PROJECT 1

CS 6301.004

BBC TEXT CLASSIFICATION CHALLENGE

ARPITA DUTTA **AXD170025**

NIKITA VISPUTE **NXV170005**

REPORT

DATA SET:

Data Source: BBC news articles dataset has 2225 records and two fields.

1. Category: The articles are categorized in five categories.
    a. Tech
    b. Sport
    c. Business
    d. Entertainment
    e. Politics
2. Text: Text field is the title and body of the article concatenated.

The dataset looks like below:

response

Filter

	class	response_text
1	tech	tv future in the hands of viewers with home theatre syst...
2	business	worldcom boss left books alone former worldcom boss ...
3	sport	tigers wary of farrell gamble leicester say they will not b...
4	sport	yeading face newcastle in fa cup premiership side newc...
5	entertainment	ocean s twelve raids box office ocean s twelve the crime ...
6	politics	howard hits back at mongrel jibe michael howard has sa...
7	politics	blair prepares to name poll date tony blair is likely to na...
8	sport	henman hopes ended in dubai third seed tim henman sl...
9	sport	wilkinson fit to face edinburgh england captain jonny ...
10	entertainment	last star wars not for children the sixth and final star wa...
11	entertainment	berlin cheers for anti-nazi film a german movie about an...
12	business	virgin blue shares plummet 20% shares in australian bud...
13	business	crude oil prices back above $50 cold weather across part...
14	politics	hague given up his pm ambition former conservative lea...
15	sport	moya emotional after davis cup win carlos moya describ...
16	business	s korean credit card firm rescued south korea s largest cr...

PREDICTED LABEL AND TRUE LABEL

25 test samples have been displayed with their true and predicted labels in the output tab of the view.html page obtained from running the code on google cloud platform.

result

Filter

	true_labels	predicted_labels	text
1	business	business	fannie mae should restate books us mortgage company ...
2	business	business	yukos unit buyer faces loan claim the owners of embattl...
3	entertainment	entertainment	disputed nirvana box set on sale a box set featuring 68 ...
4	business	business	weak end-of-year sales hit next next has said its annual ...
5	politics	politics	blair and blunkett sheffield trip tony blair is to join hom...
6	politics	politics	baron kinnock makes lords debut former labour leader ...
7	tech	tech	warnings on woeful wi-fi security companies are getting...
8	tech	tech	halo fans hope for sequel xbox video game halo 2 has b...
9	entertainment	entertainment	johnny and denise lose passport johnny vaughan and d...
10	politics	politics	hitler row over welsh arts cash an artist critical of welsh ...
11	entertainment	entertainment	corbett attacks dumbed-down tv ronnie corbett has join...
12	business	business	qantas considers offshore option australian airline qant...
13	politics	politics	howard and blair tax pledge clash tony blair has said vo...
14	sport	sport	robertson out to retain euro lure hearts manager john r...
15	sport	sport	yelling takes cardiff hat-trick european cross-country ch...
16	business	business	cash gives way to flexible friend spending on credit and ...
17	tech	tech	go-ahead for new internet names the internet could so...
18	business	business	wipro beats forecasts once again wipro india s third-big...
19	tech	tech	tough rules for ringtone sellers firms that flout rules on ...
20	sport	sport	agassi fear for melbourne andre agassi s involvement in ...
21	business	business	worldcom ex-boss launches defence lawyers defending ...
22	business	business	us gives foreign firms extra time foreign firms have been...
23	politics	politics	school sport is back says pm tony blair has promised tha...
24	sport	sport	henman overcomes rival rusedski tim henman saved a m...
25	sport	sport	saint-andre anger at absent stars sale sharks director of ...

PARAMETER TESTING AND TUNING:

ITERATION	PARAMETERS	TRAINING AND TEST ACCURACY
1.	No of layers: 3 1. Embedding Layer: Input dimension= 10000 Output dimension= 32 Input length= 1000 2. Simple RNN Layer (32 units) 3. Dense Layer Output units= 512 4. Dense classifier Layer Output units= 5 Activation function: ReLu, sigmoid Optimizer: Adam Loss/ Error Function : Categorical Cross-entropy Epochs: 10 Batch-size: 32	Train accuracy: 45.05% Test accuracy: 50.36%
2.	No of layers: 3 1. Embedding Layer: Input dimension= 10000 Output dimension= 32 Input length= 1000 2. LSTM Layer (320 Layers) 3. Dense Layer Output units= 512 4. Dense classifier Layer Output units= 5 Activation function: ReLu, sigmoid Optimizer: Adam Loss/ Error Function : Categorical Cross-entropy Epochs: 10 Batch-size: 32	Train accuracy: 84.62% Test accuracy: 83.54%

3.	No of layers: 3 1. Embedding Layer: Input dimension= 10000 Output dimension= 32 Input length= 1000 2. Dense Layer Output units= 512 3. Dense classifier Layer Output units= 5 Activation function: ReLu, softmax Optimizer: Adam Loss/ Error Function : Categorical Cross-entropy Epochs: 10 Batch-size: 32	Train accuracy: 97.8% Test accuracy: 97.58%

The best model is the 3rd entry in the parameter and tuning table as it gives highest train and test accuracy.

This is displayed in the output view.html. of cloudml_2019_04_01_221717160 runs folder.

All the 3 models are displayed in the output view.html. cloudml_2019_04_02_041306029

runs folder.