WRNN: News text classification based on a Weighted RNN

PRAMAY SINGHVI – 2016B2A30759P

TARUN KUMAR – 2016A7PS0005P

Contents

Problem Statement

NEWS TEXT CLASSIFICATION BASED ON A WEIGHTED RNN

Overview – Text Classification

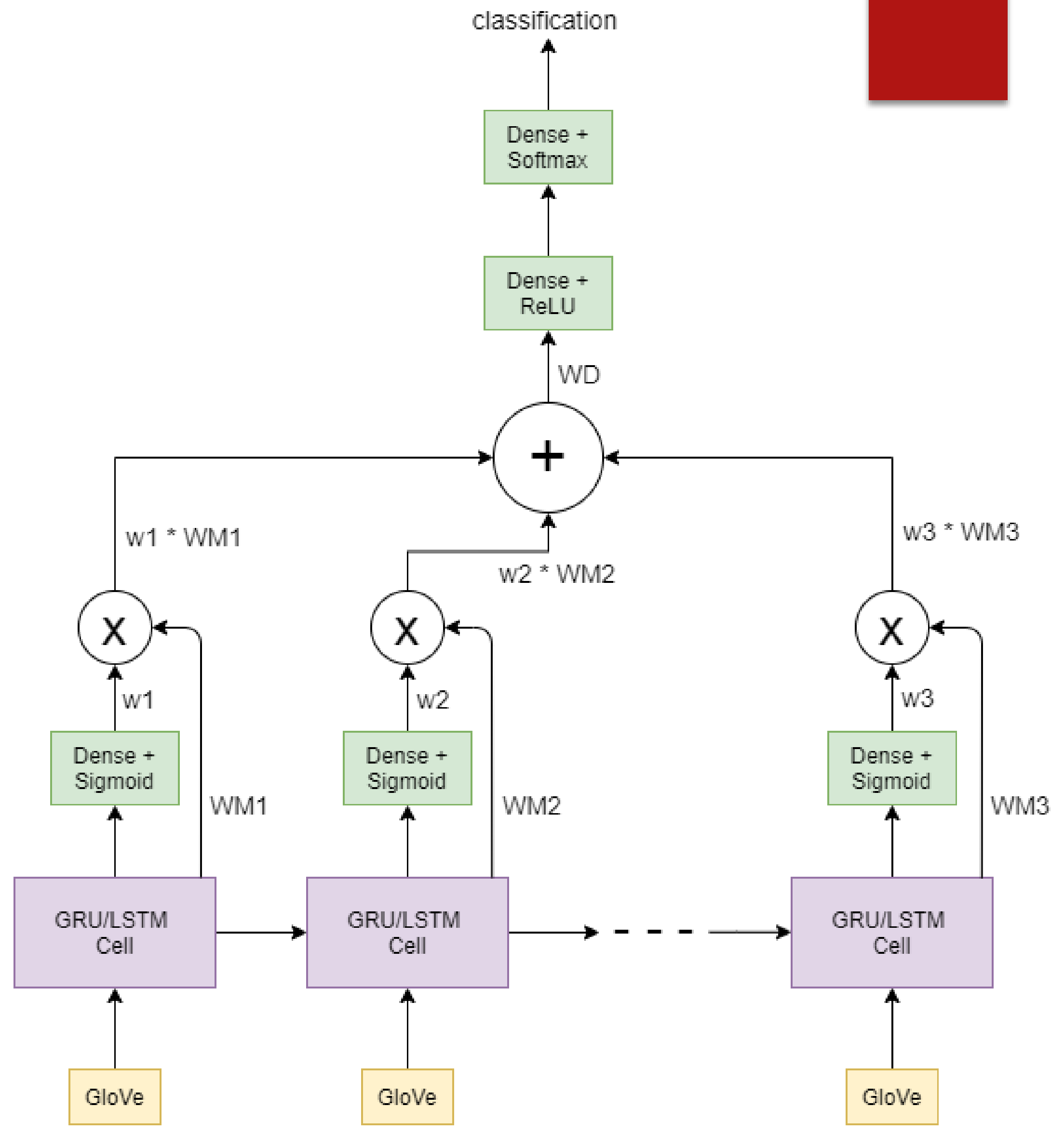
- Text classification forms the backbone for many advance tasks like Q&A systems, emotional analysis, analyzing posts on many blogging sites etc.
- A lot of research is happening, specifically in this domain.
- Some recent developments include shifting to GloVe and word2vec from traditional Bag of Words technique for embedding.
- Weighting the RNN output proves to be helpful for long documents establishing a new way of creating a document vector.

Dataset – 20newsgroups

- The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics split in two subsets: one for training (or development) and the other one for testing (or for performance evaluation). The split between the train and test set is based upon a messages posted before and after a specific date.
- We loaded this dataset using **sklearn.**
- The sequences were tokenized and one hot encoding was created for the labels before passing them to the model.
- Data was randomly split into 90:10 train-test partition (random seed is set in code)
- 5% train set was used for validation.

Architecture Details

- The task is to classify news articles into one of the 20 categories. Therefore, it is a sentence classification task.
- The data is first sent to an embedding layer which creates the word embeddings.
- Now these embeddings are sent to RNN cells to extract features of the serialized text data
- Further the resultant word vectors are weighted and summed and the passed to a fully connected layers for the classification purpose.
- Dropout is added to both after RNN and fully-connected layer 1.
- L2 regularization is employed for the fully-connected layer 1 weights.



Our Approach

- We used GloVe embeddings(300d) weights for our embedding layer. Although the paper mentions a word2vec 200d embedding.
- We also created a 200d word2vec embedding of our dataset only but we could see a clear overfitting and hence decided to proceed with GloVe .
- We also experimented with various architectures according to our understanding, apart from the architecture mentioned in the paper.
- For example - We passed the outputs of RNN layer to a Conv1D layer + Global average pooling and then the output to a Dense layer for classification.
- We got at par results with this architecture too(test accuracy 85%) but we decided not to deviate from the paper.
- We used Adam optimizer and added early stopping to optimize network weights.
- We also compared the WRNN model with baseline models like GRU and BiGRU.

Results and comparisons.

We report results on the test set for WRNN (with GRU unit), WRNN (with LSTM unit), GRU baseline and BiGRU baseline models.

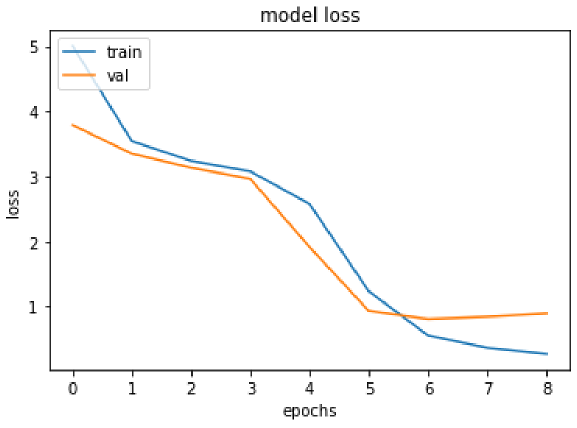
Model	Test Accuracy	Macro avg. precision	Macro avg. recall	Macro avg. F1	Avg. time/epoch	Convergence rate
WRNN (GRU unit)	85.73	86.0	85.0	85.0	40.02	7
WRNN (LSTM unit)	87.69	88.0	87.0	87.0	34.99	11
GRU baseline	81.37	82.0	80.0	81.0	41.31	3
BiGRU baseline	77.88	79.0	77.0	78.0	75.66	3

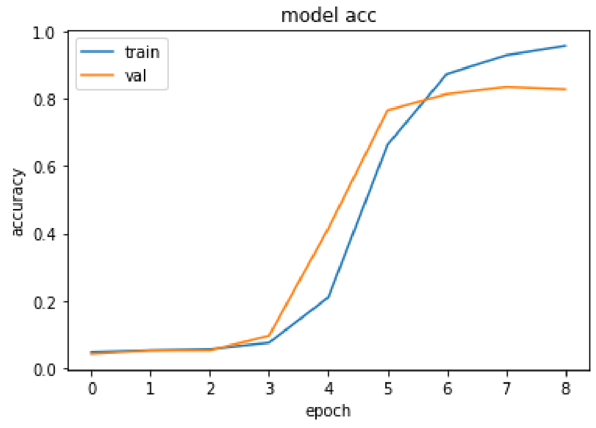
WRNN - GRU

	precision	recall	f1-score
0	0.83	0.86	0.84
1	0.69	0.86	0.77
2	0.88	0.80	0.84
3	0.71	0.78	0.75
4	0.77	0.88	0.82
5	0.87	0.84	0.86
6	0.80	0.84	0.82
7	0.92	0.84	0.88
8	0.92	0.94	0.93
9	0.93	0.91	0.92
10	0.97	0.96	0.97
11	0.91	0.91	0.91
12	0.84	0.80	0.82
13	0.85	0.93	0.88
14	0.86	0.82	0.84
15	0.86	0.93	0.89
16	0.87	0.77	0.81
17	0.96	0.94	0.95
18	0.85	0.73	0.79
19	0.92	0.67	0.77

Confusion Matrix

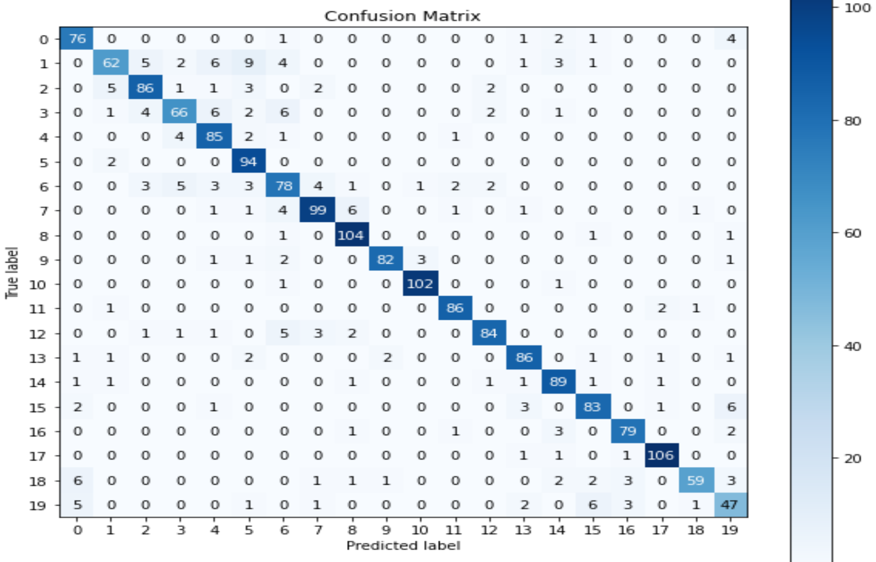
True label \ Predicted label	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	73	0	0	0	1	0	0	0	0	1	0	2	0	1	0	4	1	1	0	1
1	0	80	2	2	3	4	0	0	0	0	0	0	1	0	1	0	0	0	0	0
2	0	7	80	9	1	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	0	5	1	69	6	1	3	0	0	0	0	1	0	1	0	1	0	0	0	0
4	0	1	1	1	82	1	5	0	0	0	0	1	1	0	0	0	0	0	0	0
5	0	6	3	0	1	81	0	1	0	0	1	0	0	3	0	0	0	0	0	0
6	0	0	1	3	4	0	86	4	1	0	0	0	3	0	0	0	0	0	0	0
7	0	0	0	4	3	0	4	96	2	1	0	0	2	1	1	0	0	0	0	0
8	0	0	0	1	0	0	1	0	101	1	0	0	1	1	0	0	1	0	0	0
9	0	0	1	0	1	0	1	0	0	82	0	0	2	2	0	0	0	0	0	1
10	0	1	1	0	0	2	0	0	0	0	100	0	0	0	0	0	0	0	0	0
11	0	3	0	0	0	0	0	0	0	1	0	82	0	1	1	0	0	0	2	0
12	0	5	0	7	1	0	5	0	0	0	0	1	78	0	0	0	0	0	0	0
13	0	0	0	0	0	0	2	1	0	0	0	1	1	88	0	1	0	0	0	1
14	1	6	0	0	2	0	0	1	1	0	0	0	3	0	79	1	0	0	2	0
15	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	89	0	2	1	0
16	1	0	0	0	0	0	0	1	3	0	0	0	1	1	7	0	66	1	5	0
17	2	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	103	0	0
18	4	0	1	0	0	0	0	0	1	2	1	1	0	3	2	0	5	0	57	1
19	6	0	0	0	0	1	0	0	1	0	1	1	0	2	0	7	3	0	0	44

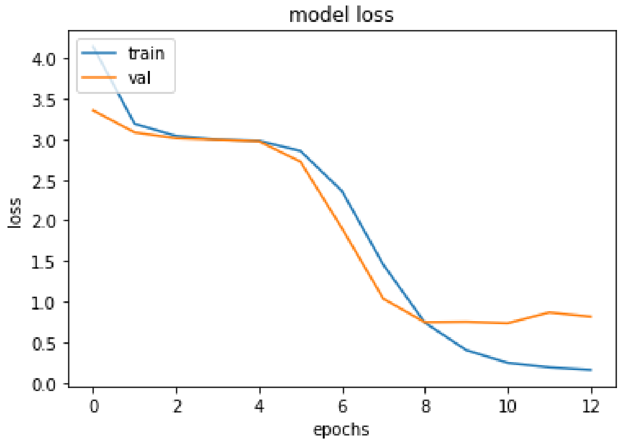


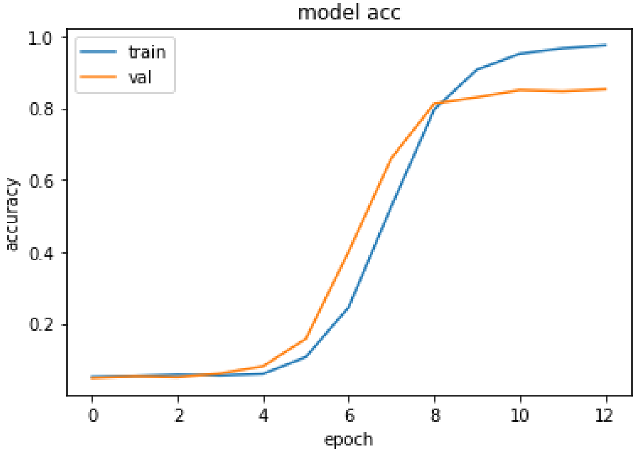


WRNN - LSTM

	precision	recall	f1-score
0	0.84	0.89	0.86
1	0.85	0.67	0.75
2	0.87	0.86	0.86
3	0.84	0.75	0.79
4	0.81	0.91	0.86
5	0.80	0.98	0.88
6	0.76	0.76	0.76
7	0.90	0.87	0.88
8	0.90	0.97	0.93
9	0.96	0.91	0.94
10	0.96	0.98	0.97
11	0.95	0.96	0.95
12	0.92	0.87	0.89
13	0.90	0.91	0.90
14	0.87	0.93	0.90
15	0.86	0.86	0.86
16	0.92	0.92	0.92
17	0.95	0.97	0.96
18	0.95	0.76	0.84
19	0.72	0.71	0.72

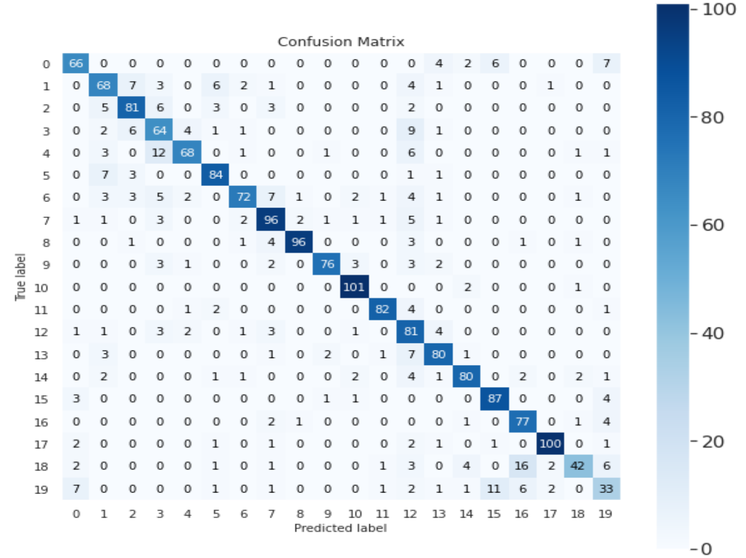


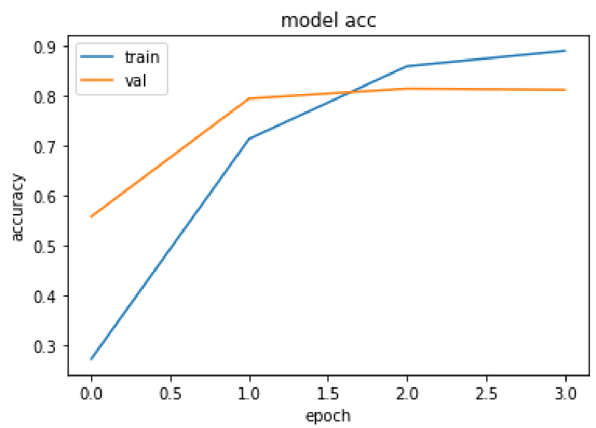


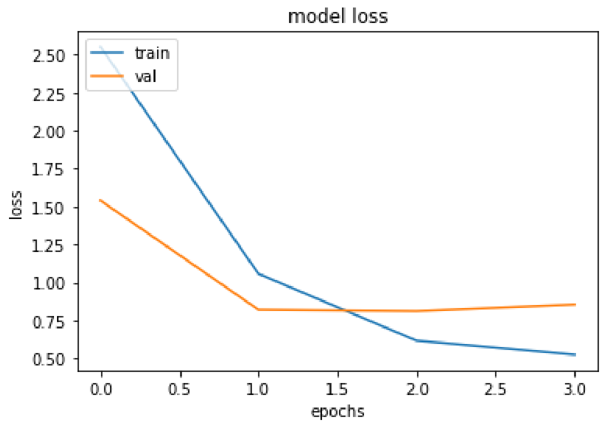


Simple GRU

	precision	recall	f1-score
0	0.80	0.78	0.79
1	0.72	0.73	0.72
2	0.80	0.81	0.81
3	0.65	0.73	0.68
4	0.87	0.73	0.80
5	0.84	0.88	0.86
6	0.89	0.71	0.79
7	0.79	0.84	0.81
8	0.96	0.90	0.93
9	0.94	0.84	0.89
10	0.91	0.97	0.94
11	0.94	0.91	0.93
12	0.58	0.84	0.68
13	0.82	0.84	0.83
14	0.88	0.83	0.86
15	0.83	0.91	0.87
16	0.75	0.90	0.82
17	0.95	0.92	0.93
18	0.86	0.54	0.66
19	0.57	0.50	0.53

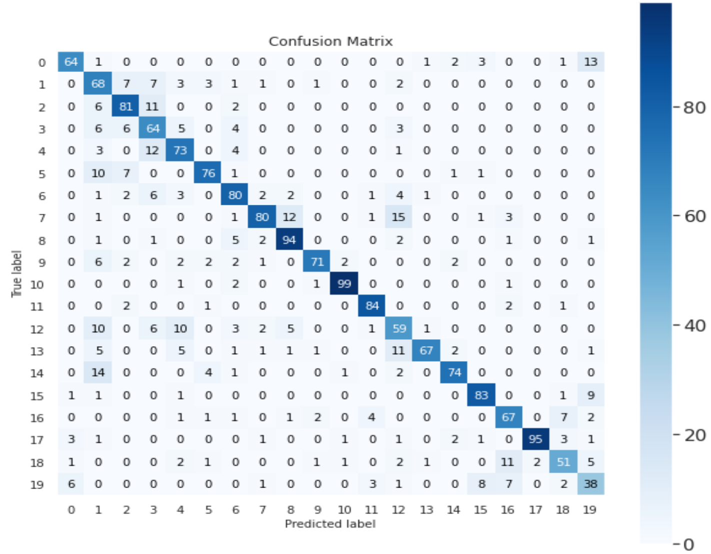


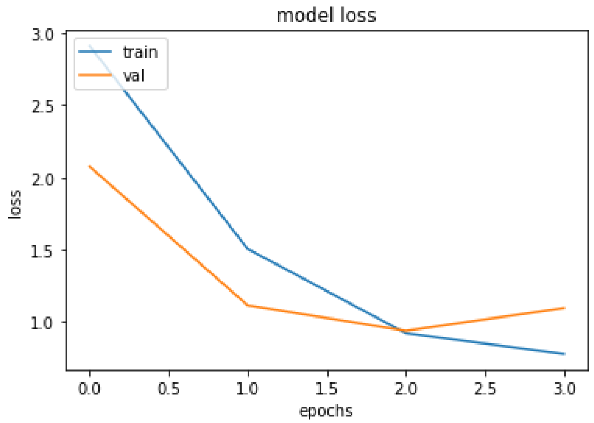


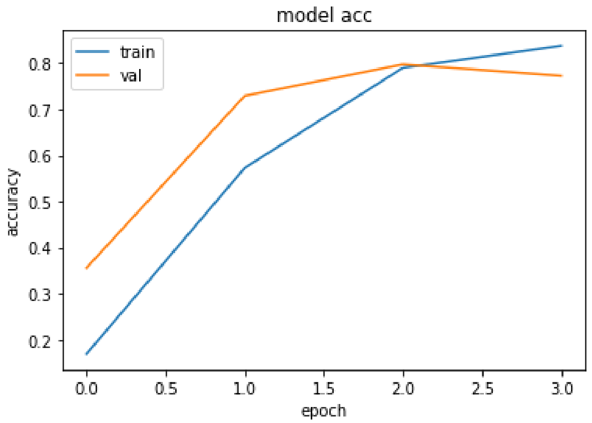


Bi GRU

	precision	recall	f1-score
0	0.85	0.75	0.80
1	0.51	0.73	0.60
2	0.76	0.81	0.78
3	0.60	0.73	0.66
4	0.69	0.78	0.73
5	0.86	0.79	0.83
6	0.74	0.78	0.76
7	0.88	0.70	0.78
8	0.82	0.88	0.85
9	0.92	0.79	0.85
10	0.95	0.95	0.95
11	0.89	0.93	0.91
12	0.57	0.61	0.59
13	0.94	0.71	0.81
14	0.89	0.77	0.83
15	0.86	0.86	0.86
16	0.73	0.78	0.75
17	0.98	0.87	0.92
18	0.77	0.65	0.71
19	0.54	0.58	0.56







Predictions on Test set

	True Label	Predicted Label	True/False
0	rec.sport.baseball	rec.sport.baseball	TRUE
1	talk.politics.mideast	talk.politics.mideast	TRUE
2	talk.politics.mideast	talk.politics.mideast	TRUE
3	talk.politics.misc	talk.politics.guns	FALSE
4	talk.politics.mideast	talk.politics.mideast	TRUE
5	rec.autos	rec.autos	TRUE
6	talk.religion.misc	sci.med	FALSE
7	comp.sys.mac.hardware	comp.sys.mac.hardware	TRUE
8	comp.windows.x	comp.windows.x	TRUE
9	comp.sys.mac.hardware	comp.sys.mac.hardware	TRUE
10	comp.graphics	comp.graphics	TRUE
11	comp.windows.x	comp.windows.x	TRUE
12	sci.med	sci.med	TRUE
13	comp.sys.mac.hardware	comp.sys.mac.hardware	TRUE
14	comp.windows.x	comp.windows.x	TRUE
15	talk.politics.mideast	talk.politics.mideast	TRUE
16	sci.crypt	comp.graphics	FALSE
17	rec.autos	rec.autos	TRUE
18	talk.politics.misc	talk.politics.misc	TRUE
19	alt.atheism	alt.atheism	TRUE
20	sci.crypt	sci.crypt	TRUE

Sample Prediction

- From: rickert@NeXTwork.Rose-Hulman.Edu (John H. Rickert) Subject: mile high runs Article-I.D.: master.1psq90INNh93 Reply-To: rickert@NeXTwork.Rose-Hulman.Edu (John H. Rickert) Organization: Computer Science Department at Rose-Hulman Lines: 35 NNTP-Posting-Host: g215a-1.nextwork.rose-hulman.edu How many runs will be scored in Denver? I don't know. but some idea can be gotten by looking at the runs scored in Mile High Stadium during the last few years of the Bears/Zephyrs tenure in the American Association. Here's the total runs scored per game in Zephyrs games, all league games and the ratio. I found the same ratios for HR. Year rpg lea ratio hrpg lea ratio 1992 10.22 9.10 1.12 1.65 1.58 1.04 1991 9.53 8.87 1.07 1.41 1.26 1.12 1990 10.71 8.72 1.23 1.49 1.24 1.20 1989 9.07 8.34 1.09 1.27 1.11 1.14 1988 9.90 8.37 1.18 1.29 1.08 1.19 1987 12.55 10.70 1.17 2.39 1.92 1.24 1986 9.45 9.33 1.01 1.35 1.38 .98 1985 9.50 8.54 1.11 1.53 1.34 1.14 1984 9.99 9.10 1.10 1.55 1.59 .97 1983 10.60 9.99 1.06 2.03 1.74 1.17 1982 11.29 10.35 1.09 2.24 1.91 1.17 1981 10.29 9.25 1.11 1.43 1.49 .96 1980 10.59 9.43 1.12 1.63 1.46 1.12 1446/13-->1.11 1444/13-->1.11 It seems pretty clear that Denver will have a large effect on runs scored (I'll stick with my prediction from last year that it'll be one of the top 3 in the NL this year) and a fairly large effect on Homeruns - though apparently not as large as Atlanta, Wrigley, Cincinnati and San Diego. Still it ought to be a pretty decent home run park. john rickert rickert@nextwork.rose-hulman.edu

Predicted - **rec.sport.baseball**

True Labels - rec.sport.baseball

Conclusion

- We get a somewhat better results as compared to the original paper. This might be due to two reasons –
  - We have used GloVe embeddings.
  - The results reported in paper are clearly over fitting the training dataset whereas our results have employed proper measures to tackle this and hence achieved better test accuracy.
- WRNN model with LSTM cells outperforms all other models.
- GRU baseline is better than Bidirectional GRU baseline and WRNN-LSTM performs better than WRNN-GRU.
- It seems that dynamically weighing the outputs of RNN to form a combined and contextual representation helps in boosting the performance on the task as opposed to just taking the output at the last time-step and using that for classification.