**Title : Spam Text Detection using Naïve Bayes and various Deep learning Models**

**Problem to be Solved :** Spam Text Detection using Naïve bayes and comparing its performance with that of the Deep learning models such as CNN(Convolutional model), LSTM(Long short term memory) , GRU( Gated recurrent Unit) and Bidirectional LSTM .

**Small background on models Used :**

**Naïve Bayes:** Based on finding the probability an SMS is spam, given it contains certain words, which performs well as it is very simple, and easy to use, needs less training data, Handles continuous and discrete data and can make predictions even if some feature is missing by altering decision rules.

**Convolutional Neural Network:** CNN performs excellently in extracting n-gram features at different positions of a message through the use of multiple layers. Firstly, the input sentence split into embedding-words which are low-dimensional representations created by models such as word2vec or GloVe. Words are then broken up into certain features and fed into a convolutionary layer which vectorizes the features. The output is then pooled to a number. This number is then fed to a fully connected neural network which makes a classification decision based on the weights assigned to each feature within the text. The ReLU activation function is used at every layer of the filtering except the last one which uses sigmoid function to classify an email either as a ham or a spam.

**Long Short Term Memory:** LSTM is able to handle word sequences of any length and capture long-term dependencies. This can be useful for identifying dependencies between words in a spam message and, hence, LSTM can determine whether the message is spam or not from the initial words of the message

**Gated Recurrent Unit:** GRU performs best as it can be trained to keep information from long ago, without washing it through time or remove information which is irrelevant to the prediction.

**Bidirectional LSTM:** BiLSTM works excellently because for some of the sentences, the context information is at the end of the sentence and without the context info, ambiguities can raise, so reading the sentence in Bi direction helps the model to determine the exact meaning of the word.

***Data Collection and Preparation:***

**Data Set and its Background :**

* Corpus: SMS Spam data collection v1 contains one set of SMS messages in English of **5,572** messages, classified as **ham (4825)** or **spam (747)**.

**Classification of messages on v1 column with message content in v2**

****

**Data Pre-Processing :**

1. **Labelling :** SPAM and HAM categories as 0 and 1
2. **Case Folding :** All text is converted to lowercase to homogenize the data
3. **Punctuation and StopWord removal :** Removing Stopwords using nltk <https://www.nltk.org/book/ch02.html> and punctuation such as (!”#$%&()\*+,-./:;<=>?@[\]^\_’{|}~\t\n).
4. **Lemmatization** : Converting every word into meaningful base form in the given corpus of data <https://www.datacamp.com/community/tutorials/stemming-lemmatization-python>
5. **Word Tokenization and Sequencing** : Assuming space as a delimiter, tokenizing the sentences into tokens(word) and giving a number to each word
6. **Post-Sequence Padding and Truncation**: Using the maxlen 20 as , padding 0 at the end of the sequences for sentences of length less than maxlen and truncating the end of the sentences when it exceeds maxlen.

***Experimental Design :***

**Partitioning**

Splitting the SMS Spam detection Dataset into **80-20** and **70-30** datasets

Where training sets are of split 80 and 70 and test sets are of split 20 and 30.

**Training phase** : Propose 5 different deep learning classifiers with train-test split: Multinomial NB classifier, CNN Model, LSTM , Bidirectional LSTM and GRU

**Scoring Phase :**

**Confusion Matrix**: Prediction of model vs Actual data

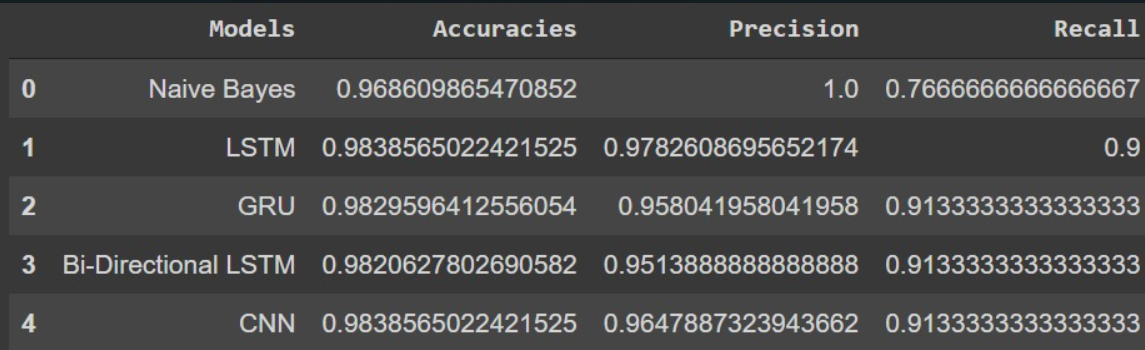
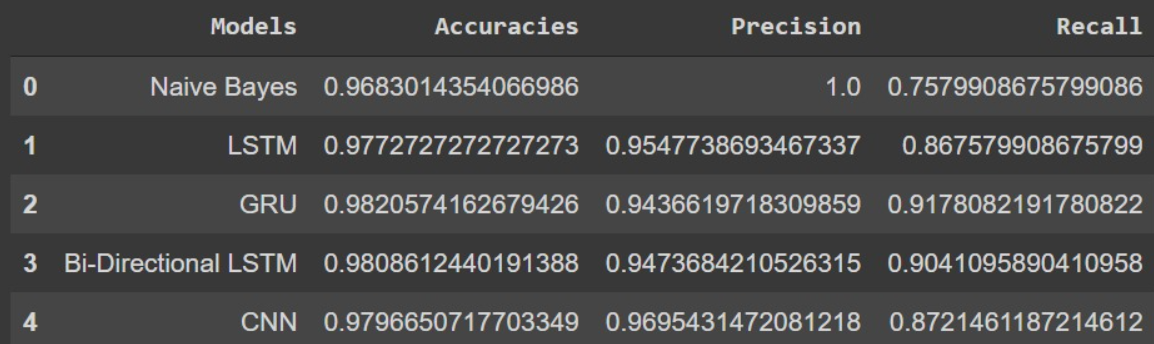
**Accuracy**: ’ Accuracy of model in classifying the data set

**ROC**: , metric used to attain optimal models from suboptimal ones

**Precision-Recall**: If no-match set is very large relative to test set, this metric might be in favor of ROC

***Naïve Bayes classifier vs other classifiers for SMS Spam detection***

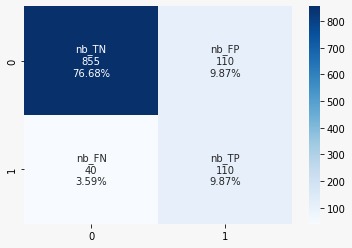
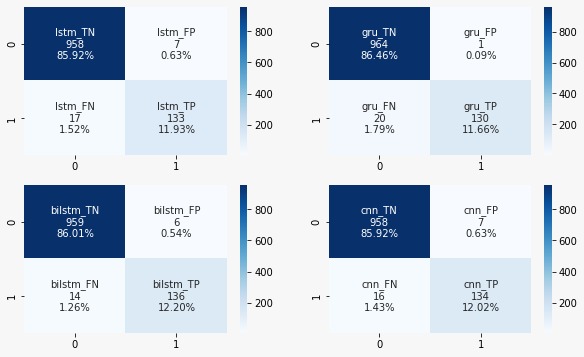
**80-20 Split 70-30 Split**

** **

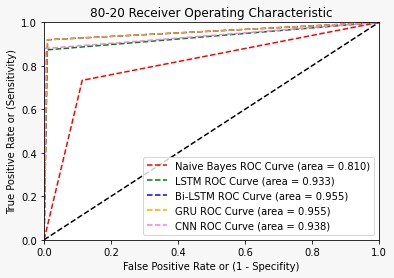
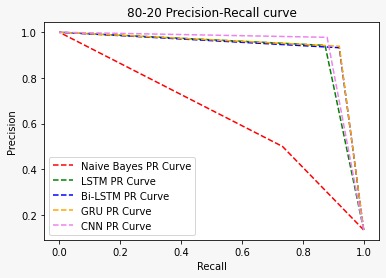
**Determining data split for further comparisons**

Based on above splits, it is determined that 80-20 split has equivalent or better results on all metrics and further analysis is based on this split

**Confusion Matrix for: Naïve Bayes Confusion Matrices for – CNN, LSTM, GRU, BiLSTM:**

** **

**ROC Curves - AUC Under Curves : PR Curves - AUC Under PR Curves**

** **

**Results : Accuracies , Precision , Recall Values for all the 5 models are shown below :**

**Model validation against literature model :**

Based on \_\_\_\_\_\_\_, using the data corpus, SMS corpus, the best metric was Accuracy, followed by Recall and then Precision (Accuracy > Recall > Precision). In our model, we achieved the same trend of Accuracy > Recall > Precision. However, the absolute numbers vary because the Naïve Bayes model used in literature was modified with FP growth and our Model used a multinomial Naïve Bayes.

**Changes During Implementation :**

**Comparing Literature vs Our model:**

**CNN:**

From 1Gauri Jain et al. where CNN, amongst other classifier models, was used on social media platforms’ text messages such as Facebook and Twitter, the CNN beat out NB on all evaluation metrics – Precision, Recall, Accuracy and F score.

However, it has to be noted that the paper was based on short-texts spam identification, which is what tends to be the length of messages in social media as opposed to long texts in emails. This difference influences the models differently as there is little or more data to classify a text message as spam or ham.

From 2Wael Hassan Gomaa’s paper, where he again used short-text spam detection, which is spam SMS, it is seen that deep learning techniques outperformed classical machine learning techniques. The best machine learning technique (Gradient Boosted Trees) which gave an accuracy of 96.86% was much lower than the best deep learning technique (RDML) which gave an accuracy of 99.26%. Hence we are able to conclude, at least in general, for short text message filtering, deep learning techniques perform better than machine learning.

3Zhuo Li et al. in their paper did long-text spam detection on email classification. In their paper they compared Naïve Bayes with three-layer Neural Network and logistic regression. It was determined here however that Naive Bayes algorithm outperformed logistic regression and neural networks. The best accuracy (across all metrics) from Naïve Bayes came at 91.5% with the three-layer Neural Network 88.3%.

4Pradeep Kumar Roy had a difficult task as well when he tried to filter out spam questions in internet Question-and-Answer websites such as Quora, Yahoo! Answers, Stack Exchange etc. This means that his data set ranges from short to long and the spam filtering algorithm has to accommodate for these differing lengths. From his results, it shows 2CNN still performs optimally with an F1 score of 98% in filtering out Insincere questions. Naïve Bayes has only 53% F1 score on filtering out insincere questions and the best machine learning technique still stays at Gradient Boost Trees with an F1 score of 66%.

**LSTM:**

From SMS Spam Detection Based on Long Short-Term Memory and Gated Recurrent Unit , where LSTM outperformed all the other models (NB, SVM,GRU) in terms of Accuracy (98.18%), precision(90.96)% and catches normal messages as a spam message with 0.74% error, when used on SMS spam dataset proposed by Almeida and Hidalgo , contains SMS text messaging conversations in English language. Here in our project , LSTM outperformed all other deep learning models used ( CNN, BiLSTM, GRU) and machine learning model Multinomial NB in terms of Accuracy (98.3)% and Precision (97.8)%. We can claim LSTM have a performance over other techniques in deep learning particularly for sequential data analysis.

**Future Research:**

1. For future enhancement, we can improve the model with improving the contextual based understanding and also extend the model to perform well on medium-long and long sequences, from the literature analysis the following two combination can lead to future scope.
2. Combination of deep learning models such as Bi-LSTM and CNN with attention can help the deeper understanding of the contextual meaning and improve the recall.
3. BERT(Bidirectional Transformers) model approach with combination of Bi-LSTM can improve the contextual embedding of text in the sentences.

**References :**

1**Gauri Jain et al.** (2019 Jan 2), Spam detection in social media using convolutional and long short term memory neural network, *Annals of Mathematics and Artificial Intelligence (2019)*

2**Wael Hassan Gomaa** (2020), The Impact of Deep Learning Techniques on SMS Spam Filtering, *International Journal of Advanced Computer Science and Applications*

3**Zhuo Li et al.** (2020), Spam Email Detection: Comparison Between Naïve Bayes and Neural Network, *IEEE*

4**Pradeep Kumar Roy** (2020 Jun 21), Multilayer Convolutional Neural Network to Filter Low Quality Content from Quora, *Neural Processing Letters*

**9Pumrapee Poomka et al.** (2019 Mar 19) , SMS Spam Detection Based on Long Short-Term Memory and Gated Recurrent Unit , *International Journal of Future Computer and Communication*

**10Yeshwanth Zagabathuni.** (2021 Aug 25) , Spam text classification using LSTM Recurrent Neural Network, *International Journal of Emerging Trends in Engineering Research*

**11Gauri Jain et al.** (2019 Jan 02) , Spam detection in social media using convolutional and long short term memory neural network, *Annals of Mathematics and Artificial Intelligence*

**12Lingyun Xiang** **et al.** (2020 Sept 12) , Spam Detection in Reviews Using LSTM-Based Multi-Entity Temporal Features, *Intelligent Automation & Soft Computing*