better accuracy which had data loading from dictionaries with the blend of positive, negative, incremental, decremental and inverse words, then labeling was carried out after the POS tagging, followed by sentiment classification after the preprocessing of data, tokenization and training the model. Al-Shabi et al. [11], contributed the work on evaluating the lexicons for sentiment analysis, the data from the twitter was extracted and preprocessed, two set of datasets such as Stanford and Sandars were the source. sentiments were computed with respect to different opinion lexicon models such as Vader, Sentiwordnet, Afinn-111,

Sentistrength, Liu and Hu opinion lexicon. Sentiments and polarity is compared and determined with respect to these models and accuracy and F1-measure was tabulated. Vader lexicon produces the better accuracy up to 72% compared to other models. Sentiwordnet was tabulated with the minimum accuracy of 53%.

Irfan Ali Kandhro et al. [12], applied the naive bayes group of classifiers and developed a tool for sentiment analysis termed as PSAT, PSAT analyzed the sentiment about the New Zealand Cricket team cancelling the tour to Pakistan, the Arabic dataset amounting to 3000 were randomly picked from different social networking sites such as Facebook, Instagram, Cricinfo and cricbuzz and used for training and testing the data, noise removal, stemming was carried out. The accuracy was computed to 75%. The model failed to differentiate unigram, bigram and n-grams. The sentiment of cricket being played in different places like inside and outside the country were analyzed, The conclusion was, the place to play the cricket match does not matter much to the fans than the Cricket game itself. Satheesh Kumar et al. [13], described the exploration of sentiments, semantic based feature selection was observed via lexical resource of WordNet, the objective was to analysis the best feature selection methods for sentiment classification, Naive bayes, FLR, and AdaBoost classifiers were used in the research work. the contemporary dimensional view of person's perception towards the opinion mining were discussed. Two different datasets such as movie review and medical query was sources from online and highest accuracy was tabulated to 85% for medical query dataset. The future works involves the study and investigation of enhancement of accuracy in classification problems. Hoong - Cheng Soong et al. [14], proposed the study on essential of opinion mining and sentiment analysis, which is a detailed survey on recent approaches in sentiment analysis and machine learning techniques, according to the survey the sentiment analysis broadly carried out through two methods, i.e. Lexicon based and machine learning models. The sentiment analysis itself carried out in different levels based on the context such as opinion, emotion, granularity and so on. The research work asserts the lexicon based model outperforms the machine learning models. It was proposed to use semi-supervised model to yield the highest accuracy.

4. Methodology

The research work comprises of classifying the sentiments of four different domain of datasets that is Bengaluru Traffic, Fake news, Movie review and Finance Sentiments through four different classifiers XGB, Extra trees, KNN, and LSTM on different performance metrics like F1-score, Precision, Sensitivity, Accuracy, MSE and Log loss. The figure 1 shows the sentiment classification methodology used through different machine learning models and deep learning model. The LSTM model uses the one hot representation and word embedding techniques for data preparation. The machine learning model uses the vectorization technique for data preparation before the classification.

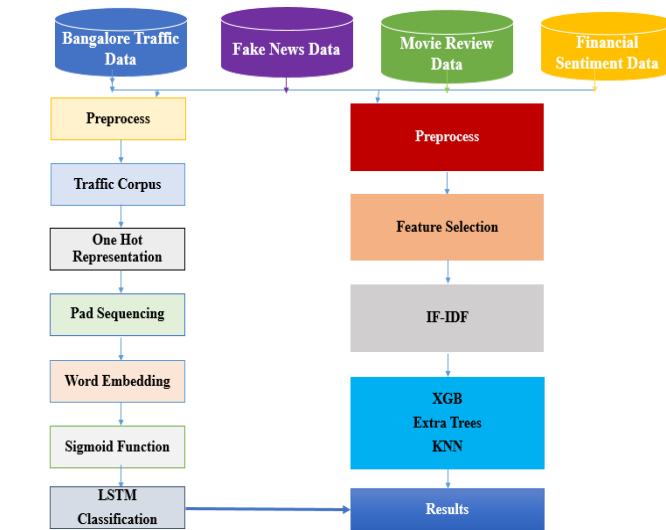


Figure 1: Sentiment Classification

4.1. Data

The Research work comprises of 4 divergent dataset, 3 datasets financial sentiment dataset, movie review dataset, fake news classification dataset has been extracted from the Kaggle and the Bengaluru Traffic dataset has been constructed through the tweets extracted from the twitter API, after the twitter API authentication. Once the data has been gathered, the pre-processing was carried out, the noise in the data such as stop words, hashtags, punctuations, white spaces, user, mention, emotions were eliminated through R tool, textblob and python. The Bengaluru Traffic data was well refined and built based on the Bengaluru traffic context.

4.2. XGBoost Classifier

XGB classifier is known as Extreme Gradient Boost decision tree based ensemble model similar to extra trees classifier. The ensemble based models are two types, one for bagging and another for boosting. The random forest is like bagging model, which was illustrated in the previous research work where the tree construction and data training was carried out collateral. In sequential model we construct a tree and based on the residuals all other trees are built. XGBoost is also used to classify multi class classification problem. In XGB first the base model is constructed based on probability of 0.5 value of residual, followed by binary classification with two categorization such as positive sentiment and negative sentiment. After the base model the summation of residual will be taken.

$$\text{Output} = \frac{\sum (\text{Residual})^2}{\sum (p(1-p) + \lambda)}$$

$\lambda \rightarrow$ hyper parameter (λ will be 0 for the computation)

Once the summation is completed, the same will be repeated for different trees finally the gain value is calculated through the similarity rate of all the trees by subtracting the root value from the sum. During these phase whichever branch value gives highest value split that

tree will be selecting the denominator, will also yield to cover value, to help the post pruning branch out trees.

The log of odds are used to compare the predicted probability value with base model, again the sigmoid function is applied to find the new probability of residuals, through which again the new decision tree is created.

$$\sigma[(O + \alpha(T_1) + \alpha(T_2) + \dots + \alpha(T_n))]$$

This process continues till we arrive at the minimal probability value and XGboost has produced very good results in our experiment following LSTM Model. The figure 2 shows the extreme gradient boost classifier graphically. The figure 3 shows the computational results from PyCharm on extreme gradient boost classifier.

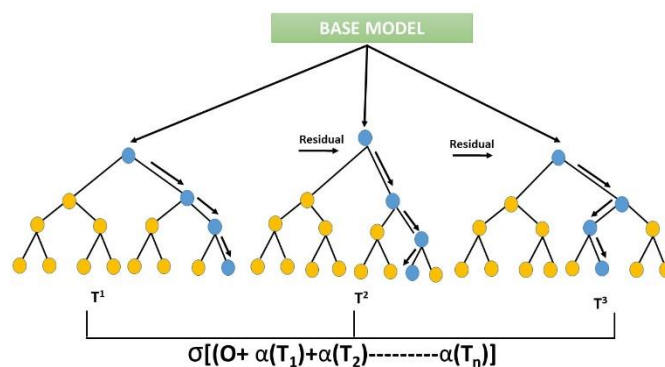


Figure 2: Extreme Gradient Boost Classifier

```

Run: paper10 (1) x
XGB CLASSIFIER for Bengaluru Traffic
Training Accuracy : 0.9171560664750719
Testing Accuracy : 0.8627948262744104
      precision    recall  f1-score   support
0         0.87      0.79      0.82      1619
1         0.86      0.92      0.89      2524

 accuracy          0.86      0.86      3943
 macro avg         0.86      0.85      0.86      3943
 weighted avg      0.86      0.86      0.86      3943

Loss value for XGB(Traffic) 4.7389689805623965
Mse value for XGB(Traffic) 0.13720517372558966
  
```

Figure 3: Computation of XGB

4.3. Extra tree Classifier

Extra tree is known as extremely randomized trees first proposed by Pierre Geurts et al. [15], in 2006, it mainly worked on the concept of randomizing the attribute selection and cut point in decision trees, in other words it is a randomized decision trees, the main goal of the algorithm was to increase the accuracy in classification problem. The computation results of extra tree classifier from python is represented in the figure 4.

Table 1: Extra tree Algorithm

Extra tree Algorithm:

To_split_node(N)

Input: subset of data

Output: if attribute a is < ac split

if split==true

then

attribute select a random k attribute {a1,a2,...ak} among N nodes

pick a random splits from maximal and minimal value

return a split

	precision	recall	f1-score	support
0	0.95	0.91	0.93	1619
1	0.94	0.97	0.95	2324
accuracy			0.94	3943
macro avg	0.95	0.94	0.94	3943
weighted avg	0.94	0.94	0.94	3943

Loss value for extra trees(Traffic) 1.953402758224604
Mse value for extra trees(Traffic) 0.056555921886888155

Figure 4: Extra tree classifier for Bengaluru Traffic

4.4. KNN Classifier

K-Nearest Neighbor is a classification technique used to find the nearest neighbor, based on K value. The values is set to determine the maximum number of nearest neighbor. The K value is set to 5. Euclidian Distance is used to find the distance between two points, i.e., $p(x_1, y_1)$ and $p(x_2, y_2)$

Euclidian Distance

$$ED = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K value is assigned as odd to perform the balanced classification. KNN can also be applied for regression based data implementation, KNN Value will assigned as 1 and aggregation of weights will be computed when the regression model is activated. KNN will also get influenced by outliers if the outliers are more than the classification will become imbalanced. The KNN is used to classify the four different domains of datasets such as Bangalore Traffic,

Fake news classification, movie reviews dataset and financial sentiments classification. All these data sets contains the sentiments as positive, negative, which is represented as 1 and 0 respectively. The results of which are completed in distinct performance metrics as mentioned in the table 2. KNN library is implemented from `sklearn.neighbors.KNeighborsClassifier` from `scikit`. The figure 5 shows the PyCharm computational results of Bengaluru traffic.

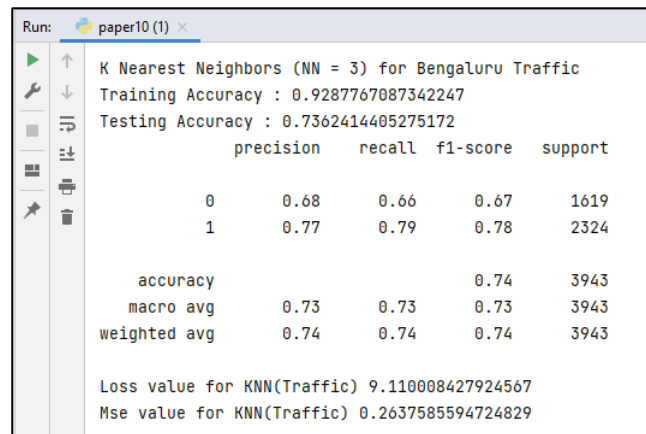


Figure 5: KNN Computation Results on Bengaluru Traffic

4.5. LSTM

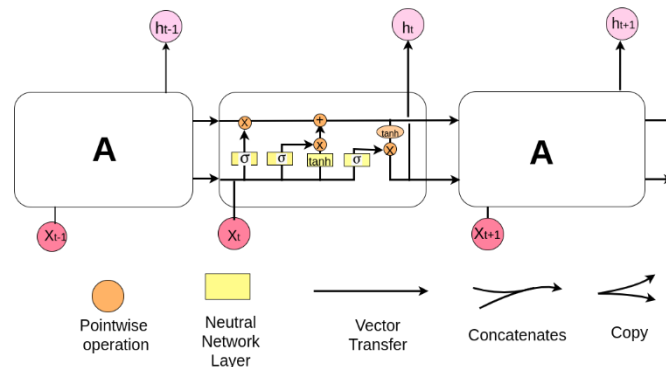


Figure 5: LSTM Model

LSTM abbreviated as Long Short Term Memory networks, broadly used in deep learning models is special type of Recurrent Neural Network works on the memory of previous gates which is implemented to avoid gradient descent problems occur in Recurrent Neural Network, The model comprises of three gates such as forget gate, input gate, output gate and one cell known as memory cell. The figure 6 Illustrates the LSTM model, X is the features are h is the output of each gate, the point wise operator receives the output of previous gate. The figure 7 shows the LSTM computational results through python.

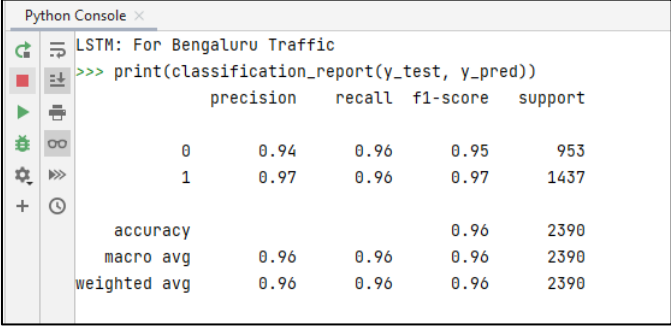
4.6. Epochs

Epochs are measure of model performance in terms of batches, when the dataset traverses forward and backward in the neural network, the loss rate can be reduced by optimizing the problem. The computer will automatically divide epoch into different batches to

accommodate the huge calculations. In LSTM model the epoch was experimented with samples, the epoch size was ranged from of the size of 5, 10, 20 and 50, the epoch size was 50 the while experimenting the LSTM model for all the datasets.

4.7. Batch

In deep learning the dataset is too colossal, it is divided into different smaller assemblage and each of these assemblage will be sent for processing on different number of iterations.



```

Python Console x
LSTM: For Bengaluru Traffic
>>> print(classification_report(y_test, y_pred))
              precision    recall  f1-score   support

     0       0.94         0.96         0.95         953
     1       0.97         0.96         0.97        1437

 accuracy          0.96         0.96         0.96        2390
 macro avg         0.96         0.96         0.96        2390
 weighted avg      0.96         0.96         0.96        2390
  
```

Figure 7: LSTM computational results

5. Results and Discussions

The table 2 represents the comparison of different datasets such as Bengaluru Traffic (KBT), Fake news, movie review, and financial classification through different machine learning models such as XGB, Extra trees, KNN and deep learning model LSTM. The contribution for the research remains the novel dataset of Bengaluru traffic(KBT) pre-processed and fine-tuned producing highest results in terms of accuracy, MSE, Log loss, precision, sensitivity, F1-score for both positive and negative classes for models such as extra trees and LSTM.

Dataset	Classifier	Class	F1-Score	Precision	Sensitivity	Accuracy	MSE	Log Loss
Bengaluru Traffic(KBT)	LSTM	Positive	97%	97%	96%	96%	0.04	1.4
		Negative	95%	94%	96%			
	XGB	Positive	89%	86%	92%	86%	0.13	4.73
		Negative	82%	87%	79%			
	EXTRATREES	Positive	95%	94%	97%	94%	0.05	1.95
		Negative	93%	95%	91%			
KNN	Positive	78%	77%	79%	74%	0.26	9.11	
	Negative	67%	68%	66%				
Fake News	LSTM	Positive	81%	80%	82%	80%	0.19	6.79
		Negative	80%	81%	79%			
	XGB	Positive	75%	80%	70%	76%	0.23	8.11
		Negative	78%	73%	83%			
	EXTRATREES	Positive	82%	86%	78%	83%	0.17	5.90
		Negative	84%	80%	87%			
	KNN	Positive	73%	72%	74%	73%	0.27	9.44
		Negative	72%	73%	71%			

Movie Review	LSTM	Positive	74%	75%	74%	74%	0.25	8.97
		Negative	74%	73%	74%			
	XGB	Positive	86%	84%	87%	85%	0.14	5.02
		Negative	85%	87%	84%			
	EXTRATREES	Positive	85%	85%	85%	85%	0.14	5.07
		Negative	85%	85%	86%			
KNN	Positive	73%	69%	76%	71%	0.28	9.85	
	Negative	70%	74%	67%				
Financial Sentiments	LSTM	Positive	88%	86%	89%	78%	0.21	7.44
		Negative	21%	23%	19%			
	XGB	Positive	90%	86%	94%	82%	0.17	6.21
		Negative	24%	36%	18%			
	EXTRATREES	Positive	88%	84%	93%	80%	0.20	7.07
		Negative	12%	19%	9%			
KNN	Positive	90%	85%	96%	82%	0.17	6.10	
	Negative	16%	32%	10%				

Table 2: Comparison of different classifier with the datasets

5.1. Performance Metrics

The results of the model is tabulated in the table, the performance of the model is computed in terms of accuracy, precision, sensitivity, f1-score, log loss and mean square error

5.2. Confusion matrix

The confusion matrix represents the model performance in terms of actual values and predicted values as true positive(TP), false positive(FP), false negative(FN), true negative(TN). The figure 8 shows the confusion matrix of LSTM model, which produces the true positive as 914 and true negative as 1379 and capitulating the accuracy to 96%.

```
>>> confusion_matrix(y_test, y_pred)
array([[ 914,   39],
       [  58, 1379]], dtype=int64)
```

Figure 8: Confusion Matrix for LSTM

Accuracy: It is a measure of precise classification

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FN+FT)}$$

Precision: Model's ability to classify positive value correctly

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Sensitivity: Model's ability to predict positive value correctly

$$\text{Sensitivity} = (TP / (TP + FN))$$

F1-Score: Harmonic mean of Sensitivity and Precision

$$\text{F1-Score} = (2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall}))$$

Log Loss: Log Loss is used to measure the performance of classification model in terms of being able to predict the expected outcome. It is the average of algorithm losses.

$$\text{Log loss} = \frac{1}{n} \sum (y * \log(p) + (1 - y) * \log(-p))$$

n : Number of values

y : Actual Values

p : Probability of predicted values

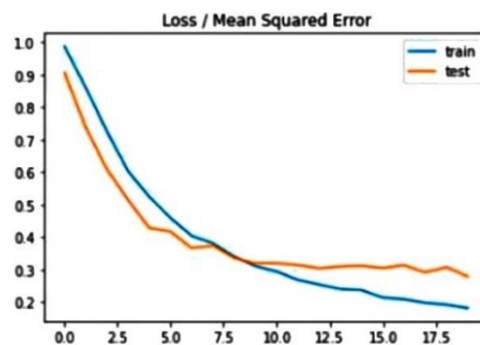


Figure 9: Mean squared error Computed

5.3. Mean squared error (MSE)

MSE is used to find the difference between predicted and input values. IT provides the average of squared difference. The figure 9 shows the MSE for LSTM model as each loss value is compared.

$$\text{MSE} = \frac{1}{n} \sum (y - y^{\wedge})^2$$

n : Number of values

y : Actual Values

y : Predicated values

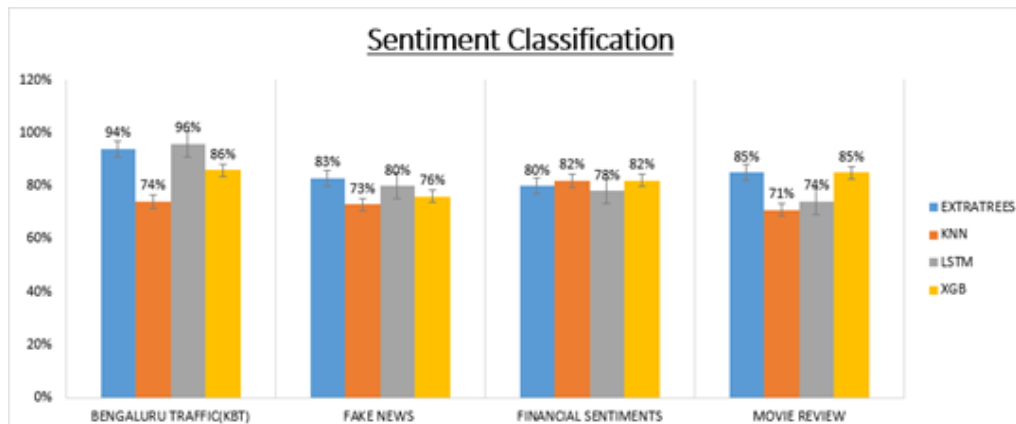


Figure 10: Sentiment classification

The figure 10 shows sentiment classification of machine learning techniques producing better results for Bengaluru Traffic dataset LSTM procuring 96% and extra trees 94% compared with other datasets.

6. Conclusion:

Though it is possible to test the model with ensemble methods, addressing the problem with basic model was adequate to find the polarity of tweets of Bengaluru traffic. In this research paper, the various ML algorithms such as XGB classifier, extra tree Classifier, K-Nearest neighbor classifier, and long short term memory networks are compared and results are completed with respect to four different datasets such as Bengaluru traffic, Fake News, Movie Review Data and financial data from which LSTM yields 96% of results for Bengaluru Traffic dataset. It is noteworthy that all the classifiers capitulate better results for Bengaluru traffic as tabulated in the table. The Bengaluru traffic data is also termed as KBT i.e. Knowledge based techniques because of its Novelty of context based data stored and classified for better results. From the tweets, the different reasons for traffic congestion such as road repair, accident and route diversions are identified. The future works in the field can be simulation based model for traffic predictions

References and notes

1. Sandeepa Kannangara, "Mining twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection," in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 751-752.
2. Yuexiong Ding, Jie Ma, and Xiaowei Luo, "Applications of natural language processing in construction," Automation in Construction, vol. 136, p. 104169, 2022.
3. Gaurika Jaitly and Manoj Kapil, "An Improved Learning Approach For Medical Sentiment Analysis And Opinion Mining For High Classifications," International Journal of Research in Engineering and Science (IJRES), vol. 9, no. 3, pp. 37-49, 2021.
4. Khadijha-Kuburat Adebisi Abdullah, Sodimu Segun Michael, and Odule Tola John, "Opinion Mining And Event Detection Analysis Of Coronavirus Twitter Data Using Ensemble Deep Learning Models," FUW Trends in Science & Technology Journal, vol. 6, no. 3, pp. 921-927, September 2021.

5. Christina Nousi and Christos Tjortjis, "A Methodology for Stock Movement Prediction Using Sentiment Analysis on Twitter and StockTwits Data," in 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), 2021, pp. 1-7.
6. Adebayo Abayomi-Alli, Olusola Abayomi-Alli, Sanjay Misra, and Luis Fernandez-Sanz, "Study of the Yahoo-Yahoo Hash-Tag Tweets Using Sentiment Analysis and Opinion Mining Algorithms," *Information*, vol. 13, no. 3, p. 152, 2022.
7. Siti Khotijah, Jimmy Tirtawangsa, and Arie A. Suryani, "Using lstm for context based approach of sarcasm detection in twitter," in Proceedings of the 11th International Conference on Advances in Information Technology, 2020, pp. 1-7.
8. Rajesh Bose, Raktim Kumar Dey, Sandip Roy, and Debabrata Sarddar, "Analyzing political sentiment using Twitter data," in *Information and communication technology for intelligent systems.*: Springer, 2019, pp. 427-436.
9. Rubul Kumar Bania, "Heterogenous Ensemble Learning Framework for Sentiment Analysis on COVID-19 Tweets," *INFOCOMP Journal of Computer Science*, vol. 20, no. 2, 2021.
10. Mohammed Boukabous and Mostafa Azizi, "Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 2, pp. 1131-1139, 2022.
11. M. A. Al-Shabi, "Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining," *IJCSNS*, vol. 20, no. 1, p. 1, 2020.
12. Irfanali Kandhro et al., "PSAT-Based Sentiment Analysis: For Text And Data Mining," *Journal of Tianjin University Science and Technology*, vol. 55, no. 4, pp. 576-591, 2022.
13. R. Satheesh Kumar et al., "Exploration of sentiment analysis and legitimate artistry for opinion mining," *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 11989-12004, 2022.
14. Hoong-Cheng Soong, Norazira Binti A. Jalil, Ramesh Kumar Ayyasamy, and Rehan Akbar, "The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and techniques," in 2019 IEEE 9th symposium on computer applications \& industrial electronics (ISCAIE), 2019, pp. 272-277.
15. Pierre Geurts, Damien Ernst, and Louis Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3-42, 2006.
16. Pengfei Liu et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.