# NLP Final Project Report

Chitrang Goyani | cbg5586@psu.edu

## Task Description and Abstract

The task proposed in this project is a text generation task, specifically the generation of job descriptions. This task is an open-ended task and there are various methods of achieving it, this particular project proposed auto-generation of job descriptions using seed text. This is carried out using a Long-Short Term Memory Networks as shown in Figure 1. LSTM's are combinations of RNN's with gates that are capable of learning long-term dependencies specifically in sequence generation tasks such as the one discussed here.

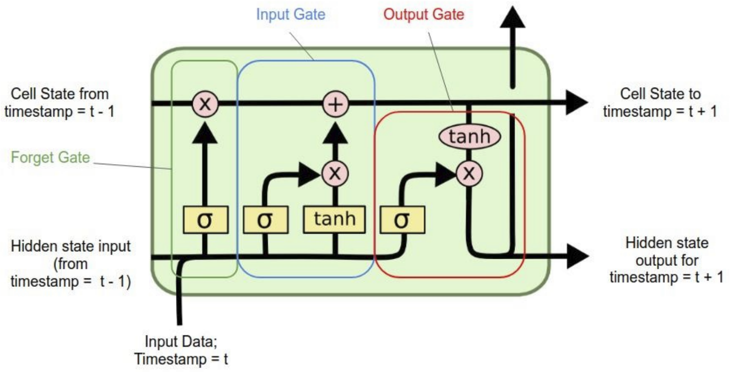
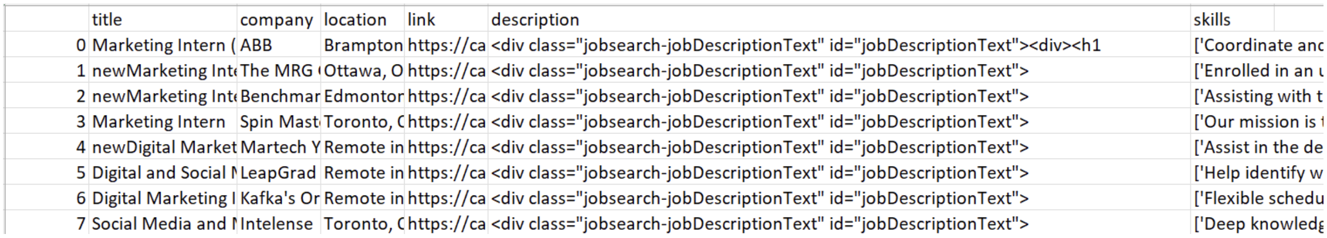

Input Gate
Output Gate
Cell State from timestamp = t - 1
Cell State to timestamp = t + 1
Forget Gate
tanh
σ
σ
tanh
σ
Hidden state input (from timestamp = t - 1)
Hidden state output for timestamp = t + 1
Input Data; Timestamp = t

Figure 1: LSTM unit

## Dataset

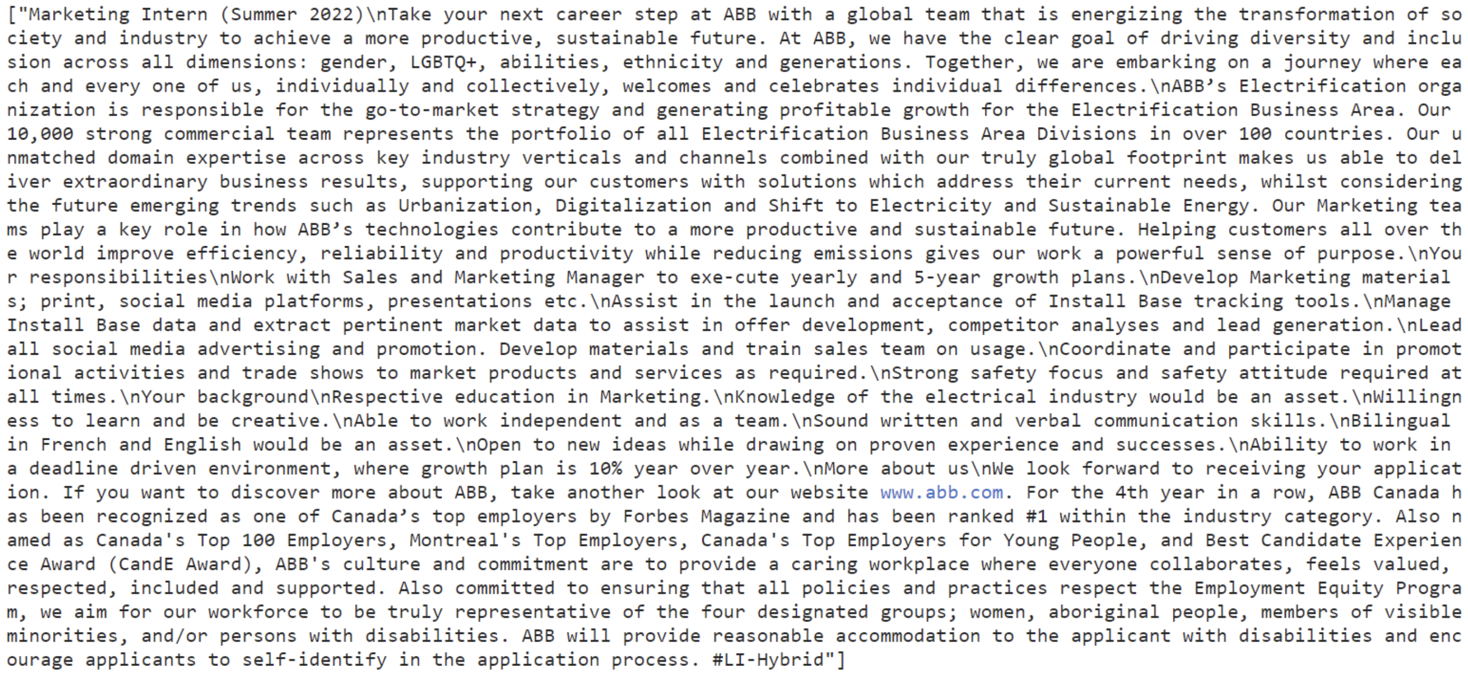
The Dataset used for this project is a non-standard dataset scraped from the web and improved upon using various filters. It is scraped for LinkedIn and Indeed.com, specifically for US and Canada based jobs. Since it is a scraped dataset, initially the dataset consists of a lot of HTML tags and hence needs to be filtered. Figure 2 shows the dataset with its columns and HTML tags specifically in the 'descriptions' column. This dataset is specific to the role of a 'Marketing Intern' and therefore all the tests done are with relation to generating job descriptions for a Marketing Intern Position.



| | title | company | location | link | description | skills |
|---|---|---|---|---|---|---|
| 0 | Marketing Intern ( | ABB | Brampton | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"><div><h1 | ['Coordinate anc |
| 1 | newMarketing Int | The MRG ( | Ottawa, O | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"> | ['Enrolled in an u |
| 2 | newMarketing Int | Benchmar | Edmonton | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"> | ['Assisting with t |
| 3 | Marketing Intern | Spin Mast | Toronto, C | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"> | ['Our mission is t |
| 4 | newDigital Market | Martech Y | Remote in | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"> | ['Assist in the de |
| 5 | Digital and Social M | LeapGrad | Remote in | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"> | ['Help identify w |
| 6 | Digital Marketing I | Kafka's Or | Remote in | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"> | ['Flexible schedu |
| 7 | Social Media and M | Intelense | Toronto, C | https://ca | <div class="jobsearch-jobDescriptionText" id="jobDescriptionText"> | ['Deep knowledg |

Figure 2: Un-filtered Dataset

For this project, we load **100 rows** from this dataset with a maximum sequence length of **986 tokens.** This dataset is used for training and there are no validation or test sets involved. Since the output is an unseen (or atleast a different sequence) of tokens it does not make sense to use a validation set or a test set. Figure 3 shows filtered description after parsing and removing the HTML. Additionally the data is also filtered to remove '\n' on tokenization.



```
["Marketing Intern (Summer 2022)\nTake your next career step at ABB with a global team that is energizing the transformation of so
ciety and industry to achieve a more productive, sustainable future. At ABB, we have the clear goal of driving diversity and inclu
sion across all dimensions: gender, LGBTQ+, abilities, ethnicity and generations. Together, we are embarking on a journey where ea
ch and every one of us, individually and collectively, welcomes and celebrates individual differences.\nABB’s Electrification orga
nization is responsible for the go-to-market strategy and generating profitable growth for the Electrification Business Area. Our
10,000 strong commercial team represents the portfolio of all Electrification Business Area Divisions in over 100 countries. Our u
nmatched domain expertise across key industry verticals and channels combined with our truly global footprint makes us able to del
iver extraordinary business results, supporting our customers with solutions which address their current needs, whilst considering
the future emerging trends such as Urbanization, Digitalization and Shift to Electricity and Sustainable Energy. Our Marketing tea
ms play a key role in how ABB’s technologies contribute to a more productive and sustainable future. Helping customers all over th
e world improve efficiency, reliability and productivity while reducing emissions gives our work a powerful sense of purpose.\nYou
r responsibilities\nWork with Sales and Marketing Manager to exe-cute yearly and 5-year growth plans.\nDevelop Marketing material
s; print, social media platforms, presentations etc.\nAssist in the launch and acceptance of Install Base tracking tools.\nManage
Install Base data and extract pertinent market data to assist in offer development, competitor analyses and lead generation.\nLead
all social media advertising and promotion. Develop materials and train sales team on usage.\nCoordinate and participate in promot
ional activities and trade shows to market products and services as required.\nStrong safety focus and safety attitude required at
all times.\nYour background\nRespective education in Marketing.\nKnowledge of the electrical industry would be an asset.\nWillingn
ess to learn and be creative.\nAble to work independent and as a team.\nSound written and verbal communication skills.\nBilingual
in French and English would be an asset.\nOpen to new ideas while drawing on proven experience and successes.\nAbility to work in
a deadline driven environment, where growth plan is 10% year over year.\nMore about us\nWe look forward to receiving your applicat
ion. If you want to discover more about ABB, take another look at our website www.abb.com. For the 4th year in a row, ABB Canada h
as been recognized as one of Canada’s top employers by Forbes Magazine and has been ranked #1 within the industry category. Also n
amed as Canada's Top 100 Employers, Montreal's Top Employers, Canada's Top Employers for Young People, and Best Candidate Experien
ce Award (CandE Award), ABB's culture and commitment are to provide a caring workplace where everyone collaborates, feels valued,
respected, included and supported. Also committed to ensuring that all policies and practices respect the Employment Equity Progra
m, we aim for our workforce to be truly representative of the four designated groups; women, aboriginal people, members of visible
minorities, and/or persons with disabilities. ABB will provide reasonable accommodation to the applicant with disabilities and enc
ourage applicants to self-identify in the application process. #LI-Hybrid"]
```

Figure 3: Filtered Data

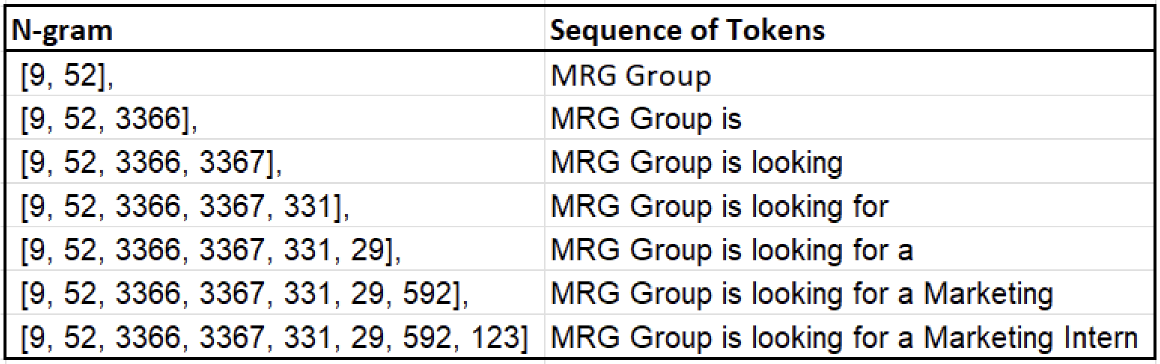
## Implementation

The tools used for the implementation of this project mainly consist of utilizing Kaggle’s P100 GPU’s with the model implemented using Tenserflow’s Keras. The dataset was loaded into a pandas dataframe and the tokenizer used was loaded from Keras as well. The evaluation was done using ROUGE (Recall Oriented Understudy for Gisting Evaluation) score. Pre-trained word-2-vec word embeddings were used from wikinews 300d subwords.

LSTMs can both predict the next word in a sequence and generate it. During training, an LSTM model is trained to predict the next word in a sequence based on the input sequence and the previous words in the sequence. During inference, or when generating new sequences, the LSTM model can be used to generate the next word in the sequence by sampling from the predicted probability distribution over the vocabulary.

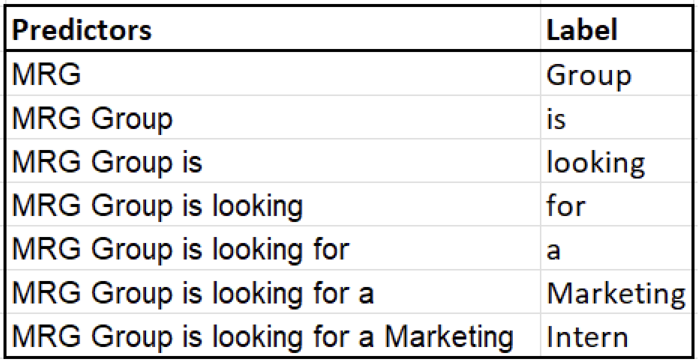
To generate a sequence, the LSTM model is typically given a starting sequence (also called a prompt or seed), and then iteratively generates the next word in the sequence based on the previous words and the predicted probability distribution over the vocabulary. The process of generating the next word is repeated until a stopping criterion is met (e.g., a maximum sequence length is reached, or a special end-of-sequence token is generated)

Below is a figure 4 that lists integers representing the ngram phrases generated from the corpus.
Figure 5 shows Pad sequences transform lists of integers into a 2D Numpy array of shape (num_samples, maxlen).



| N-gram | Sequence of Tokens |
|---|---|
| [9, 52], | MRG Group |
| [9, 52, 3366], | MRG Group is |
| [9, 52, 3366, 3367], | MRG Group is looking |
| [9, 52, 3366, 3367, 331], | MRG Group is looking for |
| [9, 52, 3366, 3367, 331, 29], | MRG Group is looking for a |
| [9, 52, 3366, 3367, 331, 29, 592], | MRG Group is looking for a Marketing |
| [9, 52, 3366, 3367, 331, 29, 592, 123] | MRG Group is looking for a Marketing Intern |

Figure 4: Ngram sequences



| Predictors | Label |
|---|---|
| MRG | Group |
| MRG Group | is |
| MRG Group is | looking |
| MRG Group is looking | for |
| MRG Group is looking for | a |
| MRG Group is looking for a | Marketing |
| MRG Group is looking for a Marketing | Intern |

Figure 5: Predictors and labels

Figure 6 represents the workflow of the implementation of this project.

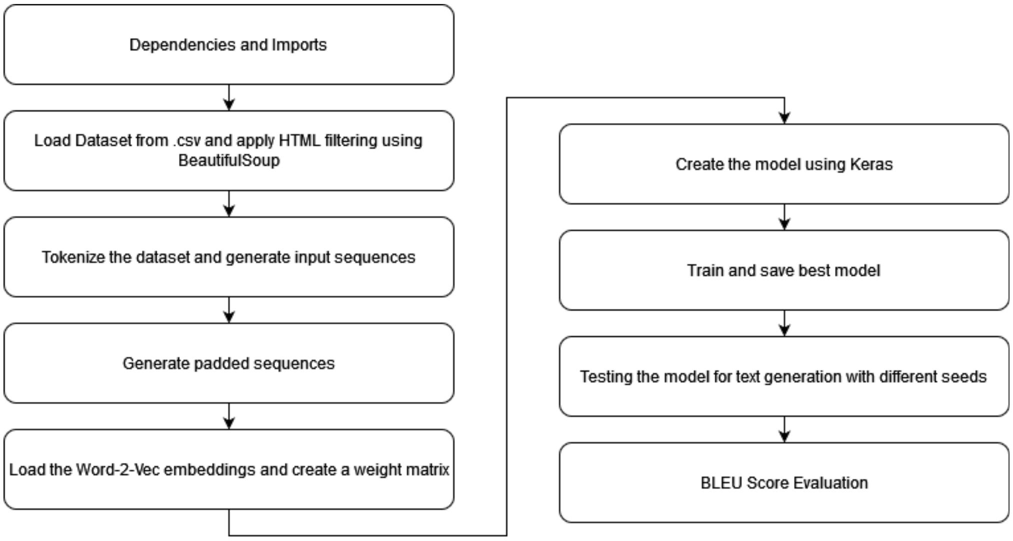
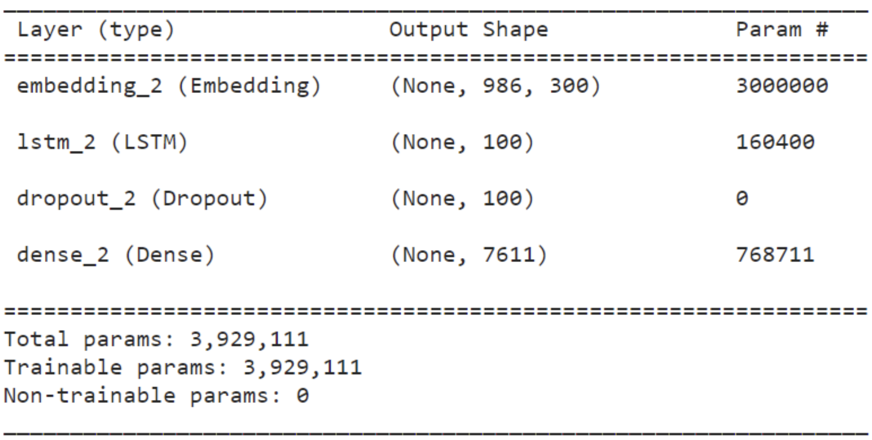

Dependencies and Imports
Load Dataset from .csv and apply HTML filtering using BeautifulSoup
Tokenize the dataset and generate input sequences
Generate padded sequences
Load the Word-2-Vec embeddings and create a weight matrix
Create the model using Keras
Train and save best model
Testing the model for text generation with different seeds
BLEU Score Evaluation

Figure 6: Implementation Workflow

Figure 7 shows the model summary. It consists of **3.9M** parameters with **100 LSTM units** and a **dropout rate of 20%**. It consists of an embedding layer as the input layer and a fully connected dense layer as the output layer with **softmax** as the activation function. Here (986,300) represents embeddings of 986 tokens (max sequence length) with 300 features per token.



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_2 (Embedding)     (None, 986, 300)          3000000

 lstm_2 (LSTM)               (None, 100)               160400

 dropout_2 (Dropout)         (None, 100)               0

 dense_2 (Dense)             (None, 7611)              768711

=================================================================
Total params: 3,929,111
Trainable params: 3,929,111
Non-trainable params: 0
_________________________________________________________________
```

Figure 7: Model Summary

## Evaluation and Results

In sequence generation tasks, it can be challenging to determine when to stop training because there is no fixed target or ground truth for the generated sequences. Without a validation set, it is not possible to use metrics such as validation loss or accuracy to monitor the performance of the model on unseen data and decide when to stop training.

One common approach to deciding when to stop training a sequence generation model is to use a fixed number of training iterations or epochs. This approach is based on the assumption that the model will continue to learn from the training data up to a certain point, after which the performance on the training data will plateau or start to deteriorate due to overfitting. In practice, the number of training iterations or epochs is often determined based on a combination of heuristics, computational resources, and previous experience with similar tasks. In this project we train the model for 100 epochs.

The evaluation metric for this model is the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score which is a set of evaluation metrics commonly used in natural language processing to measure the similarity between generated text and a set of reference texts. It is often used to evaluate the performance of text summarization and machine translation systems. In this project, we get the result of rouge scores as **0.04417.** It is important to note here that each row description in the dataset might differ in terms of company name, work location, roles and responsibilities etc. and therefore the ROUGE might not be the best metric for such a task but it is used to evaluate the performance of SOTA models is therefore incorporated in this project as well.

Figure 8 shows the training loss as the number of epochs increases. Loss at epoch 100 is **0.1458.**

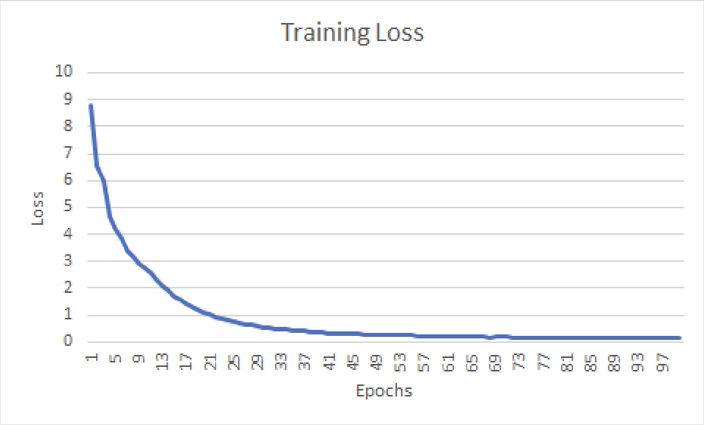
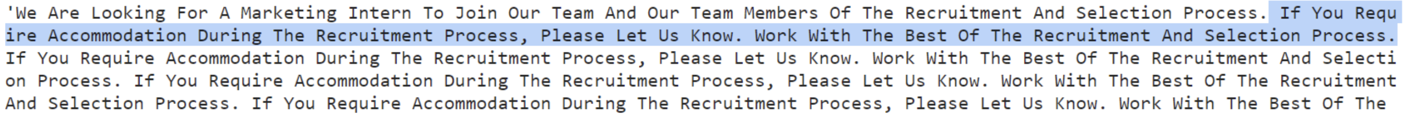
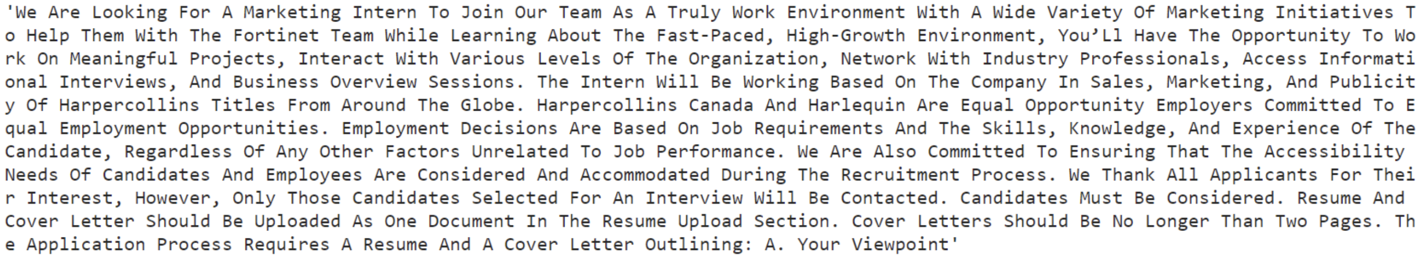


Figure 8(a-d) show the results of descriptions generated from the model trained at different epochs.



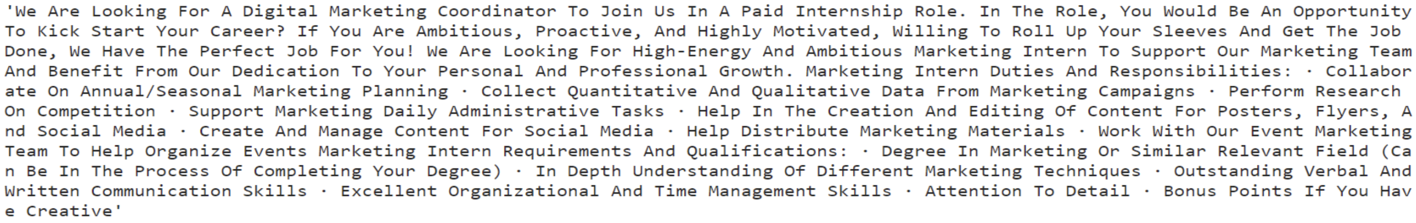
```
'We Are Looking For A Marketing Intern To Join Our Team And Our Team Members Of The Recruitment And Selection Process. If You Requ
ire Accommodation During The Recruitment Process, Please Let Us Know. Work With The Best Of The Recruitment And Selection Process.
If You Require Accommodation During The Recruitment Process, Please Let Us Know. Work With The Best Of The Recruitment And Selecti
on Process. If You Require Accommodation During The Recruitment Process, Please Let Us Know. Work With The Best Of The Recruitment
And Selection Process. If You Require Accommodation During The Recruitment Process, Please Let Us Know. Work With The Best Of The
```

Figure 7a: Results after 10-epochs



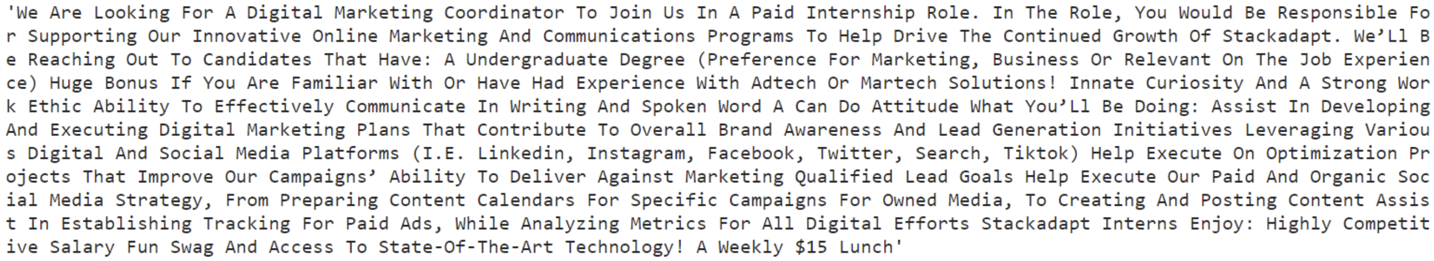
```
'We Are Looking For A Marketing Intern To Join Our Team As A Truly Work Environment With A Wide Variety Of Marketing Initiatives T
o Help Them With The Fortinet Team While Learning About The Fast-Paced, High-Growth Environment, You'Ll Have The Opportunity To Wo
rk On Meaningful Projects, Interact With Various Levels Of The Organization, Network With Industry Professionals, Access Informati
onal Interviews, And Business Overview Sessions. The Intern Will Be Working Based On The Company In Sales, Marketing, And Publicit
y Of Harpercollins Titles From Around The Globe. Harpercollins Canada And Harlequin Are Equal Opportunity Employers Committed To E
qual Employment Opportunities. Employment Decisions Are Based On Job Requirements And The Skills, Knowledge, And Experience Of The
Candidate, Regardless Of Any Other Factors Unrelated To Job Performance. We Are Also Committed To Ensuring That The Accessibility
Needs Of Candidates And Employees Are Considered And Accommodated During The Recruitment Process. We Thank All Applicants For Thei
r Interest, However, Only Those Candidates Selected For An Interview Will Be Contacted. Candidates Must Be Considered. Resume And
Cover Letter Should Be Uploaded As One Document In The Resume Upload Section. Cover Letters Should Be No Longer Than Two Pages. Th
e Application Process Requires A Resume And A Cover Letter Outlining: A. Your Viewpoint'
```

Figure 7b: Results after 20-epochs



```
'We Are Looking For A Digital Marketing Coordinator To Join Us In A Paid Internship Role. In The Role, You Would Be An Opportunity
To Kick Start Your Career? If You Are Ambitious, Proactive, And Highly Motivated, Willing To Roll Up Your Sleeves And Get The Job
Done, We Have The Perfect Job For You! We Are Looking For High-Energy And Ambitious Marketing Intern To Support Our Marketing Team
And Benefit From Our Dedication To Your Personal And Professional Growth. Marketing Intern Duties And Responsibilities: · Collabor
ate On Annual/Seasonal Marketing Planning · Collect Quantitative And Qualitative Data From Marketing Campaigns · Perform Research
On Competition · Support Marketing Daily Administrative Tasks · Help In The Creation And Editing Of Content For Posters, Flyers, A
nd Social Media · Create And Manage Content For Social Media · Help Distribute Marketing Materials · Work With Our Event Marketing
Team To Help Organize Events Marketing Intern Requirements And Qualifications: · Degree In Marketing Or Similar Relevant Field (Ca
n Be In The Process Of Completing Your Degree) · In Depth Understanding Of Different Marketing Techniques · Outstanding Verbal And
Written Communication Skills · Excellent Organizational And Time Management Skills · Attention To Detail · Bonus Points If You Hav
e Creative'
```

Figure 7c: Results after 70-epochs



```
'We Are Looking For A Digital Marketing Coordinator To Join Us In A Paid Internship Role. In The Role, You Would Be Responsible Fo
r Supporting Our Innovative Online Marketing And Communications Programs To Help Drive The Continued Growth Of Stackadapt. We'Ll B
e Reaching Out To Candidates That Have: A Undergraduate Degree (Preference For Marketing, Business Or Relevant On The Job Experien
ce) Huge Bonus If You Are Familiar With Or Have Had Experience With Adtech Or Martech Solutions! Innate Curiosity And A Strong Wor
k Ethic Ability To Effectively Communicate In Writing And Spoken Word A Can Do Attitude What You'Ll Be Doing: Assist In Developing
And Executing Digital Marketing Plans That Contribute To Overall Brand Awareness And Lead Generation Initiatives Leveraging Variou
s Digital And Social Media Platforms (I.E. Linkedin, Instagram, Facebook, Twitter, Search, Tiktok) Help Execute On Optimization Pr
ojects That Improve Our Campaigns' Ability To Deliver Against Marketing Qualified Lead Goals Help Execute Our Paid And Organic Soc
ial Media Strategy, From Preparing Content Calendars For Specific Campaigns For Owned Media, To Creating And Posting Content Assis
t In Establishing Tracking For Paid Ads, While Analyzing Metrics For All Digital Efforts Stackadapt Interns Enjoy: Highly Competit
ive Salary Fun Swag And Access To State-Of-The-Art Technology! A Weekly $15 Lunch'
```

Figure 7d: Results after 100-epochs

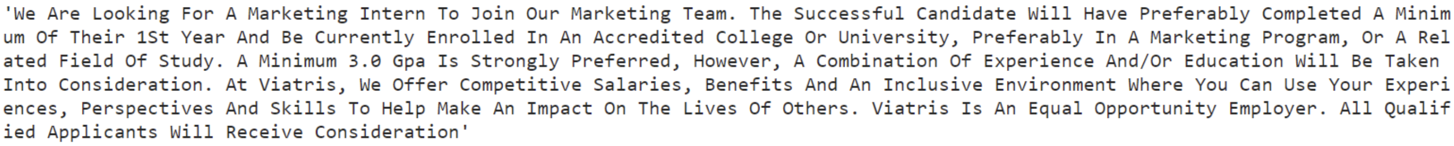
| Epochs | Results |
|---|---|
| 10 | We see that at 10 epochs, the sequence is infinitely repetitive after a point. |
| 20 | Results get better at 20 epochs and the infinite repetition seems to be gone but might occur for longer sequence generations. |
| 70 | At 70 epochs the model is generating good sequences of sentences but the grammar is off-point. |
| 100 | At 100 epochs the model is generating good sequences with good grammar and also capturing currency and price properly. |

Table 8 provides a summary of the results at various epochs.

## Analysis

1. **Using a prompt or seed text**

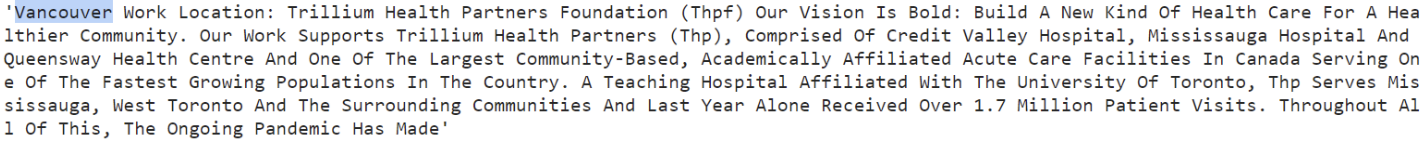
   a. “We are looking a marketing intern”



```
'We Are Looking For A Marketing Intern To Join Our Marketing Team. The Successful Candidate Will Have Preferably Completed A Minim
um Of Their 1St Year And Be Currently Enrolled In An Accredited College Or University, Preferably In A Marketing Program, Or A Rel
ated Field Of Study. A Minimum 3.0 Gpa Is Strongly Preferred, However, A Combination Of Experience And/Or Education Will Be Taken
Into Consideration. At Viatris, We Offer Competitive Salaries, Benefits And An Inclusive Environment Where You Can Use Your Experi
ences, Perspectives And Skills To Help Make An Impact On The Lives Of Others. Viatris Is An Equal Opportunity Employer. All Qualif
ied Applicants Will Receive Consideration'
```

Here we can see that on providing the direct seed text, the model generates a good description with coherent grammar.

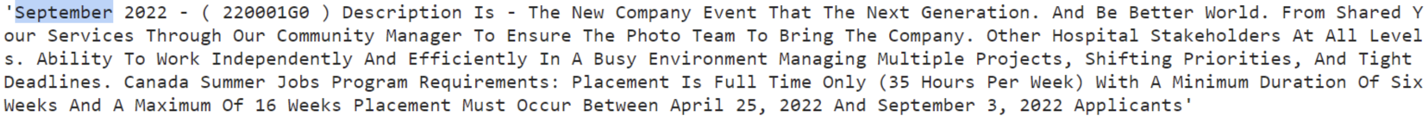
   b. Location as a prompt



```
'Vancouver Work Location: Trillium Health Partners Foundation (Thpf) Our Vision Is Bold: Build A New Kind Of Health Care For A Hea
lthier Community. Our Work Supports Trillium Health Partners (Thp), Comprised Of Credit Valley Hospital, Mississauga Hospital And
Queensway Health Centre And One Of The Largest Community-Based, Academically Affiliated Acute Care Facilities In Canada Serving On
e Of The Fastest Growing Populations In The Country. A Teaching Hospital Affiliated With The University Of Toronto, Thp Serves Mis
sissauga, West Toronto And The Surrounding Communities And Last Year Alone Received Over 1.7 Million Patient Visits. Throughout Al
l Of This, The Ongoing Pandemic Has Made'
```

Here we can see that on providing Vancouver as the prompt, the model correctly captures it as the work location and also generates Canada as a country but the overall description is based on Toronto and hence the model is not aware of the geographical distance between the two cities in the same country.

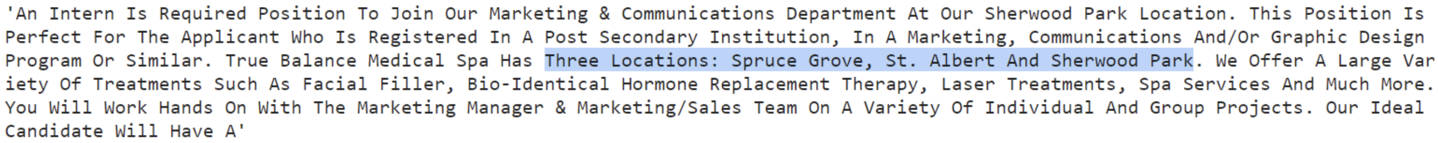
   c. Month as a prompt



```
'September 2022 - ( 220001G0 ) Description Is - The New Company Event That The Next Generation. And Be Better World. From Shared Y
our Services Through Our Community Manager To Ensure The Photo Team To Bring The Company. Other Hospital Stakeholders At All Level
s. Ability To Work Independently And Efficiently In A Busy Environment Managing Multiple Projects, Shifting Priorities, And Tight
Deadlines. Canada Summer Jobs Program Requirements: Placement Is Full Time Only (35 Hours Per Week) With A Minimum Duration Of Six
Weeks And A Maximum Of 16 Weeks Placement Must Occur Between April 25, 2022 And September 3, 2022 Applicants'
```

We can see that the model has correctly recognized September as a date and appended the year 2022.

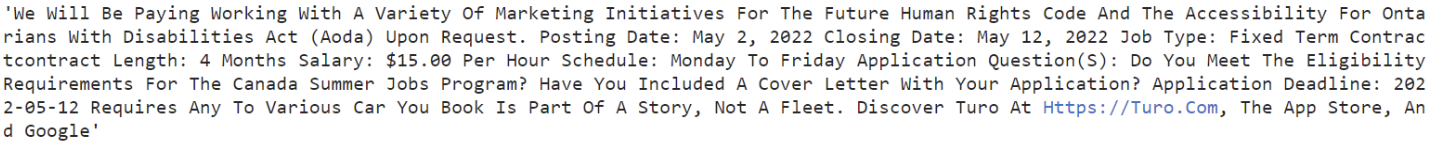
2. **Location generation**



```
'An Intern Is Required Position To Join Our Marketing & Communications Department At Our Sherwood Park Location. This Position Is
Perfect For The Applicant Who Is Registered In A Post Secondary Institution, In A Marketing, Communications And/Or Graphic Design
Program Or Similar. True Balance Medical Spa Has Three Locations: Spruce Grove, St. Albert And Sherwood Park. We Offer A Large Var
iety Of Treatments Such As Facial Filler, Bio-Identical Hormone Replacement Therapy, Laser Treatments, Spa Services And Much More.
You Will Work Hands On With The Marketing Manager & Marketing/Sales Team On A Variety Of Individual And Group Projects. Our Ideal
Candidate Will Have A'
```

Model is able to capture locations.

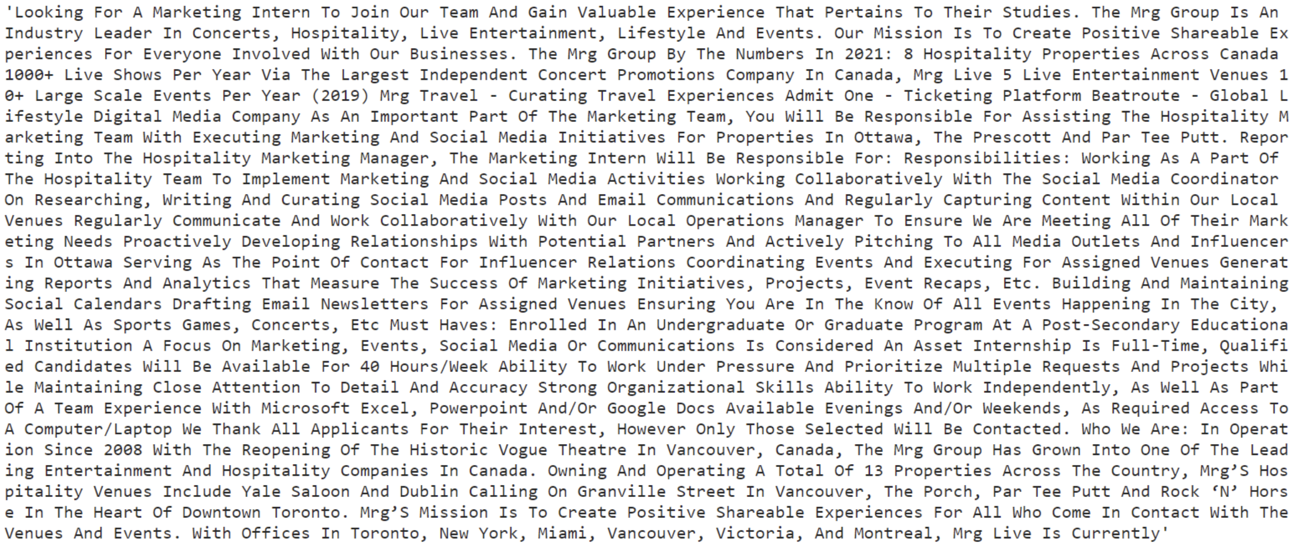
3. **Link capture, salary capture and date capture**



```
'We Will Be Paying Working With A Variety Of Marketing Initiatives For The Future Human Rights Code And The Accessibility For Onta
rians With Disabilities Act (Aoda) Upon Request. Posting Date: May 2, 2022 Closing Date: May 12, 2022 Job Type: Fixed Term Contrac
tcontract Length: 4 Months Salary: $15.00 Per Hour Schedule: Monday To Friday Application Question(S): Do You Meet The Eligibility
Requirements For The Canada Summer Jobs Program? Have You Included A Cover Letter With Your Application? Application Deadline: 202
2-05-12 Requires Any To Various Car You Book Is Part Of A Story, Not A Fleet. Discover Turo At Https://Turo.Com, The App Store, An
d Google'
```

Model is able to capture links, salary and date.

4. **Longer sequence generation**



```
'Looking For A Marketing Intern To Join Our Team And Gain Valuable Experience That Pertains To Their Studies. The Mrg Group Is An
Industry Leader In Concerts, Hospitality, Live Entertainment, Lifestyle And Events. Our Mission Is To Create Positive Shareable Ex
periences For Everyone Involved With Our Businesses. The Mrg Group By The Numbers In 2021: 8 Hospitality Properties Across Canada
1000+ Live Shows Per Year Via The Largest Independent Concert Promotions Company In Canada, Mrg Live 5 Live Entertainment Venues 1
0+ Large Scale Events Per Year (2019) Mrg Travel - Curating Travel Experiences Admit One - Ticketing Platform Beatroute - Global L
ifestyle Digital Media Company As An Important Part Of The Marketing Team, You Will Be Responsible For Assisting The Hospitality M
arketing Team With Executing Marketing And Social Media Initiatives For Properties In Ottawa, The Prescott And Par Tee Putt. Repor
ting Into The Hospitality Marketing Manager, The Marketing Intern Will Be Responsible For: Responsibilities: Working As A Part Of
The Hospitality Team To Implement Marketing And Social Media Activities Working Collaboratively With The Social Media Coordinator
On Researching, Writing And Curating Social Media Posts And Email Communications And Regularly Capturing Content Within Our Local
Venues Regularly Communicate And Work Collaboratively With Our Local Operations Manager To Ensure We Are Meeting All Of Their Mark
eting Needs Proactively Developing Relationships With Potential Partners And Actively Pitching To All Media Outlets And Influencer
s In Ottawa Serving As The Point Of Contact For Influencer Relations Coordinating Events And Executing For Assigned Venues Generat
ing Reports And Analytics That Measure The Success Of Marketing Initiatives, Projects, Event Recaps, Etc. Building And Maintaining
Social Calendars Drafting Email Newsletters For Assigned Venues Ensuring You Are In The Know Of All Events Happening In The City,
As Well As Sports Games, Concerts, Etc Must Haves: Enrolled In An Undergraduate Or Graduate Program At A Post-Secondary Educationa
l Institution A Focus On Marketing, Events, Social Media Or Communications Is Considered An Asset Internship Is Full-Time, Qualifi
ed Candidates Will Be Available For 40 Hours/Week Ability To Work Under Pressure And Prioritize Multiple Requests And Projects Whi
le Maintaining Close Attention To Detail And Accuracy Strong Organizational Skills Ability To Work Independently, As Well As Part
Of A Team Experience With Microsoft Excel, Powerpoint And/Or Google Docs Available Evenings And/Or Weekends, As Required Access To
A Computer/Laptop We Thank All Applicants For Their Interest, However Only Those Selected Will Be Contacted. Who We Are: In Operat
ion Since 2008 With The Reopening Of The Historic Vogue Theatre In Vancouver, Canada, The Mrg Group Has Grown Into One Of The Lead
ing Entertainment And Hospitality Companies In Canada. Owning And Operating A Total Of 13 Properties Across The Country, Mrg'S Hos
pitality Venues Include Yale Saloon And Dublin Calling On Granville Street In Vancouver, The Porch, Par Tee Putt And Rock 'N' Hors
e In The Heart Of Downtown Toronto. Mrg'S Mission Is To Create Positive Shareable Experiences For All Who Come In Contact With The
Venues And Events. With Offices In Toronto, New York, Miami, Vancouver, Victoria, And Montreal, Mrg Live Is Currently'
```

We see that the model is able to capture relevant information i.e. when generating a sequence it is mentioning only one company, capturing the responsibilities and locations of work very well.
However, one can evidently notice that:

- Each word is capitalized even though all words were converted to lowercase during tokenization.
- The use of punctuations can be a hit or a miss (especially full stops) and also the sentence continuity is not maintained.
- The description generated does not maintain the sequence or pattern that is usually followed in job descriptions i.e locations are mentioned but they are scattered, legalities and responsibilities are mentioned but also scattered.
- Sentence formation is not the best.

## Conclusion and Future Work

The project successfully implemented a LSTM model on job description dataset and even though the model performs well in-terms of raw generation power, it lacks grammar capability and sentence formation as well as structure.

Other models such as transformer based encoder-decoder stacks might perform better due to the presence of an attention mechanism. The approach of utilizing LSTM in this project was solely due to resource constraints as transformer training is compute intensive and time consuming.

But this does not mean that the LSTM architecture or the model itself cannot be improved. Since the dataset is web-scraped it is not of the highest quality, hence dataset quality can be improved further and be scaled to incorporate more job roles. The current model itself can be trained for more epochs as the loss is only at 0.14 and can be further reduced.

## References

https://www.turing.com/kb/comprehensive-guide-to-lstm-rnn
https://www.analyticsvidhya.com/blog/2022/02/explaining-text-generation-with-lstm/
https://pypi.org/project/rouge-score/
https://www.tensorflow.org/api_docs/python/tf/keras