ARXIV TOPIC CLASSIFICATION AND PAPER RECOMMENDATION SYSTEM

Team Members

Nirajan Bekoju

Nirajan Thakuri

Rijan Pokhrel

Abhisek Timilsina

INTRODUCTION

- Multi-label classification of Arxiv Research Paper
- Development of recommendation system for the paper

PROBLEM

2,309,544
Total number of articles
August 9, 2023

- **Difficult to get the recommendation of similar paper especially for beginners.**

DATASETS

	title	authors	date	abstract	categories
0	Calculation of prompt diphoton production cros...	C. Bal\'azs, E. L. Berger, P. M. Nadolsky, C.-...	2008-11-26	A fully differential calculation in perturba...	hep-ph
1	Sparsity-certifying Graph Decompositions	Ileana Streinu and Louis Theran	2008-12-13	We describe a new algorithm, the (k,ℓ)-...	math.CO cs.CG
2	The evolution of the Earth-Moon system based o...	Hongjun Pan	2008-01-13	The evolution of Earth-Moon system is descri...	physics.gen-ph
3	A determinant of Stirling cycle numbers counts...	David Callan	2007-05-23	We show that a determinant of Stirling cycle...	math.CO
4	From dyadic Λ_{α} to $\Lambda_{\a...	Wael Abu-Shammala and Alberto Torchinsky	2013-10-15	In this paper we show how to compute the $\L...	math.CA math.FA

DATASETS

156
UNIQUE LABELS

1 000 000
Data Points

METHODOLOGY

PREPROCESSING

- Remove all punctuations
- Replace number with <NUM tokens>
- Lower case
- Remove all stop words
- Stemming

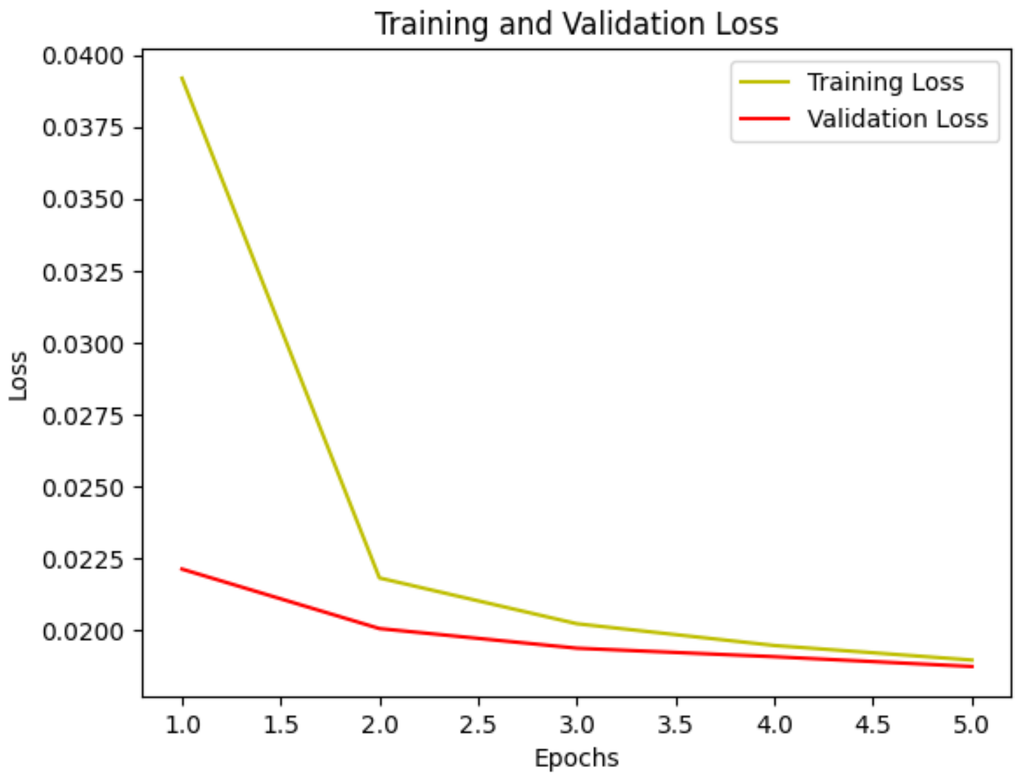
TOKENIZATION

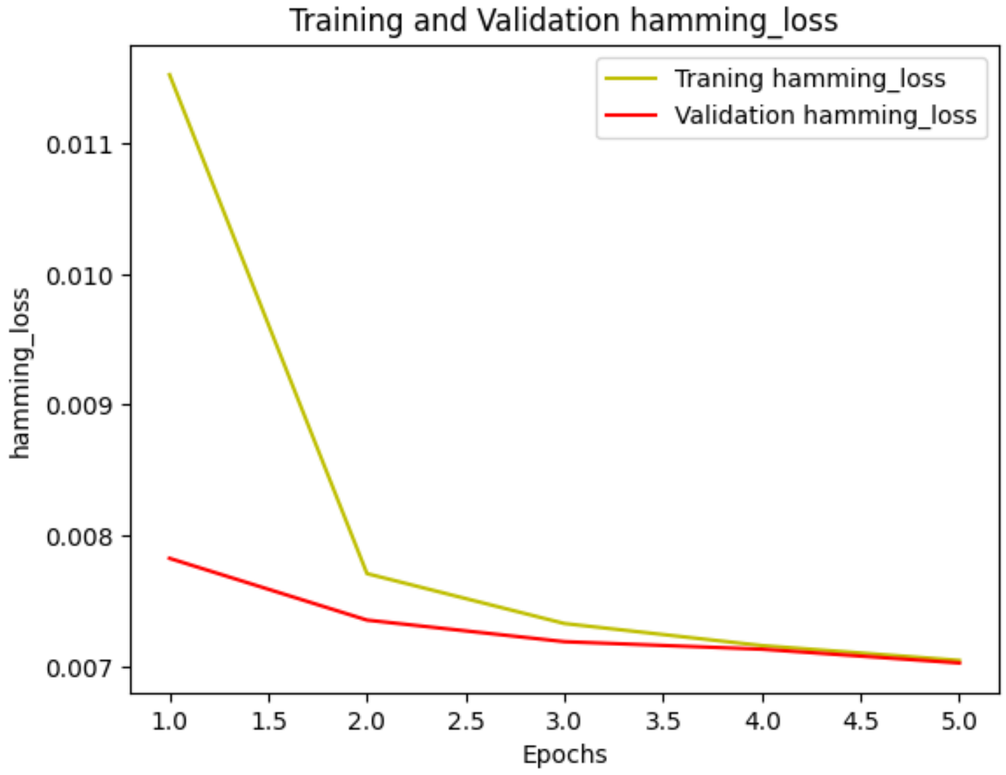
- UNIQUE WORD COUNT = 25 000
- MAX_PADDING_LENGTH = 210

MODEL DEVELOPMENT

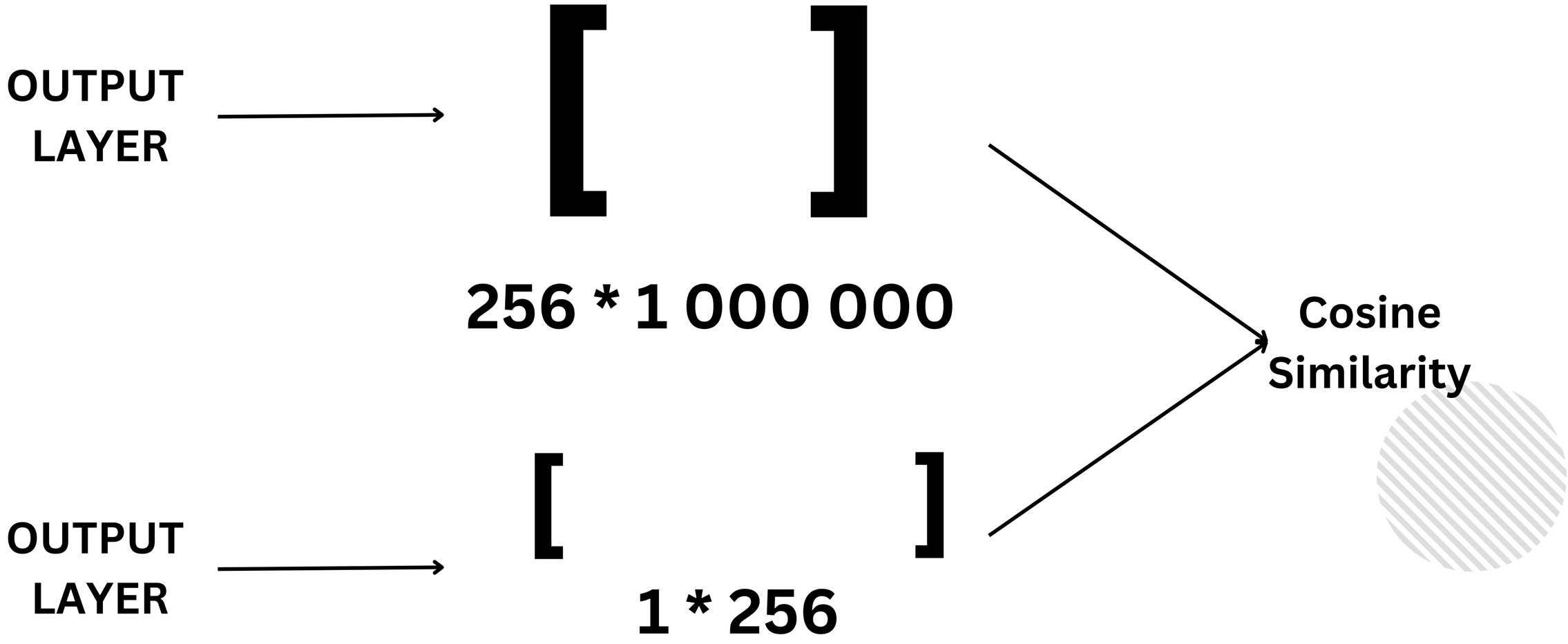
```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 210)]             0

 token_and_position_embeddin  (None, 210, 32)          806720
 g (TokenAndPositionEmbeddin
 g)

 transformer_block (Transfor  (None, 210, 32)          10656
 merBlock)

 global_average_pooling1d (G  (None, 32)               0
 lobalAveragePooling1D)

 dropout_2 (Dropout)         (None, 32)                0

 dense_2 (Dense)             (None, 256)               8448

 dropout_3 (Dropout)         (None, 256)               0

 dense_3 (Dense)             (None, 156)               40092

=================================================================
Total params: 865,916
Trainable params: 865,916
Non-trainable params: 0
_________________________________________________________________
```

RESULT


Training and Validation Loss
Training Loss
Validation Loss
Loss
0.0400
0.0375
0.0350
0.0325
0.0300
0.0275
0.0250
0.0225
0.0200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Epochs


Training and Validation hamming_loss
Traning hamming_loss
Validation hamming_loss
hamming_loss
0.011
0.010
0.009
0.008
0.007
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Epochs

RECOMMENDATION



THANK YOU