**Contents**https://docs.google.com/drawings/d/sc6m-OSucjdL2OCe64fU0pw/image?w=752&h=1&rev=1&ac=1&parent=16r4Jbrqmh3B6oGDHaV7UHkf0ZWN8xHS-HQw5Nv_KQO0

1. Introduction and Problem Statement
2. Dataset Details
3. Preprocessing Dataset
4. Exploratory Data Analysis
5. Model Creation
6. Results
7. References

https://docs.google.com/drawings/d/sIcXK6U4XIu-gsmQ3zokbAg/image?w=752&h=1&rev=1&ac=1&parent=16r4Jbrqmh3B6oGDHaV7UHkf0ZWN8xHS-HQw5Nv_KQO0**1.Introduction**:

Source data from public data set on BBC news articles.

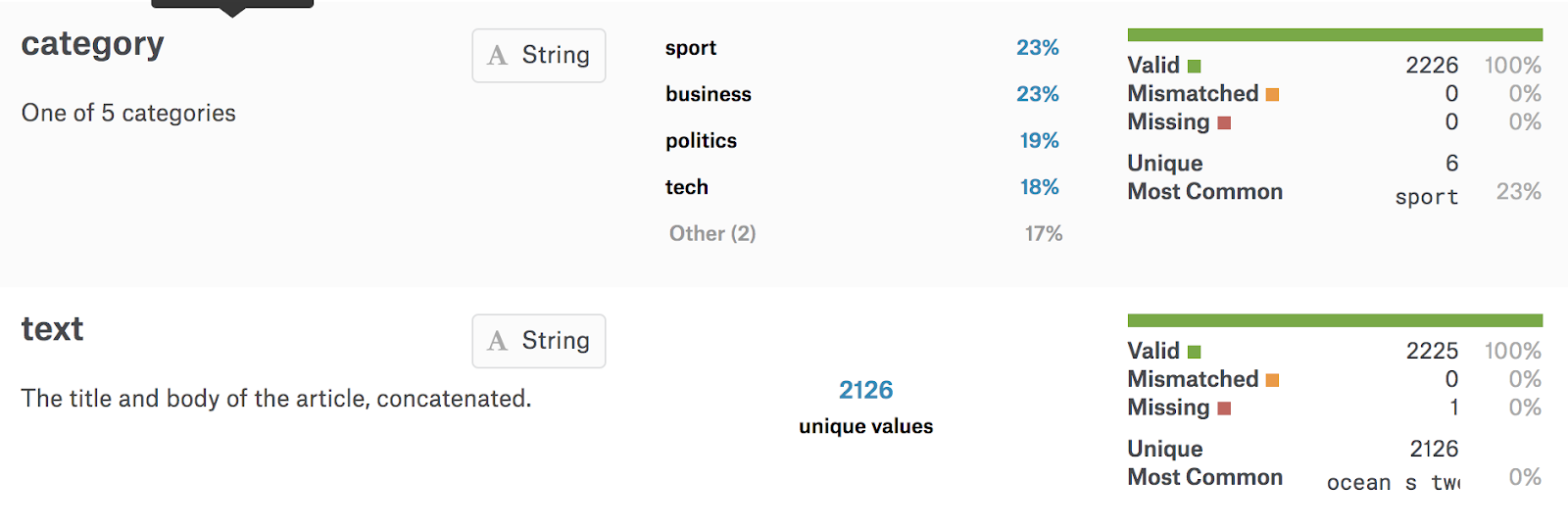
<http://mlg.ucd.ie/datasets/bbc.html>

**Problem Statement**: Theobjective of this project is text classification and predict the category to which the text belongs.

**2. Dataset Description**:

The data set has bbc-text.csv file which is the data from BBC news article.

The dataset has 5 categories, which are Sport, business, politics, Tech, entertainment. And the text as the title and body of the article concatenated.

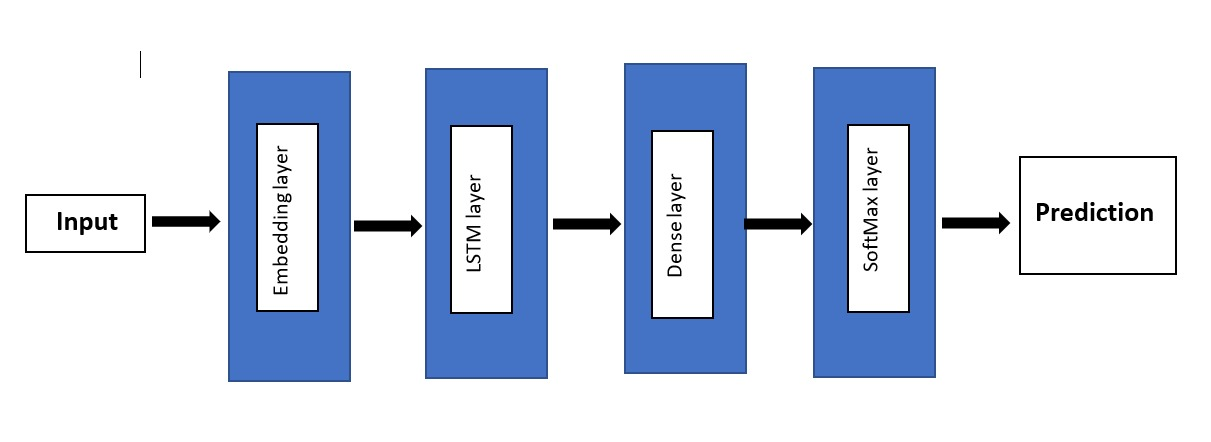


1. **Pre-Processing:**
   * The data set has to be preprocessed to get the best results. We have following steps in preprocessing the dataset, which is the first step of loading the dataset
   * For pre processing we have used the packages which are in R they are:

* library(tidyverse) -> metapackage with lots of helpful functions
* library(tidytext) ->working with text
* library(keras) ->deep learning with keras
* suppressMessages(library(stringr))
* suppressMessages(library(DT))
* suppressMessages(library(igraph))
* suppressMessages(library(ggraph))
* suppressMessages(library(tm))
* suppressMessages(library(wordcloud2))
* suppressMessages(library(wordcloud))
* suppressMessages(library(caret))
* suppressMessages(library(dplyr))
* suppressMessages(library(magrittr))
* suppressMessages(library(plyr))
* suppressMessages(library(tidyverse)) ->general utility & workflow functions
* suppressMessages(library(tidytext)) ->tidy implementation of NLP methods
* suppressMessages(library(topicmodels)) -> for LDA topic modelling
* suppressMessages(library(tm)) -> general text mining functions, making document term matrixes
* suppressMessages(library(SnowballC)) -> for stemming
* suppressMessages(library(glue)) -> for pasting strings

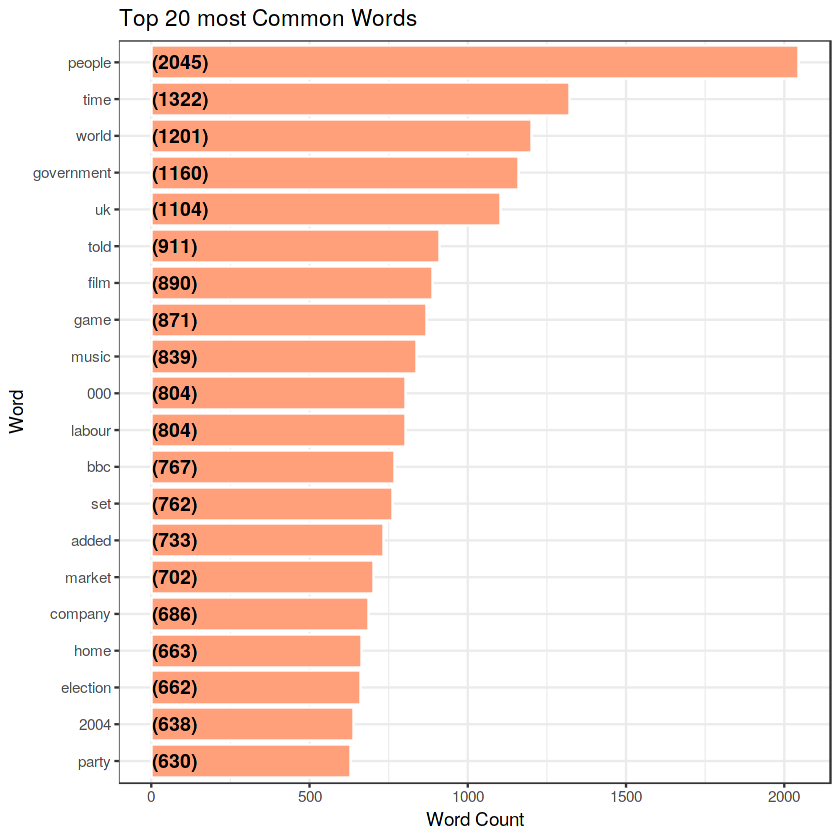
* + The dataset has no missing values. The dataset consists of text and categories.
  + There are five categories :
  + "business"=0,"tech"=1,"sport"=2, "entertainment" = 3, "politics" =4
  + We have used one hot encoding technique to change categorical labels to one hot vector using keras package.

1. **Network Design**



**Exploratory Data Analysis :**

* + The Top 20 most Common words used in the BBC wiki news data can be visualized from the plot below.



As we can see the words **people**, **time** and **world** are the three most commonly used words. We can infer that most the news data revolves around people. Time and world related data are common parameters used in news.

**Sentiment Analysis:**

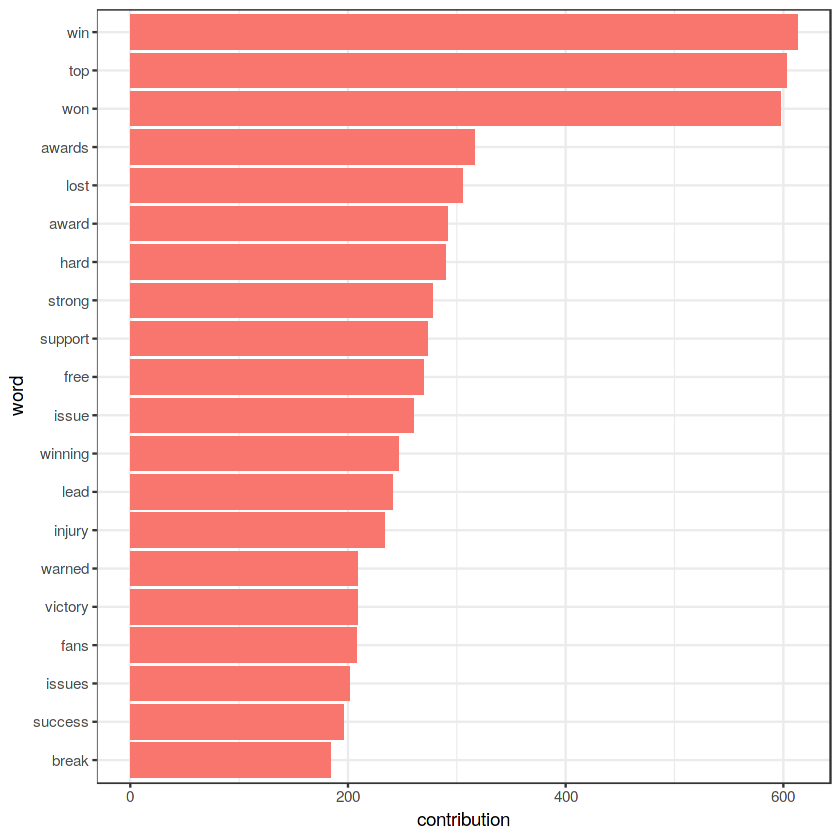
Using the sentiments dataset included as a part of the tidytext package we can obtain several distinct lexicons. They are usually dictionaries of words with an assigned sentiment category or value. In the tidytext package we are provided with three general purpose lexicons:

**AFINN** - > This lexicon assigns words with a score that runs between **-5 and 5**, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

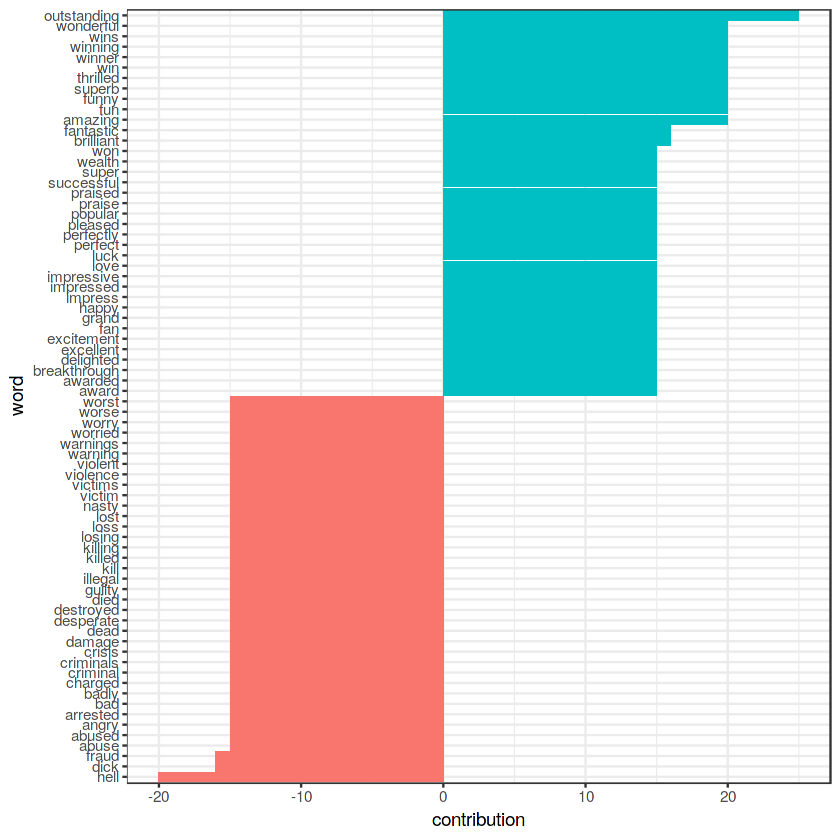
**Bing** -> This lexicon assigns words into **positive and negative** categories based on various parameters.

**NRC** -> This lexicon assigns words into one or more of the following ten categories: **positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust**

* The Top Contributing words and their corresponding NRC sentiment score based on the category



From the plot above we know that the words **‘win’** and **‘top’** are the highest and second highest sentiment score. But from NRC sentiment overall contribution we don’t know which one is having +ve impact(+ve score ) and which one is having -ve impact(-ve score) on reviews. **To distinguish b/w +ve and -ve score we need AFINN Sentiment score.**

* The Top Contributing words and their corresponding AFINN sentiment score based on the category.
  + 

Here we can clearly distinguish the +ve & -ve impact of words on insecurity questions.

**Green** marked words are having **+ve sentiment scores** & **red** marked are having **-ve scores**.

"**Outstanding**" is having the most +ve score whereas **“hell”** is most -ve.

* + The Most frequently used words in the dataset can be visualized using the following code :

train %>%

unnest\_tokens(word, text) %>%

filter(!word %in% stop\_words$word) %>%

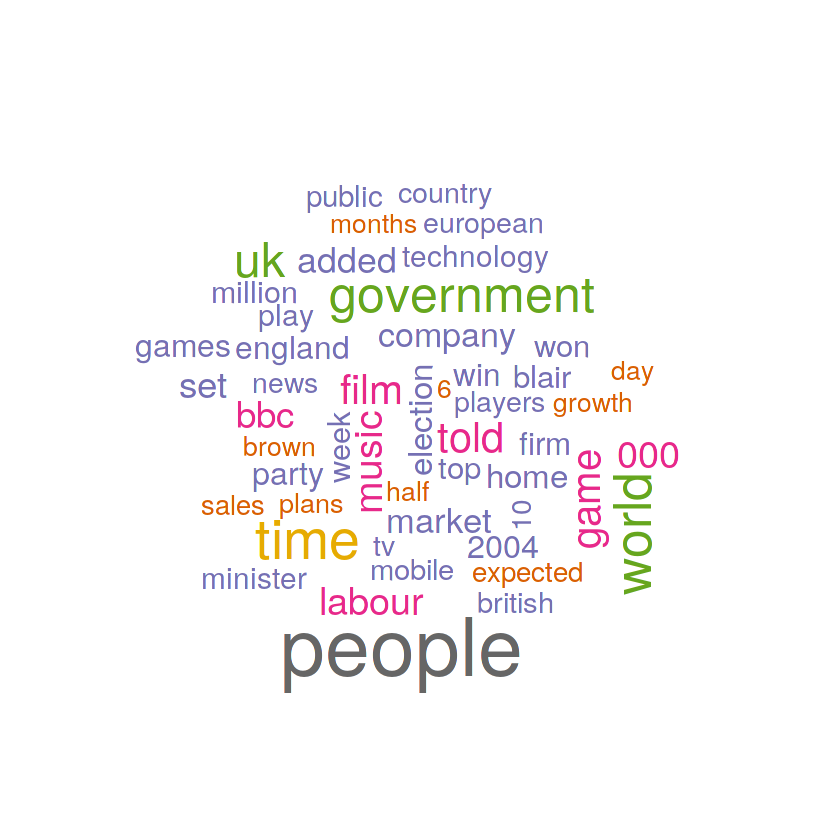
filter(!word %in% custom\_stop\_words$word) %>%

dplyr::count(word,sort = TRUE) %>%

ungroup() %>%

head(50) %>%

with(wordcloud(word, n, max.words = 50,colors=brewer.pal(8, "Dark2")))



**Feature Engineering:**

Word Frequencies with Tf-idf Vectorize

This helps in creating e tf-idf matrix of features in R. It has fit, transform methods to generate its features.

* Term Frequency: This tells how frequently a word appears in the dataset.
* Inverse Document Frequency: This downscales words that appear a lot across document

The [TfidfVectorizer](http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html) will tokenize documents, learn the vocabulary and inverse document frequency weightings, and allow you to encode new documents.

It helps to normalize the dataset for better prediction.

* + We have used word Embeddings Technique which is described as below :

**Word Embeddings :**

[text\_tokenizer()](https://tensorflow.rstudio.com/keras/reference/text_tokenizer.html) was used for pre processing. The tokenizer will transform each review into a sequence of integer tokens. In Keras, the embedding matrix is represented as a "layer", and maps positive integers into dense vectors of fixed size.

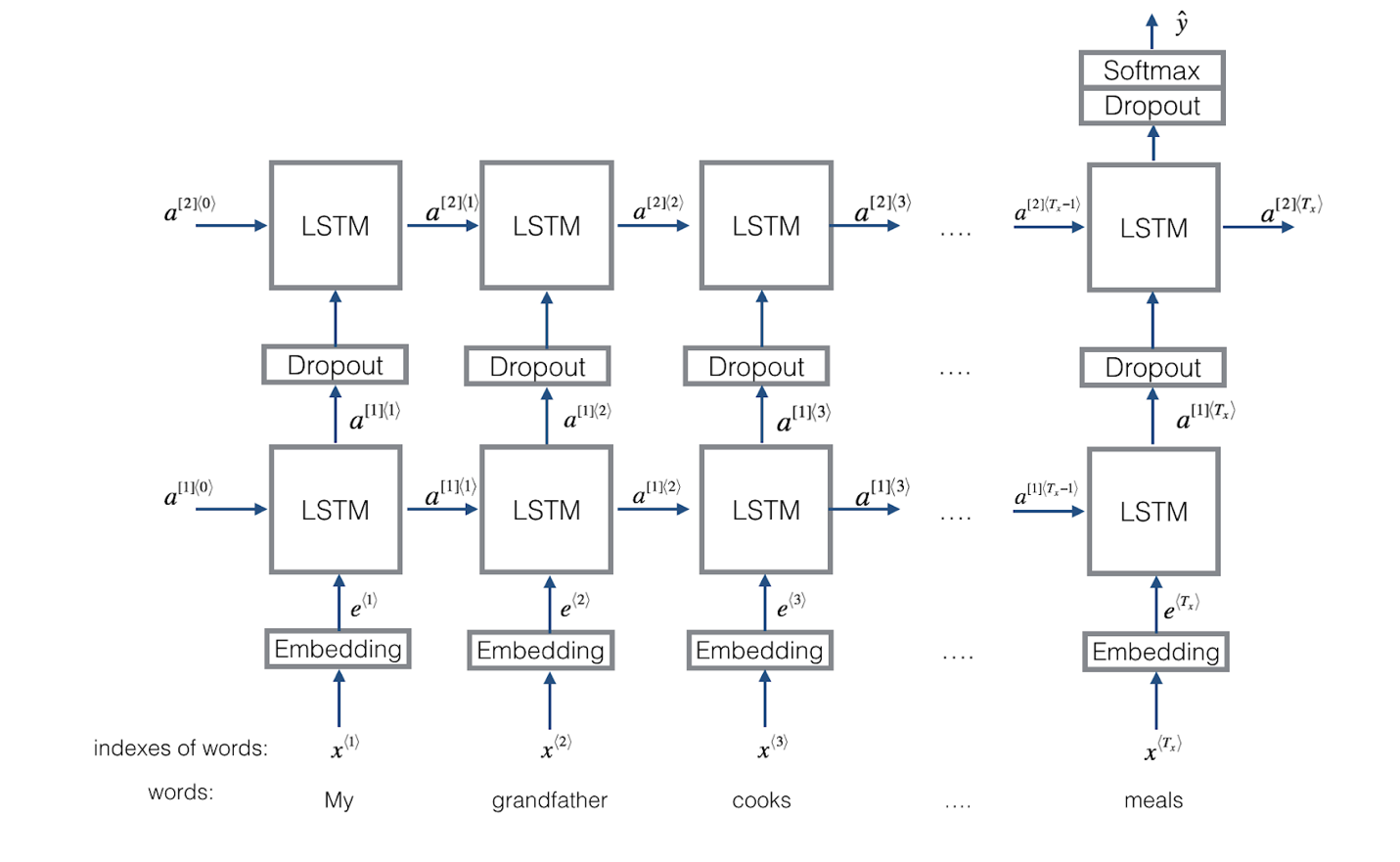
It can be trained or initialized with a pretrained embedding.

We have used pretrained Wiki-news-300d-1M.Vec for getting the word similarities of words which helps in getting better results

The Embedding() layer takes an integer matrix of size (batch size, max input length) as input. This corresponds to sentences converted into lists of indices (integers).

We have used Deep Learning technique for predictive accuracy.

LSTM : The architecture of LSTM can be seen as below :



**7.Results:**

**METRICS:**

We tried our data set on various parameters and found the best accuracy. We have performed several iterations on the dataset to get the perfect model.

The metrics we used:

1. Accuracy
2. Mean-squared- error
3. Mean\_squared\_logarithmic\_error

**Results Table:**

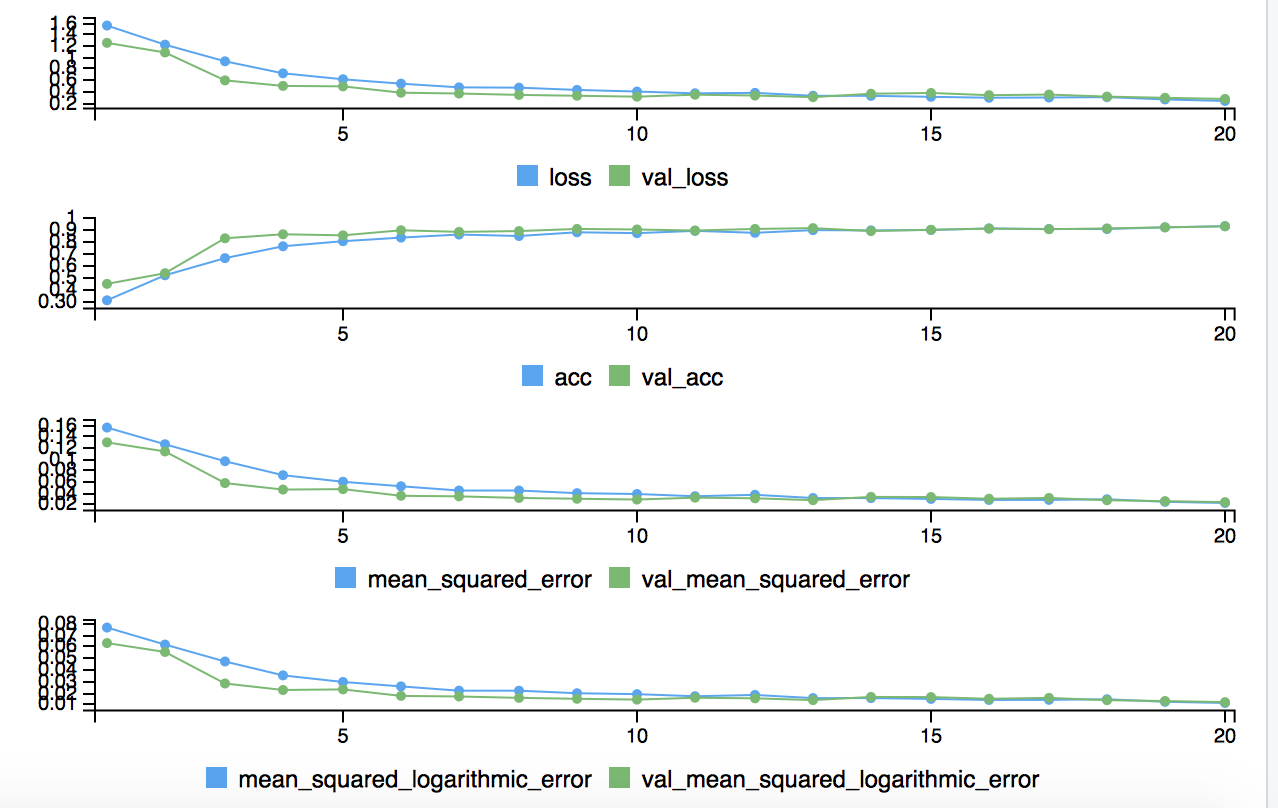
|  |  |  |  |
| --- | --- | --- | --- |
| **Iteration** | **Parameters** | **Metrics for train data** | **metrics for test data** |
| **1** | **layers = 4**  **optimizer = adam**  **activation = relu** | **acc: 0.9309**  **loss: 0.2195**  **mean\_squared\_error: 0.02179**  **mean\_squared\_logarithmic\_error: 0.01062** | **acc: 0.9303**  **loss: 0.2548**    **mean\_squared\_error: 0.02322**  **mean\_squared\_logarithmic\_error: 0.01134** |
| **2** | **layers = 5**  **optimizer = adam**  **activation = relu** | **acc: 0.9275**  **loss: 0.222**  **mean\_squared\_error: 0.02181**  **mean\_squared\_logarithmic\_error: 0.01062** | **acc: 0.8966**  **loss: 0.3228**  **mean\_squared\_error: 0.032292**  **mean\_squared\_logarithmic\_error: 0.01569** |
| **3** | **layers = 4**  **optimizer = rmsprop**  **activation = relu** | **loss: 0.255**  **acc: 0.9213**  **mean\_squared\_error: 0.02487**  **mean\_squared\_logarithmic\_error: 0.01209** | **acc: 0.9124**  **loss: 0.2723**  **mean\_squared\_error: 0.02542**  **mean\_squared\_logarithmic\_error: 0.01245** |
| **4** | **layers = 5**  **optimizer = rmsprop**  **activation = relu** | **acc: 0.9152**  **loss: 0.2845**  **mean\_squared\_error: 0.02686**  **mean\_squared\_logarithmic\_error: 0.0131** | **acc: 0.9034**  **loss: 0.3673**  **mean\_squared\_error: 0.03222**  **mean\_squared\_logarithmic\_error: 0.01564** |
| **5** | **layers = 5**  **Optimizer used = rmsprop**  **activation = tanh** | **acc: 0.9163**  **loss: 0.2668**  **mean\_squared\_error: 0.02626**  **mean\_squared\_logarithmic\_error: 0.01276** | **acc: 0.8854**  **loss: 0.376**  **mean\_squared\_error: 0.03479**  **mean\_squared\_logarithmic\_error: 0.01697** |

Hence by trying multiple times with different parameters we get the best accuracy in the first iteration with layers as 4, optimizer as adam and activation being relu.

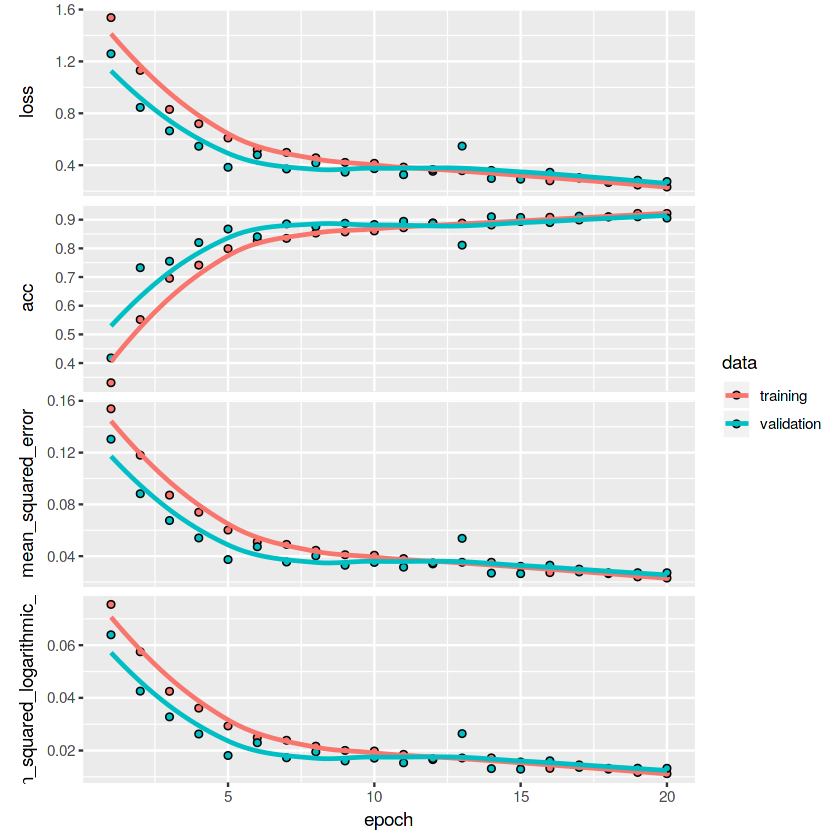
**Screenshot of Sample results:**

The metrics graphs are given as below : We have done the plots both on R studio and kaggle kernel environment.

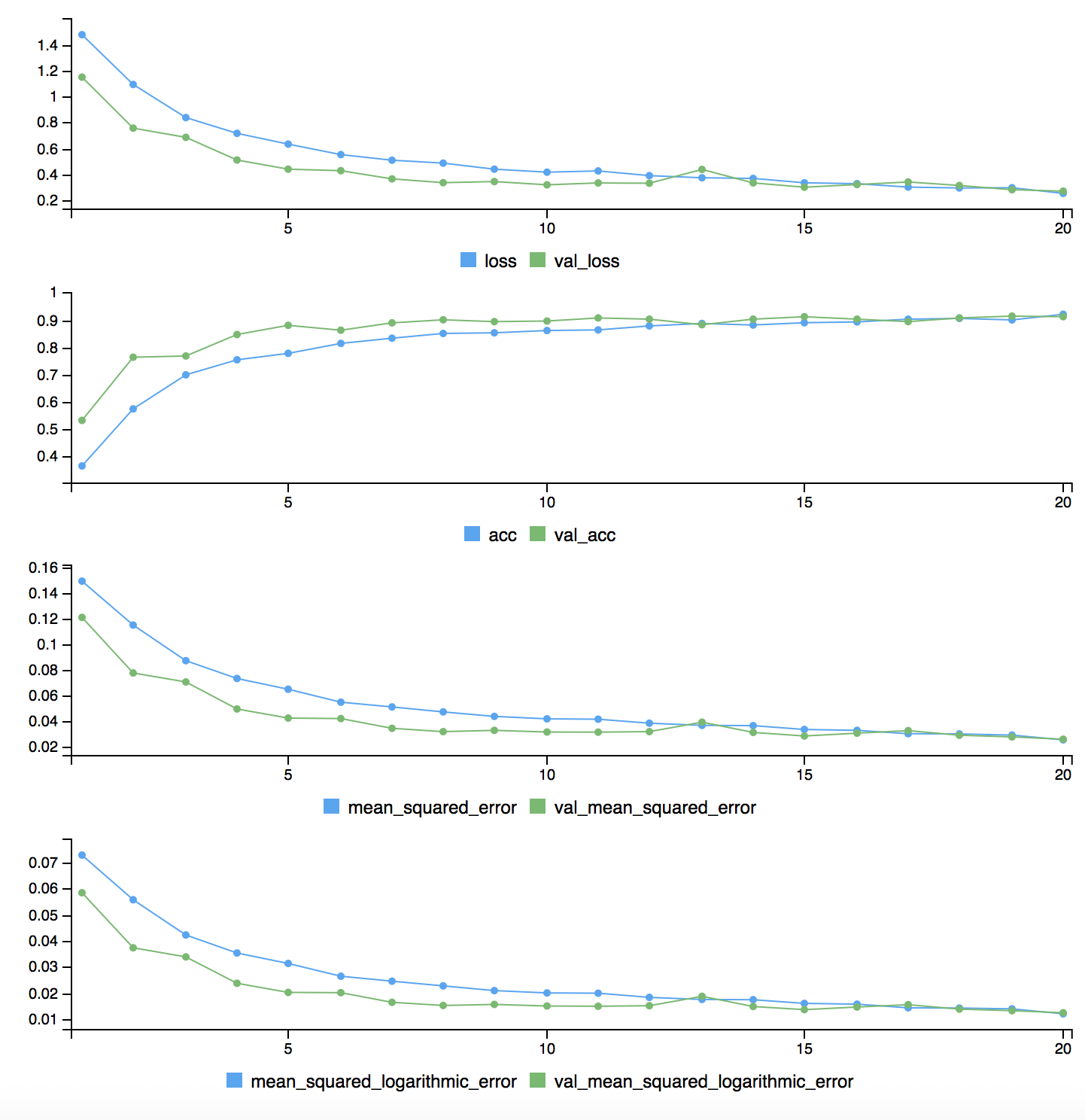
**Iteration -1 :**



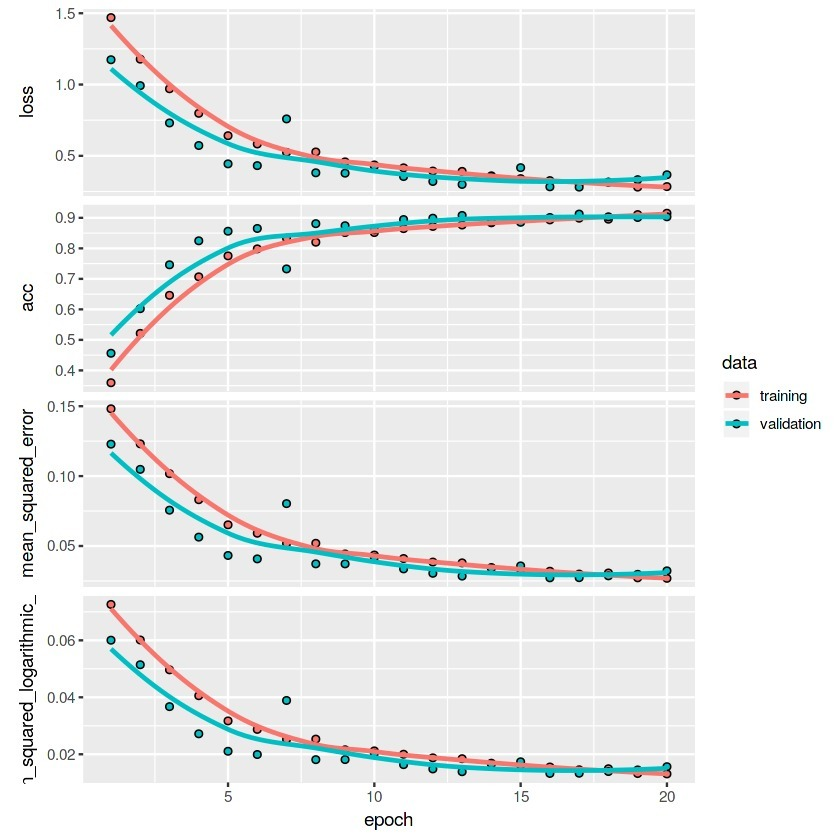
**Iteration -2 :**

****

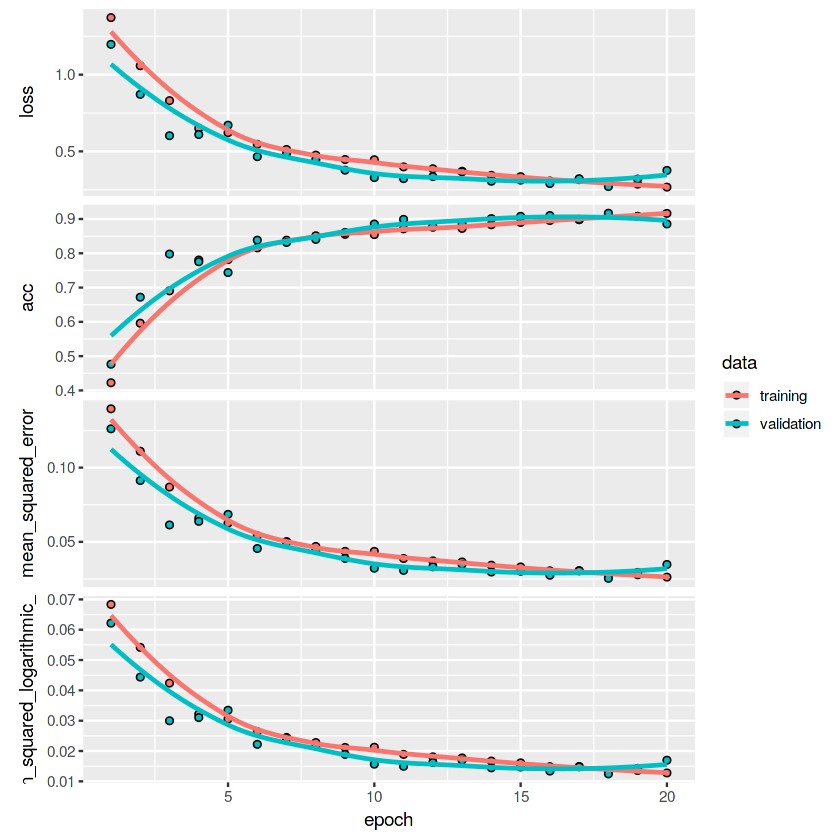
**Iteration -3 :**

****

**Iteration 4:**



**Iteration 5:**



Observations : achieved best results with adam optimizer with four layers network design and with activation function relu.

**Outputs:**

Since the paragraphs are too big, we are printing only the first sentence in the output data for results.

|  |  |  |
| --- | --- | --- |
| Text | Actual | Predicted |
| fox attacks blair s tory lies tony blair lied when he took the uk to war so has no qualms about lying in the election campaign say the tories. | politics | politics |
| us woman sues over cartridges a us woman is suing hewlett packard (hp) saying its printer ink cartridges are secretly programmed to expire on a certain date. | tech | tech |
| china had role in yukos split-up china lent russia $6bn (£3.2bn) to help the russian government renationalise the key yuganskneftegas unit of oil group yukos it has been revealed. | business | business |
| slovakia seal hopman cup success slovakia clinched the hopman cup for the second time by beating argentina 3-0 in saturday s final in perth. | sport | sport |
| analysis: no pain no gain he called it his masochism strategy in the run-up to the iraq war and now tony blair has signed up for another dose of pain. | politics | politics |
| show over for mtv s the osbournes rock star ozzy osbourne has said his family will not make any more episodes of reality tv show the osbournes. | entertainment | entertainment |
| bush to get tough on deficit us president george w bush has pledged to introduce a tough federal budget next february in a bid to halve the country s deficit in five years. | business | business |
| honour for uk games maker leading british computer games maker peter molyneux has been made an obe in the new year honours list. | tech | entertainment |
| the sound of music is coming home the original stage production of the sound of music is to be performed for the first time in the austrian capital 40 years after the film was released. | entertainment | entertainment |
| ban hits half-life 2 pirates hard about 20 000 people have been banned from playing the half-life 2 game. | tech | tech |
| crucial decision on super-casinos a decision on whether to allow westminster to legislate on super-casinos is set to be made by the scottish parliament. | politics | politics |
| quiksilver moves for rossignol shares of skis rossignol the world s largest ski-maker have jumped as much as 15% on speculation that it will be bought by us surfwear firm quiksilver. | business | business |
| us casino tricks face ban in uk controversial new uk casinos will be banned from using american tricks of the trade to ensure they are socially responsible it has been suggested. | business | politics |
| dawson joins england injury list scrum-half matt dawson is an injury doubt for england s six nations opener against wales next weekend. | sport | sport |
| fockers keeps us box office lead film comedy meet the fockers sequel to meet the parents has topped the us box office for a third week. | entertainment | entertainment |
| tobacco giants hail court ruling us tobacco companies have welcomed an appeal court s decision to reject the government s $280bn (£155bn) claim for alleged deceit about smoking dangers. | business | politics |
| vibe awards back despite violence the us vibe awards will be held again next year despite a stabbing which happened during the ceremony. | entertainment | politics |
| dance music not dead says fatboy dj norman cook - aka fatboy slim - has said that dance music is not dead but has admitted it is currently going through a fallow patch . | sport | entertainment |
| barbarians 19-47 new zealand new zealand proved too strong for an australian-dominated barbarians to round off their unbeaten northern hemisphere tour with an easy win. | sport | sport |
| england coach faces rap after row england coach andy robinson is facing disciplinary action after criticising referee jonathan kaplan in his side s six nations defeat to ireland | sport | sport |
| opposition grows to house arrests the conservatives have expressed serious misgivings about government plans for keeping uk and foreign terror suspects under house arrest. | politics | politics |
| scots suffer another injury blow scotland s back row crisis has worsened ahead of the rbs six nations with news that scott gray will miss out on the opening matches. | sport | sport |

**References**:

* 1. <https://www.kaggle.com/yesbutwhatdoesitmean/wikinews300d1mvec#wiki-news-300d-1M.vec>
  2. <https://www.r-bloggers.com/time-series-deep-learning-forecasting-sunspots-with-keras-stateful-lstm-in-r/>
  3. <https://www.kaggle.com/yufengdev/bbc-fulltext-and-category>
  4. <https://www.springboard.com/blog/machine-learning-with-r/>
  5. <https://www.r-bloggers.com/word-embeddings-with-keras/>